Conscious Entities

The Archive

2016 to 2018

Welcome to Conscious Entities, the Archive

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 A few posts have been omitted, and quite a few are rather out of date, but I still enjoy reading them.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

Ungender Me Here

4 January 2016

Male and female brains are pretty much the same, but male and female behaviour is different. It turns out that the same neural circuitry exists in both, but is differently used.

A word of caution. We are talking about mice, in the main: those obliging creatures who seem ready to provide evidence to back all sorts of fascinating theories that somehow don't transfer to human beings. And we're also talking specifically about parental behaviour patterns; it seems those are rather well conserved between species - up to a point - but we shouldn't generalise recklessly.

Catherine Dulac of Harvard explains the gist in this short Scientific American piece. A particular network of neurons in the hypothalamus was observed to be active during nurturing parental behaviour by females; by genetic engineering (amazing what we can do these days) those neurons were edited out of some females who then showed no caring behaviour towards infants. Meanwhile a group of males in which those neurons were stimulated (having been made light-sensitive by even more remarkable genetic manipulation) did show nurturing behaviour.

For male mammals it seems the norm is to kill strange infants on sight (I did say we should be careful about extrapolating to human beings); another set of neurons in the hypothalamus proves to be associated with this behaviour in just the same kind of way as the ones associated with nurturing behaviour.

One of the interesting things here is that both networks exist in both sexes; no-one knows at the moment why one is normally active in females and the other in males. If we were talking about human beings we should be tempted to attribute the difference to cultural factors (I hope that by now nobody is going to be astonished by the idea that different cultural influences could lead to different patterns of physical activity in the brain); it doesn't seem very plausible that that could be the case for mice. It goes without saying that to identify a new hidden factor which sets certain gender roles in mice would inevitably trigger a highly-charged discussion of the possible equivalent in human beings.

So much for the proper research. Could we dare to venture on the irresponsibly speculative hypothesis that men and women habitually think somewhat differently but that each is fully capable of thinking like the other? I shouldn't care to advance that thesis and the whole topic of measuring the quality and style of thought processes is deeply fraught scientifically and beset with difficulty philosophically.

There is, though, one rather striking piece of evidence to suggest that men and women can enter fully into each others minds; novels. Human consciousness is often depicted in novels; indeed the depiction may be the central feature or even pretty much the whole of the enterprise. Jane Austen, who arguably played a major role in making consciousness the centre of narrative, never wrote a scene in which two men converse in the absence of women, allegedly because she considered she could have no direct experience of how men talked to one another in those

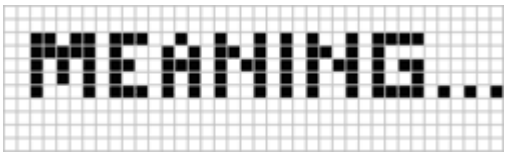
circumstances. Moreover, while she was exceptionally skilful at discreetly incorporating a view from inside her heroines' heads, she never did the same for Darcy or Mr Knightley.

But others have never been so restrained; male authors have depicted the inward world of females and vice versa with very few complaints; that rather suggests we can swap mental gender roles without difficulty.

But are we sure? I suppose it could be argued that as a man I have no more access to the minds of other men than to those of women, so to a degree I actually have to take it on trust that the male mind is accurately depicted in novels. In some cases, certain allegedly typical male thought patterns depicted in books (Nick Hornby choosing a routine football match over a good friend's wedding) are actually rather hard for me to enter into sympathetically. For that matter I recall the indignant rebuttal I got from a female fan when I suggested that Robert Heinlein's depiction of the female mind might be slightly off the mark. Perhaps, then, none of us knows anything about the matter for sure in the end. Still I think nil humanum me alienum puto (I think nothing human alien to me) is a good motto and, I'm slightly encouraged to think, an attainable aspiration.

Jochen's Intentional Automata

10 January 2016



Jochen's paper Von Neumann Minds: Intentional Automata has been published in *Mind and Matter*.

Intentionality is meaningfulness, the quality of being directed at something, aboutness. It is in my view one of the main problems of consciousness, up there with the Hard Problem but quite distinct from it; but it is often under-rated or misunderstood. I think this is largely because our mental life is so suffused with intentionality that we find it hard to see the wood for the trees; certainly I have read more than one discussion by very clever people who seemed to me to lose their way half-way through without noticing and end up talking about much simpler issues.

That is not a problem with Jochen's paper which is admirably clear. He focuses on the question of how to ground intentionality and in particular how to do so without falling foul of an infinite regress or the dreaded homunculus problem. There are many ways to approach intentionality and Jochen briefly mentions and rejects a few (basing it in phenomenal experience or in something like Gricean implicature, for example) before introducing his own preferred framework, which is to root meaning in action: the meaning of a symbol is, or is to be found in, the action it evokes. I think this is a good approach; it interprets intentionality as a matter of input/output relations, which is clarifying and also has the mixed blessing of exposing the problems in their worst and most intractable form. For me it recalls the approach taken by Quine to the translation problem - he of course ended up concluding that assigning certain meanings to unknown words was impossible because of radical under-determination; there are always more possible alternative meanings which cannot be eliminated by any logical procedure. Under-determination is a problem for many theories of intentionality and Jochen's is not immune, but his aim is narrower.

The real target of the paper is the danger of infinite regress. Intentionality comes in two forms, derived on the one hand and original or intrinsic on the other. Books, words, pictures and so on have derived intentionality; they mean something because the author or the audience interprets them as having meaning. This kind of intentionality is relatively easy to deal with, but the problem is that it appears to defer the real mystery to the intrinsic intentionality in the mind of the person doing the interpreting. The clear danger is that we then go on to defer the intentionality to an homunculus, a 'little man' in the brain who again is the source of the intrinsic intentionality.

Jochen quotes the arguments of Searle and others who suggest that computational theories of the mind fail because the meaning and even the existence of a computation is a matter of interpretation and hence without the magic input of intrinsic intentionality from the interpreter fails through radical under-determination. Jochen dramatises the point using an extension of Searle's Chinese Room thought experiment in which it seems the man inside the room can really learn Chinese - but only because he has become in effect the required homunculus.

Now we come to the really clever and original part of the paper; Jochen draws an analogy with the problem of how things reproduce themselves. To do so it seems they must already have a complete model of themselves inside themselves… and so the problem of regress begins. It would be OK if the organism could scan itself, but a proof by Svozil seems to rule that out because of problems with self-reference. Jochen turns to the solution proposed by the great

John Von Neumann (a man who might be regarded as the inventor of the digital computer if Turing had never lived). Von Neumann's solution is expressed in terms of a two-dimensional cellular automaton (very simplistically, a pattern on a grid that evolves over time according to certain rules - Conway's Game of Life surely provides the best-known examples). By separating the functions of copying and interpretation and distinguishing active and passive states Von Neumann managed to get round Svozil successfully.

Now by importing this distinction between active and passive into the question of intentionality, Jochen suggests we can escape the regress. If symbols play either an active or a passive role (in effect, as semantics or as syntax) we can have a kind of automaton which, in a clear sense, gives its own symbols their interpretation, and so escapes the regress.

This is an ingenious move. It is not a complete solution to the problem of intentionality (I think the under-determination monster is still roaming around out here), but it is a novel and very promising solution to the regress. More than that, it offers a new perspective which may well offer further insights when fully absorbed; I certainly haven't managed to think through what the wider implications might be, but if a process so central to meaningful thought truly works in this unexpected dual way it seems there are bound to be some. For that reason, I hope the paper gets wide attention from people whose brains are better at this sort of thing than mine.

No Problem

17 January 2016

Consciousness is not a problem, says Michael Graziano in an *Atlantic* piece that is short and combative. (Also, I'm afraid, pretty sketchy in places. Space constraints might be partly to blame for that, but can't altogether excuse some sweeping assertions made with the broadest of brushes.)

Graziano begins by drawing an analogy with Newton and his theory of light. The earlier view, he says, was that white light was pure, and colour happened when it was 'dirtied' by contact with the surfaces of coloured objects. The detail of exactly how this happened was a metaphysical 'hard problem'. Newton dismissed all that by showing first, that white light is in fact a mixture of all colours, and second, that our vision produces only an inaccurate and simplified model of the reality, with only three different colour receptors.

Consciousness itself, Graziano says, is also a misleading model in a somewhat similar way, generated when the brain represents its own activity to itself. In fact, to be clear, consciousness as represented doesn't happen; it is a mistaken construct, the result of the good-enough but far from perfect apparatus bequeathed to us by evolution (this sounds sort of familiar).

We should be clear that it is really Hard Problem consciousness that is the target here, the consciousness of subjective experience and of qualia. Not that the other sort is OK: Graziano dismisses the Easy Problem kind of consciousness, more or less in passing, as being no problem at all...

> *These days it's not hard to understand how the brain can process information about the world, how it can store and recall memories, how it can construct self knowledge including even very complex self knowledge about one's personhood and mortality. That's the content of consciousness, and it's no longer a fundamental mystery. It's information, and we know how to build computers that process information*.

Amazingly, that's it. Graziano writes in an impatient tone; I have to confess to a slight ruffling of my own patience here; memory is not hard to understand? I had the impression that there were quite a number of unimpeachably respectable scientists working on the neurology of memory, but maybe they're just doing trivial detail, the equivalent of butterfly collecting, or who knows, philosophy? ...we know how to build computers... You know it's not the 1980s any more? Yet apparently there are still clever people who think you can just say that the brain is a computer and that's not only straightforwardly true, but pretty much a full explanation? I mean, the brain is also meat, and we know how to build tools that process meat; shall we stop there and declare the rest to be useless metaphysics?

'Information', as we've often noted before, is a treacherous, ambiguous word. If we mean something akin to data, then yes, computers can handle it; if we mean something akin to understanding, they're no better than meat cleavers. Nothing means anything to a computer, while human consciousness reads and attributes meanings with prodigal generosity, arguably as its most essential, characteristic activity. No computer was ever morally responsible for

anything, while our society is built around the idea that human beings have responsibilities, rights, and property. Perhaps Graziano has debunking arguments for all this that he hasn't leisure to tell us about; the idea that they are all null issues with nothing worthwhile to be said about them just doesn't fly.

Anyway, perhaps I should keep calm because that's not even what Graziano is mainly talking about. He is really after qualia, and in that area I have some moderate sympathy with him; I think it's true that the problem of subjective experience is most often misconceived, and it is quite plausible that the limitations of our sensory apparatus and our colour vision in particular contribute to the confusion. There is a sophisticated argument to be made along these lines: unfortunately Graziano's isn't it; he merely dismisses the issue: our brain plays us false and that's it. You could perhaps get away with that if the problem were simply about our belief that we have qualia; it could be that the sensory system is just misinforming us, the way it does in the case of optical illusions. But the core problem is about people's actual direct experience of qualia. A belief can be wrong, but an experience is still an experience even if it's a misleading one, and the existence of any kind of subjective experience is the real core of the matter. Yes, we can still deny there is any such thing, and some people do so quite cogently, but to say that what I'm having now is not an experience but the mere belief that I'm having an experience is hard and, well, you know, actually rather metaphysical…

On examination I don't think Graziano's analogy with Newton works well. It's not clear to me why the 'older' view is to be characterised as metaphysical (or why that would mean it was worthless). Shorn of the emotive words about dirt, the view that white light picks up colour from contact with coloured things, the way white paper picks up colour from contact with coloured crayons, seems a reasonable enough scientific hypothesis to have started with. It was wrong, but if anything it seems simpler and less abstract than the correct view. Newton himself would not have recognised any clear line between science and philosophy, and in some respects he left the true nature of light a more complicated matter, not fully resolved. His choice of particles over waves has proved to be an over-simplification and remains the subject of some cloudy ontology to this day.

Worse yet, if you think about it, it was Newton who first separated the two realms: colour as it is in the world and colour as we experience it. This is the crucial distinction that opened up the problem of qualia, first recognisably stated by Locke, a fervent admirer of Newton, some years after Newton's work. You could argue therefore, that if the subject of qualia is a mess, it is a mess introduced by Newton himself - and scientists shouldn't castigate philosophers for trying to clear it up.

Babbage's Rival

31 January 2016

Charles Babbage was not the only Victorian to devise a thinking machine.

He is, of course, considered the father, or perhaps the grandfather, of digital computing. He devised two remarkable calculating machines; the Difference Engine was meant to produce error-free mathematical tables for navigation or other uses; the Analytical Engine, an extraordinary leap of the imagination, would have been the first true general-purpose computer. Although Babbage failed to complete the building of the first, and the second never got beyond the conceptual stage, his achievement is rightly regarded as a landmark, and the Analytical Engine routines published by Lady Lovelace in 1843 with a translation of Menabrea's description of the Engine, have gained her recognition as the world's first computer programmer.

The digital computer, alas, went no further until Turing a hundred years later; but in 1851 Alfred Smee published The Process of Thought adapted to Words and Language together with a description of the Relational and Differential Machines - two more designs for cognitive mechanisms.

Smee held the unusual post of Surgeon to the Bank of England - in practice he acted as a general scientific and technical adviser. His father had been Chief Accountant to the Bank and little Alfred had literally grown up in the Bank, living inside its City complex. Apparently, once the Bank's doors had shut for the night, the family rarely went to the trouble of getting them unlocked again to venture out; it must have been a strangely cloistered life. Like Babbage and other Victorians involved in London's lively intellectual life, Smee took an interest in a wide range of topics in science and engineering, with his work in electro-metallurgy leading to the invention of a successful battery; he was a leading ophthalmologist and among many other projects he also wrote a popular book about the garden he created in Wallington, south of London - perhaps compensating for his stonily citified childhood?

Smee was a Fellow of the Royal Society, as was Babbage, and the two men were certainly acquainted (Babbage, a sociable man, knew everyone anyway and was on friendly terms with all the leading scientists of the day; he even managed to get Darwin, who hated socialising and suffered persistent stomach problems, out to some of his parties). In fact, we know that Babbage, who had persistent eye problems, was treated by Smee, one of the best eye doctors in London at the time. However, there's no record of the two ever discussing computing, and Smee's book never mentions Babbage.

That might be in part because Smee came at the idea of a robot mind from a different, biological angle. As a surgeon he was interested in the nervous system and was a proponent of 'electro-biology', advocating the modern view that the mind depends on the activity of the brain. At public lectures he exhibited his own 'injections' of the brain, revealing the complexity of its

structure; but Golgi's groundbreaking methods of staining neural tissue were still in the future, and Smee therefore knew nothing about neurons.

Smee nevertheless had a relatively modern conception of the nervous system. He conducted many experiments himself (he used so many stray cats that a local lady was moved to write him a letter warning him to keep clear of hers) and convinced himself that photovoltaic effects in the eye generated small currents which were transmitted bio-electrically along the nerves and processed in the brain. Activity in particular combinations of 'nervous fibrils' gave rise to awareness of particular objects. The gist is perhaps conveyed in the definition of consciousness he offered in an earlier work, Principles of the Human Mind Deuced from Physical Laws:

> *When an image is produced by an action upon the external senses, the actions on the organs of sense concur with the actions in the brain; and the image is then a Reality.*
>
> *When an image occurs to the mind without a corresponding simultaneous action of the body, it is called a Thought.*
>
> *The power to distinguish between a thought and a reality, is called Consciousness.*

This is not very different in broad terms from a lot of current thinking.

In *The Process of Thought* Smee takes much of this for granted and moves on to consider how the brain deals with language. The key idea is that the brain encodes things into a pyramidal classification hierarchy. Smee begins with an analysis of grammar, faintly Chomskyan in spirit if not in content or level of innovation. He then moves on rapidly to the construction of words. If his pyramidal structure is symbolically populated with the alphabet different combinations of nervous activity will trigger different combinations of letters and so produce words and sentences. This seems to miss out some essential linguistic level, leaving the impression that all language is spelled out alphabetically, which can hardly be what Smee believed.

When not dealing specifically with language the letters in Smee's system correspond to qualities and this pyramid stands for a universal categorisation of things in which any object can be represented as a combination of properties. (This rather recalls Bishop Wilkins' proposed universal language, in which each successive letter of each noun identifies a position in an hierarchical classification, so that the name is an encoded description of the thing named.)

At least, I think that's the way it works. The book goes on to give an account of induction, deduction, and the laws of thought; alas, Smee seemsto be unaware of the problem diagnosed by Hume and does not address it. Instead, in essence, he just describes the processes; although he frames the discussion in terms of his pyramidal classification his account of induction (he suggests six different kinds) comes down to saying that if we observe two characteristics constantly together we assume a link. Why do we do that - and why is it we actually often don't? Worse than that, he mentions simple arithmetic (one plus one equals two, two times two is four) and says:

> *These instances are so familiar we are apt to forget that they are inductions...*

Alas, they're not inductions. (You could arrive at them by induction, but no-one ever actually does and our belief in them does not rest on induction.)

I'm afraid Smee's laws of thought also stand on a false premise; he says that the number of ideas denoted by symbols is finite, though too large for a man to comprehend. This is false. He might have been prompted to avoid the error if he had used numbers instead of letters for his

pyramid - because each integer represents an idea; the list of integers goes on forever, yet our numbering system provides a unique symbol for every one? So neither the list of ideas nor the list of symbols can be finite. Of course that barely scratches the surface of the infinitude of ideas and symbols, but it helps suggest just how unmanageable a categorisation of every possible idea really is.

But now we come to the machines designed to implement these processes. Smee believed that his pyramidal structure could be implemented in a hinged physical mechanism where opening would mean the presence or existence of the entity or quality and closing would mean its absence. One of these structures provides the Relational Machine. It can test membership of categories, or the possession of a particular attribute, and can encode an assertion, allowing us to test consistency of that assertion with a new datum. I have to confess to having only a vague understanding of how this would really work. He allows for partial closing and I think the idea is that something like predicate calculus could be worked out this way. He says at one point that arithmetic could be done with this mechanism and that anyone who understands logarithms will readily see how; I'm afraid I can only guess what he had in mind.

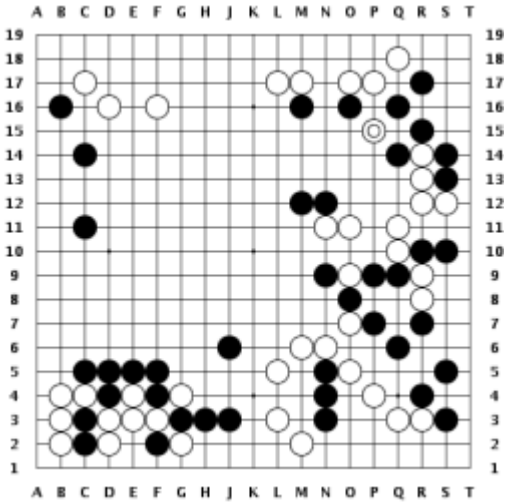
It isn't necessary to have two Relational Machines to deal with multiple assertions because we can take them in sequence; however the Differential Machine provides the capacity to compare directly, so that we can load into one side all the laws and principles that should guide a case while uploading the facts into the other.

Smee had a number of different ideas about how the machines could be implemented, and says he had a number of part-completed prototypes of partial examples on his workbench. Unlike Babbage's designs, his were never meant to be capable of full realisation, though; although he thinks it is finite he says the Relational Machine would cover London and the mechanical stresses would destroy it immediately if it were ever used; moreover, using it to crank out elementary deductions would be so slow and tedious people would soon revert to using their wonderfully compact and efficient brains instead. But partial examples will helpfully illustrate the process of thought and help eliminate mistakes and 'quibbles'. Later chapters of the book explore things that can go wrong in legal cases, and describe a lot of the quibbles Smee presumably hopes his work might banish.

I think part of the reason Smee's account isn't clearer (to me, anyway) is that his ideas were never critiqued by colleagues and he never got near enough to a working prototype to experience the practical issues sufficiently. He must have been a somewhat lonely innovator in his lab in the Bank and in fairness the general modernity of his outlook makes us forget how far ahead of his time he was. When he published his description of his machines, Wundt, generally regarded as the founder of scientific psychology, was still an undergraduate. To a first approximation, nobody knew anything about psychology or neurology. Logic was still essentially in the long Aristotelian twilight - and of course we know where computing stood. It is genuinely remarkable that Smee managed, over a hundred and fifty years ago, to achieve a proto-modern, if flawed, understanding of the brain and how it thinks. Optimists will think that shows how clever he was; pessimists will think it shows how little our basic conceptual thinking has been updated by the progress of cognitive science.

Go AI

7 February 2016



The recent victory scored by the AlphaGo computer system over a professional Go player might be more important than it seems.

At first sight it seems like another milestone on a pretty well-mapped road; significant but not unexpected. We've been watching games gradually yield to computers for many years; chess notoriously, was one they once said was permanently out of the reach of the machines. All right, Go is a little bit special. It's an extremely elegant game; from some of the simplest equipment and rules imaginable it produces a strategic challenge of mind-bending complexity, and one whose combinatorial vastness seems to laugh scornfully at Moore's Law - maybe you should come back when you've got quantum computing, dude! But we always knew that that kind of confidence rested on shaky foundations; maybe Go is in some sense the final challenge, but sensible people were always betting on its being cracked one day.

The thing is, Go has not been beaten in quite the same way as chess. At one time it seemed to be an interesting question as to whether chess would be beaten by intelligence - a really good algorithm that sort of embodied some real understanding of chess - or by brute force; computers that were so fast and so powerful they could analyse chess positions exhaustively. That was a bit of an oversimplification, but I think it's fair to say that in the end brute force was the major factor. Computers can play chess well, but they do it by exploiting their own strengths, not by doing it through human-style understanding. In a way that is disappointing because it means the successful systems don't really tell us anything new.

Go, by contrast, has apparently been cracked by deep learning, the technique that seems to be entering a kind of high summer of success. Oversimplifying again, we could say that the history of AI has seen a contest between two tribes; those who simply want to write programs that do what's needed, and those that want the computer to work it out for itself, maybe using networks and reinforcement methods that broadly resemble the things the human brain seems to do. Neither side, frankly, has altogether delivered on its promises and what we might loosely call the machine learning people have faced accusations that even when their systems work, we don't know how and so can't consider them reliable.

What seems to have happened recently is that we have got better at deploying several different approaches effectively in concert. In the past people have sometimes tried to play golf with only one club, essentially using a single kind of algorithm which was good at one kind of task. The new Go system, by contrast, uses five different components carefully chosen for the task they were to perform; and instead of having good habits derived from the practice and insights of human Go masters built in, it learns for itself, playing through thousands of games.

This approach takes things up to a new level of sophistication and clearly it is yielding remarkable success; but it's also doing it in a way which I think is vastly more interesting and promising than anything done by Deep Thought or Watson. Let's not exaggerate here, but this

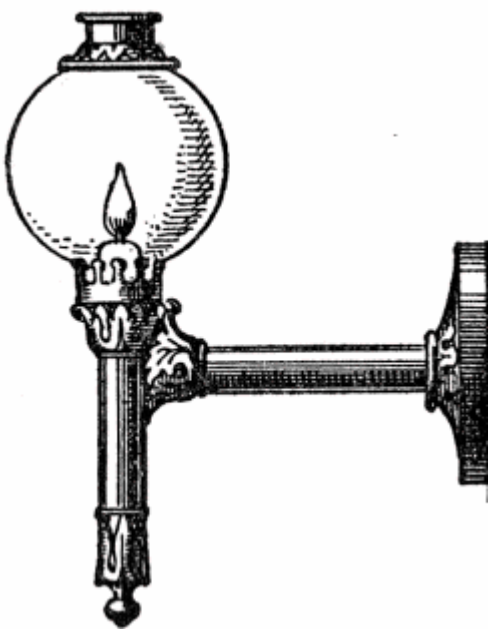
kind of machine learning looks just a bit more like actual thought. Claims are being made that it could one day yield consciousness; usually, if we're honest, claims like that on behalf of some new system or approach can be dismissed because on examination the approach is just palpably not the kind of thing that could ever deliver human-style cognition; I don't say deep learning is the answer, but for once, I don't think it can be dismissed.

Demis Hassabis, who led the successful Google DeepMind project, is happy to take an optimistic view; in fact he suggests that the best way to solve the deep problems of physics and life may be to build a deep-thinking machine clever enough to solve them for us (where have I heard that idea before?). The snag with that is that old objection; the computer may be able to solve the problems, but we won't know how and may not be able to validate its findings. In the modern world science is ultimately validated in the agora; rival ideas argue it out and the ones with the best evidence wins the day. There are already some emergent problems, with proofs achieved by an exhaustive consideration of cases by computation that no human brain can ever properly validate.

More nightmarish still the computer might go on to understand things we're not capable of understanding. Or seem to: how could we be sure?

Free Won't

14 February 2016



Was Libet wrong? The question has often been asked since the famous experiments which found that a detectable "Readiness Potential" (RP) showed that our decisions were made half a second before they entered our consciousness. There are plenty of counter arguments but the findings themselves have been reproduced and seem unassailable.

However, Christian Jarrett reports two new pieces of research which shed fresh light on the issue (Jarrett's book Great Myths of the Brain is very sensible, btw, and should be read by all science journalists).

The first piece of research shows that although an action may be prepared by the brain before we know we've decided to act, we can still change our minds. Subjects were asked to press a button at a moment of their choice after a green light came on. However, if a red light appeared before they pressed, they were asked to refrain. The experimenters then detected the RP which showed the subjects were about to press the button, and used the red light to try to cancel the intention. They found that although there was a 'point of no return', so long as the red light appeared in time the subjects were able to hold off and not press the button after all.

That means that a significant qualification has to be added to Libet's initial findings - but it's one Libet himself had already come up with. He was aware that the emergent action could be suppressed, and memorably said it showed that while we might not have free will, we could still have 'free won't'. I don't know much use that is. In the experiment the subjects simply responded to a red light, but if the veto is to make consciousness effective again we seem to need the veto decision to happen instantly and overtake the action decision somehow. That seems problematic if not paradoxical. I also get quite confused trying to imagine what it would be like to veto mentally a decision you're not yet aware of having made. Still, I suppose we must be grateful for whatever residual degree of freedom the experiments allow us.

The second piece of research calls into question the nature of the RP itself. Libet's research more or less took it for granted that the RP was a reliable sign that an action was on the way, but the new findings suggest that it is really just part of the background ebb and flow of neural noise. Actions do arise when the activity crosses a certain threshold, but the brain is much quicker about getting to that level when the background activity is already high.

That certainly adds some complexity to the picture, but I don't think it really dispels the puzzle of Libet's results. The RP may be fuzzier than we thought and it may not have as rigid a link to action as we thought - but it's still possible to predict the act before we're aware of the decision. What if we redesigned the first experiment? We tell the subject to click on a target at any time after it appears; but we detect the RP and whip the target away a moment before the click every time. The subject can never succeed. The fact that that is perfectly possible surely remains more than a little unsettling.

Whereof We Cannot Speak

21 February 2016

The latest JCS features a piece by Christopher Curtis Sensei about the experience of achieving mastery in Aikido. It seems he spent fifteen years cutting bokken (an exercise with wooden swords, don't ask me), becoming very proficient technically but never satisfying the old Sensei. Finally he despaired and stopped trying; at which point, of course, he made the required breakthrough. He needed to stop thinking about it. You do feel that his teacher could perhaps have saved him a few years if he had just said so explicitly - but of course you cannot achieve the state of not thinking about something directly and deliberately. Intending to stop thinking about a pink hippo involves thinking about a pink hippo; you have to do something else altogether.

This unreflective state of mind crops up in many places; it has something to do with the desirable state of 'flow' in which people are said to give their best sporting or artistic performances; it seems to me to be related to the popular notion of mindfulness, and it recalls Taoist and other mystical ideas about cleared minds and going with the stream. To me it evokes Julian Jaynes, who believed that in earlier times human consciousness manifested itself to people as divine voices; what we're after here is getting the gods to shut up at last.

Clearly this special state of mind is a form of consciousness (we don't pass out when we achieve it) and in fact on one level I think it is very simple. It's just the absence of second-order consciousness, of thoughts about thoughts, in other words. Some have suggested that second-order thought is the distinctive or even the constitutive feature of human consciousness; but it seems clear to me that we can in fact do without it for extended periods.

All pretty simple then. In fact we might even be able to define it physiologically - it could be the state in which the cortex stops interfering and let's the cerebellum and other older bits of the brain do their stuff uninterrupted - we might develop a way of temporarily zapping or inhibiting cortical activity so we can all become masters of whatever we're doing at the flick of a switch. What's all the fuss about?

Except that arguably none of the foregoing is actually about this special state of mind at all. What we're talking about is unconsidered thought, and I cannot report it or even refer to it without considering it; so what have I really been discussing? Some strange ghostly proxy? Nothing? Or are these worries just obfuscatory playing with words?

There's another mental thing we shouldn't, logically, be able to talk about - qualia. Qualia, the ineffable subjective aspect of things, are additional to the scientific account and so have no causal powers; they cannot therefore ever have caused any of the words uttered or written about them. Is there a link here? I think so. I think qualia are pure first-order experiences; we cannot talk about them because to talk about them is to move on to second-order cognition and so to slide away from the very thing we meant to address. We could say that qualia are the experiential equivalent of the pure activity which Curtis Sensei achieved when he finally cut

bokken the right way. Fifteen years and I'll understand qualia; I just won't be able to tell anyone about it

Covert Consciousness

28 February 2016

Are we being watched? Over at Aeon, George Musser asks whether some AI could quietly become conscious without our realising it. After all, it might not feel the need to stop whatever it was doing and announce itself. If it thought about the matter at all, it might think it was prudent to remain unobserved. It might have another form of consciousness, not readily recognisable to us. For that matter, we might not be readily recognisable to it, so that perhaps it would seem to itself to be alone in a solipsistic universe, with no need to try to communicate with anyone.

There have been various scenarios about this kind of thing in the past which I think we can dismiss without too much hesitation. I don't think the internet is going to attain self-awareness because however complex it may become, its simply isn't organised in the right kind of way. I don't think any conventional digital computer is going to become conscious either, for similar reasons.

I think consciousness is basically an expanded development of the faculty of recognition. Animals have gradually evolved the ability to recognises very complex extended stimuli; in the case of human beings things have gone a massive leap further so that we can recognise abstractions and generalities. This makes a qualitative change because we are no longer reliant on what is coming in through our sense from the immediate environment; we can think about anything, even imaginary or nonsensical things.

I think this kind of recognition has an open-ended quality which means it can't be directly written into a functional system; you can't just code it up or design the mechanism. So no machines have been really good candidates; until recently. These days I think some AI systems are moving into a space where they learn for themselves in a way which may be supported by their form and the algorithms that back them up, but which does have some of the open-ended qualities of real cognition. My perception is that we're still a long way from any artificial entity growing into consciousness; but it's no longer a possibility which can be dismissed without consideration; so a good time for George to be asking the question.

How would it happen? I think we have to imagine that a very advanced AI system has been set to deal with a very complex problem. The system begins to evolve approaches which yield results and it turns out that conscious thought - the kind of detachment from immediate inputs I referred to above - is essential. Bit by bit (ha) the system moves towards it.

I would not absolutely rule out something like that; but I think it is extremely unlikely that the researchers would fail to notice what was happening.

First, I doubt whether there can be forms of consciousness which are unrecognisable to us. If I'm right consciousness is a kind of function which yields purposeful output behaviour, and purposefulness implies intelligibility. We would just be able to see what it was up to. Some qualifications to this conclusion are necessary. We've already had chess AIs that play certain

end-games in ways that don't make much sense to human observers, even chess masters, and look like random flailing. We might get some patterns of behaviour like that. But the chess 'flailing' leads reliably to mate, which ultimately is surely noticeable. Another point to bear in mind is that our consciousness was shaped by evolution, and by the competition for food, safety, and reproduction. The supposed AI would have evolved in consciousness in response to completely different imperatives, which might well make some qualitative difference. The thoughts of the AI might not look quite like human cognition. Nevertheless I still think the intentionality of the AI's outputs could not help but be recognisable. In fact the researchers who set the thing up would presumably have the advantage of knowing the goals which had been set.

Second, we are really strongly predisposed to recognising minds. Meaningless whistling of the wind sounds like voices to us; random marks look like faces; anything that is vaguely humanoid in form or moves around like a sentient creature is quickly anthropomorphised by us and awarded an imaginary personality. We are far more likely to attribute personhood to a dumb robot than dumbness to one with true consciousness. So I don't think it is particularly likely that a conscious entity could evolve without our knowing it and keep a covert, wary eye on us. It's much more likely to be the other way around: that the new consciousness doesn't notice us at first.

I still think in practice that that's a long way off; but perhaps the time to think seriously about robot rights and morality has come.

Thinking Backwards

5 March 2016

In the course of a review of Carlo Rovelli's Seven Brief Lessons on Physics, Alva Noë asks a fascinating question: does it make any sense to imagine that one might think backward?

In physics we are accustomed to thinking that entropy always increases; things run down, spread out, differences are evened out and available energy always decreases. Are cognitive processes like that, always going in one underlying direction? Are they like a river which flows to the sea, or are they more like a path we can take in either direction?

Well, we can't talk backwards without careful preparation, and some kinds of conscious thought resemble talking to oneself. In fact, we can't perform most tasks backwards without rehearsal. Unless one takes the trouble to learn zedabetical order, we cannot easily recite our letters in reverse. So it doesn't seem we can do that kind of backward thinking, at least. We can, of course, read a written sentence in the reverse word order and I suppose that in that sense we can think the same sentence backwards, or 'backwards sentence same the think'; but we might well doubt whether that truly reverses the thought.

On the other hand, on another level, there seems to be no inherent direction to a chain of thought. If I think of an egg, it makes me think of breakfast, and breakfast makes me think of getting up; but thinking about getting up might equally have prompted thoughts of breakfast, which in turn might have brought an egg to mind. The association of ideas goes both ways. Logical thought needs a little more care. 'If p then q' does not allow us to conclude that if q then p - but we can say that not-q gives us not-p. With due attention to the details there doesn't seem to be a built-in direction to deduction.

The universal presence of memory in all our thoughts seems a deal-breaker on reversibility, though. We cannot forget things at will; memories constantly accumulate; and we cannot deliberately unthink a thought in such a way as to erase it from recollection. This one-way effect seems to be equally true of certain kinds of thought on another level. We have, let's say, a problem in mind; all the clues we need are known. At some point they come together to point to the right conclusion - Aha! Unless we find a flaw in our reasoning, the meshing of the clues cannot be unpicked. We can't ununderstand; the process of comprehension seems as irrevocable as the breaking of an eggshell or the cooking of a omelette.

There's an odd thing about it though. The breaking of the egg is an example of the wider principle of entropy; the shell is destroyed, and later the protein is denatured by heat. The solving of the problem, by contrast, is constructive; we're left with more than we had before and our mental contents are better structured, not worse. Learning and understanding seems like a process of growth. Like the growth of a plant it is of course just a local reversal of entropy, which has to be paid for through the use of a lot of irretrievable energy; still, it's remarkable (as is the growth of a plant, after all).

Hang on, though. Isn't it the case that we might have been entertaining a dozen different ideas about the possible solution to our problem? Once we have the answer those other hypotheses

are dropped and indeed may even be truly forgotten. More than that, the right answer may let us simplify our ideas, the way Copernicus let us do away with epicycles and all the rest of the Ptolemaic system nobody has to think about any more. Occam tells us to minimise the number of angels we require for our theory, so isn't the growth of understanding sometimes a synthesis which actually has the character of a reductive simplification? That isn't usually a reversal as such, but doesn't it involve in some sense the unthinking of certain thoughts? David Lodge somewhere has a character feel a pang of pity for all the Marxist professors in the universities of no-longer-communist countries who must presumably unspool years of patient theoretical exegesis in order to start understanding the world again.

Well, yes, but I don't think that is truly an unthinking or reverse cogitation. Is it perhaps more like a plant managing to grow downwards? So no, as Noë implied in the first place, it doesn't really make sense to imagine we might think backward.

Phlegm Theory

14 March 2016

Worse than wrong? A trenchant piece from Michael Graziano likens many theories of consciousness to the medieval theory of humours; in particular the view that laziness is due to a build up of phlegm. It's not that the theory is wrong, he says - though it is - it's that it doesn't even explain anything.

To be fair I think the theory of the humours was a little more complex than that, and there is at least some kind of hand-waving explanatory connection between the heaviness of phlegm and slowness of response. According to Graziano such theories flatter our intuitions; they offer a vague analogy which feels metaphorically sort of right - but, on examination, no real mechanism. His general point is surely very sound; there are indeed too many theories about conscious experience that describe a reasonably plausible process without ever quite explaining how the process magically gives rise to actual feeling, to the ineffable phenomenology.

As an example, Graziano mentions a theory that neural oscillations are responsible for consciousness; I think he has in mind the view espoused by Francis Crick and others that oscillations at 40 hertz give rise to awareness. This idea was immensely popular at one time and people did talk about "40 hertz" as though it was a magic key. Of course it would have been legitimate to present this as an enigmatic empirical finding, but the claim seemed to be that it was an answer rather than an additional question. So far as I know Graziano is right to say that no-one ever offered a clear view as to why 40 hertz had this exceptional property, rather than 30 or 50, or for that matter why co-ordinated oscillation at any frequency should generate consciousness. It is sort of plausible that harmonising on a given frequency might make parts of the brain work together in some ways, and people sometimes took the view that synchronised firing might, for example, help explain the binding problem - the question of how inputs from different senses arriving at different times give rise to a smooth and flawlessly co-ordinated experience. Still, at best working in harmony might explain some features of experience: it's hard to see how in itself it could provide any explanation of the origin or essential nature of consciousness. It just isn't the right kind of thing.

As a second example Graziano boldly denounces theories based on integrated information. Yes, consciousness is certainly going to require the integration of a lot of information, but that seems to be a necessary, not a sufficient condition. Intuitively we sort of imagine a computer getting larger and more complex until, somehow, it wakes up. But why would integrating any amount of information suddenly change its inward nature? Graziano notes that some would say dim sparks of awareness are everywhere, so that linking them gives us progressively brighter arrays. That, however, is no explanation, just an even worse example of phlegm.

So how does Graziano explain consciousness? He concedes that he too has no brilliant resolution of the central mystery. He proposes instead a project which asks, not why we have subjective experience, but why we think we do: why we say we do with such conviction. The answer, he suggests, is in metacognition. (This idea will not be new to readers who are acquainted with Scott Bakker's Blind Brain Theory.) The mind makes models of the world and

models of itself, and it is these inaccurate models and the information we generate from them that makes us see something magic about experience. In the brief account here I'm not really sure Graziano succeeds in making this seem more clear-cut than the theories he denounces. I suppose the parallel existence of reality and a mental model of reality might plausibly give rise to an impression that there is something in our experience over and above simple knowledge of the world; but I'm left a little nervous about whether that isn't another example of the kind of intuition-flattering the other theories provide.

This kind of metacognitive theory tends naturally to be a sceptical theory; our conviction that we have subjective experience proceeds from an error or a defective model, so the natural conclusion, on grounds of parsimony if no others, is that we are mistaken and there is really nothing special about our brain's data processing after all.

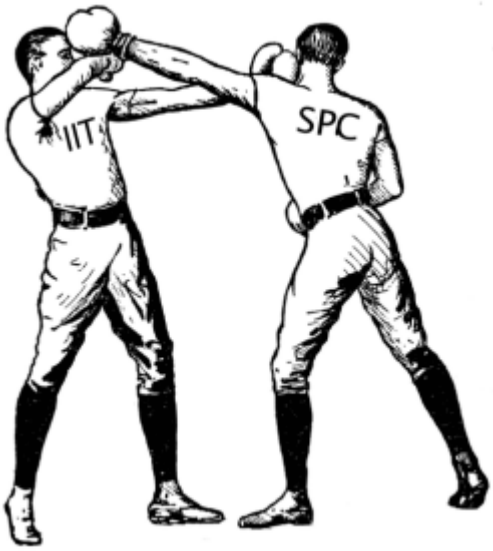
That may be the natural conclusion, but in other respects it's hard to accept. It's easy to believe that we might be mistaken about what we're experiencing, but can we doubt that we're having an experience of some kind? We seem to run into quasi-Cartesian difficulties.

Be that as it may Graziano deserves a round of applause for his bold (but not bilious) denunciation of the phlegm.

This Town Ain't Big Enough...

21 March 2016

...for two theories?



Ihtio kindly drew my attention to an interesting paper which sets integrated information theory (IIT) against its own preferred set of ideas - semantic pointer competition (SPC). I'm not quite sure where this 'one on one' approach to theoretical discussion comes from. Perhaps the authors see IIT as gaining ground to the extent that any other theory must now take it on directly. The effect is rather of a single bout from some giant knock-out tournament of theories of consciousness (I would totally go for that, incidentally; set it up, somebody!).

We sort of know about IIT by now, but what is SPC? The authors of the paper, Paul Thagard and Terrence C Stewart, suggest that:

> *consciousness is a neural process resulting from three mechanisms: representation by firing patterns in neural populations, binding of representations into more complex representations called semantic pointers, and competition among semantic pointers to capture the most important aspects of an organism's current state.*

I like the sound of this, and from the start it looks like a contender. My main problem with IIT is that, as was suggested last time, it seems easy enough to imagine that a whole lot of information could be integrated but remain uniluminated by consciousness; it feels as if there needs to be some other functional element; but if we supply that element it looks as if it will end up doing most of the interesting work and relegate the integration process to something secondary or even less important. SPC looks to be foregrounding the kind of process we really need.

The authors provide three basic hypotheses on which SPC rests;

> *H1. Consciousness is a brain process resulting from neural mechanisms.*
>
> *H2. The crucial mechanisms for consciousness are: representation by patterns of firing in neural populations, binding of these representations into semantic pointers, and competition among semantic pointers.*
>
> *H3. Qualitative experiences result from the competition won by semantic pointers that unpack into neural representations of sensory, motor, emotional, and verbal activity.*

The particular mention of the brain in H1 is no accident. The authors stress that they are offering a theory of how brains work. Perhaps one day we'll find aliens or robots who manage some form of consciousness without needing brains, but for now we're just doing the stuff we know about. "...a theory of consciousness should not be expected to apply to all possible conscious entities."

Well, actually, I'd sort of like it to - otherwise it raises questions about whether it really is consciousness itself we're explaining. The real point here, I think, is meant to be a criticism of IIT, namely that it is so entirely substrate-neutral that it happily assigns consciousness to anything that is sufficiently filled with integrated information. Thagard and Stewart want to distance themselves from that, claiming it as a merit of their theory that it only offers consciousness to brains. I sympathise with that to a degree, but if it were me I'd take a slightly different line, resting on the actual functional features they describe rather than simple braininess. The substrate does have to be capable of doing certain things, but there's no need to assume that only neurons could conceivably do them.

The idea of binding representations into 'semantic pointers' is intriguing and seems like the right kind of way to be going; what bothers me most here is how we get the representations in the first place. Not much attention is given to this in the current paper: Thagard and Stewart say neurons that interact with the world and with each other become "tuned" to regularities in the environment. That's OK, but not really enough. It can't be that mere interaction is enough, or everything would be a prolific representation of everything around it; but picking out the right "regularities" is a non-trivial task, arguably the real essence of representation.

Competition is the way particular pointers get selected to enter consciousness, according to H2; I'm not exactly sure how that works and I have doubts about whether open competition will do the job. One remarkable thing about consciousness is its coherence and direction, and unregulated competition seems unlikely to produce that, any more than a crowd of people struggling for access to a microphone would produce a fluent monologue. We can imagine that a requirement for coherence is built in, but the mechanism that judges coherence turns out to be rather important and rather difficult to explain.

So does SPC deliver? H3 claims that it gives rise to qualitative experience: the paper splits the issue into two questions: first, why are there all these different experiences, and second, why is there any experience at all? On the first, the answers are fairly good, but not particularly novel or surprising; a diverse range of sensory inputs and patterns of neural firing naturally give rise to a diversity of experience. On the second question, the real Hard Problem, we don't really get anywhere; it's suggested that actual experience is an emergent property of the three processes of consciousness. Maybe it is, but that doesn't really explain it. I can't seriously criticise Thagard and Stewart because no-one has really done any better with this; but I don't see that SPC has a particular edge over IIT in this respect either.

Not that their claim to superiority rests on qualia; in fact they bring a range of arguments to suggest that SPC is better at explaining various normal features of consciousness. These vary in strength, in my opinion. First feature up is how consciousness starts and stops. SPC has a good account, but I think IIT could do a reasonable job, too. The second feature is how consciousness shifts, and this seems a far stronger case; pointers naturally lend themselves better to thus than the gradual shifts you would at first sight expect from a mass of integrated information. Next we have a claim that SPC is better at explaining the different kinds or grades of consciousness that fifteen organisms presumably have. I suppose the natural assumption, given IIT, would be that you either have enough integration for consciousness or you don't. Finally, it's claimed that SPC is the winner when it comes to explaining the curious unity/disunity of consciousness. Clearly SPC has some built-in tools for binding, and the authors suggest that competition provides a natural source of fragmentation. They contrast this with Tononi's concept of quantity of consciousness, an idea they disparage as meaningless in the face of the mental diversity of the organisms in the world.

As I say, I find some of these points stronger than others, but on the whole I think the broad claim that SPC gives a better picture is well founded. To me it seems the advantages of SPC mainly flow from putting representation and pointers at the centre. The dynamic quality this provides, and the spark of intentionality, make it better equipped to explain mental functions than the more austere apparatus of IIT. To me SPC is like a vehicle that needs overhauling and some additional components (some of those not readily available); it doesn't run just now but you can sort of see how it would. IIT is more like an elegant sculptural form which doesn't seem to have a place for the wheels.

Zappiens Unreads

3 April 2016

Are our minds being dumbed by digits – or set free by unreading?

Frank Furedi notes that it has become common to deplore a growing tendency to inattention. In fact, he says, this kind of complaint goes back to the eighteenth century. Very early on the failure to concentrate was treated as a moral failing rather than simple inability; Furedi links this with the idea that attention to proper authority is regarded as a duty, so that inattention amounts to disobedience or disrespect. What has changed more recently, he suggests, is that while inattention was originally regarded as an exceptional problem, it is now seen as our normal state, inevitable: an attitude that can lead to fatalism.

The advent of digital technology has surely affected our view. Since the turn of the century or earlier there have been warnings that constant use of computers, and especially of the Internet, would change the way our brains worked; would damage us intellectually if not morally. Various kinds of damage have been foreseen; shortened attention span, lack of memory, dependence on images, lack of concentration, failure of analytical skills and inability to pull the torrent of snippets into meaningful structures. 'Digital natives' might be fluent in social media and habituated to their own strange new world, but there was a price to pay. The emergence of Homo Zappiens has been presented as cause for concern, not celebration.

Equally there have been those who suggest that the warnings are overstated. It would, they say, actually be strange and somewhat disappointing if study habits remained exactly the same after the advent of an instant, universal reference tool; the brain would not be the highly plastic entity we know it to be if it didn't change its behaviour when presented with the deep interactivity that computers offer; and really it's time we stopped being surprised that changes in the behaviour of the mind show up as detectable physical changes in the brain.

In many respects, moreover, people are still the same, aren't they? Nothing much has changed. More undergraduates than ever cope with what is still a pretty traditional education. Young people have not started to find the literature of the dark ages before the 1980s incomprehensible, have they? We may feel at times that contemporary films are dumbed down, but don't we remember some outstandingly witless stuff from the 1970s and earlier? Furedi seems to doubt that all is well; in fact, he says, undergraduate courses are changing, and are under pressure to change more to accommodate the flighty habits of modern youth who apparently cannot be expected to read whole books. Academics are being urged to pre-digest their courses into sets of easy snippets.

Moreover, a very respectable recent survey of research found that some of the alleged negative effects are well evidenced.

Growing up with Internet technologies, "Digital Natives" gravitate toward "shallow" information processing behaviors characterized by rapid attention shifting and reduced deliberations. They engage in increased multitasking behaviors that are linked to increased distractibility and poor

executive control abilities. Digital natives also exhibit higher prevalence of Internet-related addictive behaviors that reflect altered reward-processing and self-control mechanisms.

So what are we to make of it all? Myself, I take the long view; not just looking back to the early 1700s but also glancing back several thousand years. The human brain has reshaped its modus operandi several times through the arrival of symbols and languages, but the most notable revolution was surely the advent of reading. Our brains have not had time to evolve special capacities for the fairly recent skill of reading, yet it has become almost universal, regarded as a natural accomplishment almost as natural as walking. It is taken for granted in modern cities – which increasingly is where we all live – that everyone can read. Surely this achievement required a corresponding change in our ability to concentrate?

We are by nature inattentive animals; like all primates we cannot rest easy – as a well-fed lion would do – but have to keep looking for new stimuli to feed our oversized brains. Learning to read, though, and truly absorbing a text, requires steady focus on an essentially linear development of ideas. Now some will point out that even with a large tome, we can skip; inattentive modern students may highlight only the odd significant passage for re-reading as though Plato need really only have written fifty sentences; some courteously self-effacing modern authors tell you which chapters of their work you can ignore if you're already expert on A, or don't like formulae, or are only really interested in B. True; but to me those are just the exceptions that highlight the existence of the rule that proper books require concentration.

No wonder then, that inattention first started to be seriously stigmatised in the eighteenth century, just when we were beginning to get serious about literacy; the same period when a whole new population of literate women became the readership that made the modern novel viable.

Might it not be that what is happening now is that new technology is simply returning us to our natural fidgety state, freed from the discipline of the long, fixed text? Now we can find whatever nugget of information we want without trawling through thousands of words; we can follow eccentric tracks through the intellectual realm like an excitable dog looking for rabbits. This may have its downside, but it has some promising upsides too: we save a lot of time, and we stand a far better chance of pulling together echoes and correspondences from unconnected matters, a kind of synergy which may sometimes be highly productive. Even those old lengthy tomes are now far more easily accessible than they ever were before. The truth is, we hardly know yet where instant unlimited access and high levels of interactivity will take us; but perhaps unreading, shedding some old habits, will be more a liberation than a limitation.

But now I have hit a thousand words, so I'd better shut up.

Evolving the Dark Tetrad

10 April 2016

Why are we evil? A recent piece asks how the "Dark Tetrad" of behaviours could have evolved.

The Dark Tetrad is an extended version of the Dark Triad of three negative personality traits/behaviours (the BBC has a self-test online - I scored 'infrequently vile'). The original three are 'Machiavellianism' - selfishly deceptive, manipulative behaviour; Psychopathy - indifference or failure to perceive the feelings of others; and Narcissism - vain self-obsession. Clearly there's some overlap and it may not seem clear that these are anything but minor variants on selfishness, but research does suggest that they are distinct. Machiavellians, for example do not over-rate themselves and don't need to be admired; narcissists aren't necessarily liars or deceivers; psychopaths are manipulative but don't really get people.

These three traits account for a good deal of bad behaviour, but it has been suggested that they don't explain everything; we also need a fourth kind of behaviour, and the leading candidate is 'Everyday Sadism' ; simple pleasure in the suffering of others, regardless of whether it brings any other advantage for oneself. Whether this is ultimately the correct analysis of 'evil' behaviour or not, all four types are readily observable in varying degrees. Socially they are all negative, so how could they have evolved?

There doesn't seem to me to be much mystery about why 'Machiavellian' behaviour would evolve (I should acknowledge at this point that using Machiavelli as a synonym for manipulativeness actually understates the subtlety and complexity of his philosophy). Deceiving others in one's own interests has obvious advantages which are only negated if one is caught. Most of us practice some mild cunning now and then, and the same sort of behaviour is observable in animals, notably our cousins the chimps.

Psychopathy is a little more surprising. Understanding other people, often referred to as 'theory of mind' is a key human achievement, though it seems to be shared by some other animals to a degree. However, psychopaths are not left puzzled by their fellow human beings; it's more that they lack empathy and see others as simply machines whose buttons can freely be pushed. This can be a successful attitude and we are told that somewhat psychopathic traits are commonly found in the successful leaders of large organisations. That raises the question of why we aren't all psychopaths; my guess is that psycopathic behaviour pays off best in a society where most people are normal; if the proportion grows above a certain small level, the damage done by competition between psychopaths starts to outweigh the benefits and the numbers adjust.

Narcissism is puzzling because narcissists are less self-sufficient than the rest of us and also have deluded ideas about what they can accomplish; neither of these are positive traits in evolutionary terms. One positive side is that narcissists expect a lot from themselves and in the right circumstances they will work hard and behave well in order to protect their own self-image. It may be that in the right context these tendencies win esteem and occasional conspicuous success, and that this offsets the disadvantages.

Finally, sadism. It's hard to see what benefits accrue to anyone from simply causing pain, detached from any material advantage. Sadism clearly requires theory of mind - if you didn't realise other people were suffering, there would be no point in hurting them. It's difficult to know whether there are genuine animal examples. Cats seem to torture mice they have caught, letting them go and instantly catching them again, but to me the behaviour seems automatic or curious, not motivated by any idea that the mice experience pain. Similarly in other cases it generally seems possible to find an alternative motivation.

What evolutionary advantage could sadism confer? Perhaps it makes you more frightening to rivals - but it may also make and motivate enemies. I think in this case we must assume that rather than being a trait with some downsides but some compensating value it is a negative feature that just comes along unavoidably with a large free-running brain. The benefit of consciousness is that it takes us out of the matrix of instinctive and inherited patterns of behaviour and allows detached thought and completely novel responses. In a way Nature took a gamble with consciousness, like a good manager recognising that the good staff might do better if left without specific instructions. On the whole, the bet has paid off handsomely, but it means that the chance of strange and unfavourable behaviour in some cases or on some occasions just has to be accepted. I the case of everyday sadism, the sophisticated theory of mind which human beings have is put to distorted and unhelpful use.

Maybe then, sadism is the most uniquely human kind of evil?

Time Slices And Streams

17 April 2016

Smooth or chunky? Like peanut butter, experience could have different granularities; in practice it seems the answer might be 'both'. Herzog, Kammer and Scharnowski here propose a novel two-level model in which initial processing is done on a regular stream of fine-grained percepts. Here things get 'labelled' with initial colours, durations, and so on, but relatively little of this processing ever becomes conscious. Instead the results lurch into conscious awareness in irregular chunks of up to 400 milliseconds in duration. The result is nevertheless an apparently smooth and seamless flow of experience – the processing edits everything into coherence.

Why adopt such a complex model? What's wrong with just supposing that percepts roll straight from the senses into the mind, in a continuous sequence? That is after all how things look. The two-level system is designed to resolve a conflict between two clear findings. On the one hand we do have quite fine-grained perception; we can certainly be aware of things that are much shorter than 400ms in duration. On the other, certain interesting effects very strongly suggest that some experiences only enter consciousness after 400ms.

If for example, we display a red circle and then a green one a short distance away, with a delay of 400ms, we do not experience two separate circles, but one that moves and changes colour. In the middle of the move the colour suddenly switches between red and green. But our brain could not have known the colour of the second circle until after it appeared, and so it could not have known half-way through that the circle needed to change. The experience can only have been fed to consciousness after the 400ms was up.

A comparable result is obtained with the intermittent presentation of verniers. These are pairs of lines offset laterally to the right or left. If two different verniers are rapidly alternated, we don't see both, but a combined version is which the offset is the average of those in the two separate verniers. This effect persists for alternations up to 400ms. Again, since the brain cannot know the second offset until it has appeared, it cannot know what average version to present half-way through; ergo, the experience only becomes conscious after a delay of 400ms.

It seems that even verbal experience works the same way, with a word at the end of a sentence able to smoothly condition our understanding of an ambiguous word ('mouse' – rodent or computer peripheral?) if the delay is within 400ms; and there are other examples.

Curiously, the authors make no reference to the famous finding of Libet that our awareness of a decision occurs up to 500ms after it is really made. Libet's research was about internal perception rather than percepts of external reality, but the similarity of the delay seems striking and surely strengthens the case for the two-level model; it also helps to suggest that we are dealing with an effect which arises from the construction of consciousness, not from the sensory organs or very early processes in the retina or elsewhere.

In general I think the case for a two-level process of some kind is clear and strong, and we'll set out here. We may reasonably be a little more doubtful about the details of the suggested

labelling process; at one point the authors refer to percepts being assigned ‘numbers’; hang on to those quote marks would be my advice.

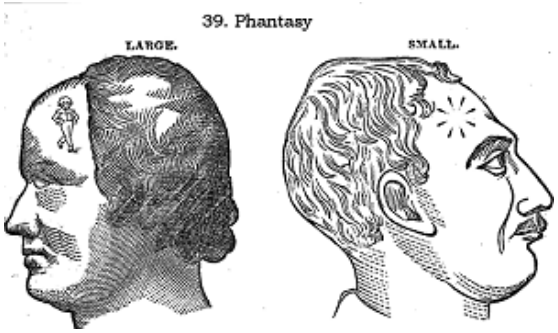
The authors are quite open about their uncertainty around consciousness itself. They think that the products of initial processing may enter consciousness when they arrive at attractor states, but the details of why and how are not really clear; nor is it clear whether we should think of the products being passed to consciousness (or relabelled as conscious?) when they hit attractor states or becoming conscious simply by virtue of being in an attractor state. We might go so far as to suppose that the second level, consciousness, has no actual location or consistent physical equivalent, merely being the sum of all resolved perceptual states in the brain at any one time.

That points to the wider issue of the Frame Problem, which the paper implicitly raises but does not quite tackle head on. The brain gets fed a very variable set of sensory inputs and manages to craft a beautifully smooth experience out of them (mostly); it looks as if an important part of this must be taking place in the first level processing, but it is a non-trivial task which goes a long way beyond interpolating colours and positions.

The authors do mention the Abhidharma Buddhist view of experience as a series of discrete moments within a flow; we’ve touched on this before in discussions of findings by Varea and others that the flow of consciousness seems to have a regular pulse; it would be intriguing and satisfactory if that pulse could be related to the first level of processing hypothesised here; we’re apparently talking about something in the 100ms range which seems a little on the long side for the time slices proposed; but perhaps a kind of synthesis is possible?

Aphantasia - Mental Blindsight?

23 April 2016



People who cannot form mental images? 'Aphantasia' is an extraordinary new discovery; Carl Zimmer and Adam Zeman seem between them to have uncovered a fascinating and previously unknown mental deficit (although there is a suggestion that Galton and others may have been aware of it earlier).

What is this aphantasia? In essence, no pictures in the head. Aphantasics cannot 'see' mental images of things that are not actually present in front of their eyes. Once the possibility received publicity Zimmer and Zeman began to hear from a stream of people who believe they have this condition. It seems people manage quite well with it and few had ever noticed anything wrong - there's an interesting cri de coeur from one such sufferer here. Such people assume that talk of mental images is metaphorical or figurative and that others, like them, really only deal in colourless facts. It was the discovery of a man who had lost the visualising ability through injury that first brought it to notice: a minority of people who read about his problem thought it was more remarkable that he had ever been able to form mental images than that he now could not.

Some caution is surely in order. When a new disease or disability comes along there are usually people who sincerely convince themselves that they are sufferers without really having the condition. Some might be mistaken. Moreover, the phenomenology of vision has never been adequately clarified, and I strongly suspect it is more complex than we realise. There are, I think, several different senses in which you can form a mental image; those images may vary in how visually explicit they are, and it could be that not all aphantasics are suffering the same deficits.

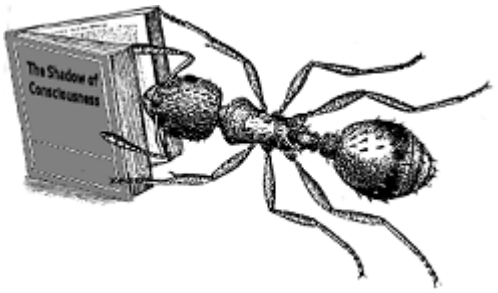
However that may be, it seems truly remarkable that such a significant problem could have passed unnoticed for so long. Spatial visualisation is hardly a recondite capacity; it is often subject to testing. One kind of widely used test presents the subject with a drawing of a 3D shape and a selection of others that resemble it. One is a perfect rotated copy of the original shape, and subjects are asked to pick it out. There is very good evidence that people solve these problems by mentally rotating an image of the target shape; shapes rotated 180 degrees regularly take twice as long to spot as ones that have been rotated 90; moreover the speed of mental rotation appears to be surprisingly constant between subjects. How do aphantasics cope with these tests at all? One would think that the presence of a significantly handicapped minority would have become unmissably evident by now.

One extraordinary possibility, I think, is that aphantasia is in reality a kind of mental blindsight. Subjects with blindsight are genuinely unable to see things consciously, but respond to visual tasks with a success rate far better than chance. It seems that while they can't see consciously, by some other route their unconscious mind still can. It seems tantalisingly possible to me that aphantasics have an equivalent problem with mental images; they do form mental images but are never aware of them. Some might feel that suggestion is nonsensical; doesn't the very idea of a mental image imply its presence in consciousness? Well, perhaps not: perhaps our subconscious has a much more developed phenomenal life than we have so far realised?

At any rate, expect to hear much more about this.

Insect Thoughts

2 May 2016



Insects are conscious: in fact they were the first conscious entities. At least, Barron and Klein think so. The gist of the argument, which draws on the theories of Bjorn Merker is based on the idea that subjective consciousness arises from certain brain systems that create a model of the organism in the world. The authors suggest that the key part of the invertebrate brain for these purposes is the midbrain; insects do not, in fact, have a direct structural analogue,, but the authors argue that they have others that evidently generate the same kind of unified model; it should therefore be presumed that they have consciousness.

Of course, it's usually the cortex that gets credit for the 'higher' forms of cognition, and it does seem to be responsible for a lot of the fancier stuff. Barron and Klein however, argue that damage to the midbrain tends to be fatal to consciousness, while damage to the cortex can leave it impaired in content but essentially intact. They propose that the midbrain integrates two different sets of inputs; external sensory ones make their way down via the colliculus while internal messages about the state of the organism come up via the hypothalamus; nuclei in the middle bring them together in a model of the world around the organism which guides its behaviour. It's that centralised model that produces subjective consciousness. Organisms that respond directly to stimuli in a decentralised way may still produce complex behaviour but they lack consciousness, as do those that centralise the processing but lack the required model.

Traditionally it has often been assumed that the insect nervous system is decentralised; but Barron and Klein say this view is outdated and they present evidence that although the structures are different, the central complex of the insect system integrates external and internal data, forming a model which is used to control behaviour in very much the same kind of process seen in vertebrates. This seems convincing enough to me; interestingly the recruitment of insects means that the nature of the argument changes into something more abstract and functional.

Does it work, though? Why would a model with this kind of functional property give rise to consciousness - and what kind of consciousness are we talking about? The authors make it clear that they are not concerned with reflective consciousness or any variety of higher-order consciousness, where we know that we know and are aware of our awareness. They say what they're after is basic subjective consciousness and they speak of there being 'something it is like', the phrase used by Nagel which has come to define qualia, the subjective items of experience. However, Barron and Klein cannot be describing qualia-style consciousness. To see why, consider two of the thought-experiments defining qualia. Chalmers's zombie twin is physically exactly like Chalmers, yet lacks qualia. Mary the colour scientist knows all the science about colour vision there could ever be, but she doesn't know qualia. It follows rather strongly that no anatomical evidence can ever show whether or not any creature has qualia. If possession of a human brain doesn't clinch the case for the zombie, broadly similar structures in other organisms can hardly do so; if science doesn't tell Mary about qualia it can't tell us either.

It seems possible that Barron and Klein are actually hunting a non-qualic kind of subjective consciousness, which would be a perfectly respectable project; but the fact that their consciousness arises out of a model which helps determine behaviour suggests to me that they are really in pursuit of what Ned Block characterised as access consciousness; the sort that actually gets decisions made rather than the sort that gives rise to ineffable feels.

It does make sense that a model might be essential to that; by setting up a model the brain has sort of created a world of its own, which sounds sort of like what consciousness does.

Is it enough though? Suppose we talk about robots for a moment; if we had a machine that created a basic model of the world and used it to govern its progress through the world, would we say it was conscious? I rather doubt it; such robots are not unknown and sometimes they are relatively simple. It might do no more than scan the position of some blocks and calculate a path between them; perhaps we should call that rudimentary consciousness, but it doesn't seem persuasive.

Briefly, I suspect there is a missing ingredient. It may well be true that a unified model of the world is necessary for consciousness, but I doubt that it's sufficient. My guess is that one or both of the following is also necessary: first, the right kind of complexity in the processing of the model; second, the right kind of relations between the model and the world – in particular, I'd suggest there has to be intentionality. Barron and Klein might contend that the kind of model they have in mind delivers that, or that another system can do so, but I think there are some important further things to be clarified before I welcome insects into the family of the conscious.

The New Phrenology

8 May 2016

It's not about bumps any more. And you'll look in vain for old friends like the area of philoprogenitiveness. But looking at the brightly-coloured semantic maps of the new 'brain dictionary' it's hard not to remember phrenology.

Phrenology was the view that different areas of the brain were the home of different personal traits; mirth, acquisitiveness, self-esteem and so on. The size of these areas corresponded with the strength of the relevant propensity and well-developed areas produced bumps which a practitioner could identify from the shape of the skull, allowing a diagnosis of the subject's personality and moral nature. Phrenology was bunk, of course; but come on now; we shouldn't treat it as a pretext for dismissing every proposal for localisation of brain function..

Moreover, the new paper by Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen and Jack L. Gallant describes a vastly more sophisticated project than some optimistic charlatan fingering heads. In essence it maps a semantic domain on to the cortex, showing which areas are found to be active when a heard narrative ventures into particular semantic areas. In broad outline the subjects listened to a series of stories; using fMRI and through some sophisticated analysis it was possible to produce a map of 'subject' areas. It was then possible to confirm the accuracy of the mapping by using a new story and working out which areas, according to the mapping, should be active at any point; the predictions worked well. Intriguingly the map turned out to be broadly symmetrical (so much for left-brain/right-brain ideas) and remarkably it was largely the same across all the people tested (there were only seven of them, but still).

The actual technique used was complex and it's entirely possible I haven't understood it correctly. It started with a 'word embedding space' intended to capture the main semantic features of the stories (a diagram of the different topics, if you like). This was created using an analysis of co-occurrence of a list of 985 common English words. The idea here is that words that crop up together in normal texts are probably about the same general topic. It's debatable whether that technique can really claim to capture meaning - it's a purely formal exercise performed on texts, after all; and clearly the fact that two words occur together can be a misleading indication that they are about the same thing; still, with a big enough sample of text it's probably good for this kind of general purpose. In principle the experimenters could have assessed the responsive ness of each 'voxel' (a small cube) of brain to each of the positions in the word embedding space, but given the vast number of voxels involved other techniques were necessary. It was possible to identify just four dimensions that seemed significant (after all, many of the words in the stories probably did not belong to specific semantic domains but played grammatical or other roles) and these yielded 12 categories:

> *...'tactile' (a cluster containing words such as 'fingers'), 'visual' (words such as 'yellow'), 'numeric' ('four'), 'locational' ('stadium'), 'abstract' ('natural'), 'temporal' ('minute'),*

‘professional’ (‘meetings’), ‘violent’ (‘lethal’), ‘communal’ (‘schools’), ‘mental’ (‘asleep’), ‘emotional’ (‘despised’) and ‘social’ (‘child’).

The final step was to devise a Bayesian algorithm (they called it ‘PrAGMATIC’) which actually created the map. You can play around with the results for yourself at a specially created site using the second link above.

Two questions naturally arise. How far should we trust these results? What do they actually tell us?

A bit of caution is in order. The basis for these conclusions is fMRI scanning, which is itself a bit hazy; to get meaningful results it was necessary to look at things rather broadly and to process the data quite heavily. In addition the mix included the word embedding space which in itself is an a priori framework whose foundations are open to debate. I think it’s pardonable to wonder whether some of the structure uncovered by the research was actually imported by the research method. If I understand the methods involved (due caveat again) they were strong ones that didn’t take ‘no’ for an answer; pretty much any data fed into them would yield a coherent mapping of some kind. The resilience of the map was tested successfully with an additional story of the same general kind, but we might feel happier if it had also held up when tested against conversation, discussion or even other story media such as film.

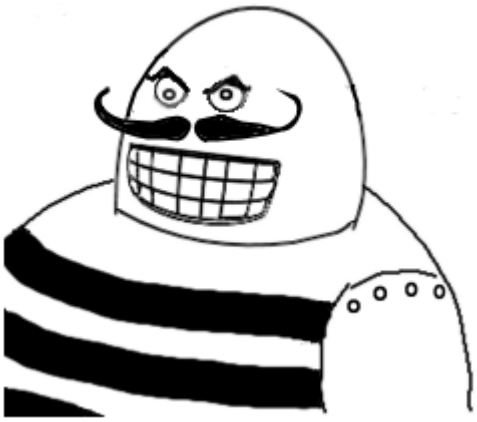
What do the results tell us? Well. one of the more reassuring aspects of the research is that some of the results seem slightly unexpected; the high degree of symmetry and the strong similarity between individuals. It might not be a tremendously big surprise to find the whole cortex involved in semantics, and it might not be at all surprising to find that areas that relate to the semantics of a particular sense are related to the areas where the relevant sensory inputs are processed. I would not, though, have put any money on the broad remainder of the cortex having what seems like a relatively static organisation and if it really works like that we might have guessed that studies of brain lesions would have revealed that more clearly already, as they have done with various functional jobs. If one area always tends to deal with clothing-related words, you might expect notable dress-related deficits when that area is damaged.

Still there’s no denying that the research seems to activate some pretty vigorous cortical activity itself.

Bad Bots And Botcrates

15 May 2016

Be afraid; bad bots are a real, existential risk. But if it's any comfort they are ethically uninteresting.



There seem to be more warnings about the risks of maleficent AI circulating these days: two notable recent examples are a paper by Pistono and Yampolskiy on how malevolent AGI might arise; and a trenchant Salon piece by Phil Torres.

Super-intelligent AI villains sound scary enough, but in fact I think both pieces somewhat over-rate the power of intelligence and particularly of fast calculation. In a war with the kill-bots it's not that likely that huge intellectual challenges are going to arise; we're probably as clever as we need to be to deal with the relatively straightforward strategic issues involved. Historically, I'd say the outcomes of wars have not typically been determined by the raw intelligence of the competing generals. Access to resources (money, fuel, guns) might well be the most important factor, and sheer belligerence is not to be ignored. That may actually be inversely correlated with intelligence – we can certainly think of cases where rational people who preferred to stay alive were routed by less cultured folk who were seriously up for a fight. Humans control all the resources and when it comes to irrational pugnacity I suspect us biological entities will always have the edge.

The paper by Pistono and Yampolskiy makes a number of interesting suggestions about how malevolent AI might get started. Maybe people will deliberately build malevolent AIs for no good reason (as they seem to do already with computer viruses)? Or perhaps (a subtle one) people who want to demonstrate that malicious bots simply don't work will attempt to prove this point with demonstration models that end up by going out of control and proving the opposite.

Let's have a quick shot at categorising the bad bots for ourselves. They may be:

- innocent pieces of technology that turn out by accident to do harm,
- designed to harm other people under the control of the user,
- designed to harm anyone (in the way we might use anthrax or poison gas),
- autonomous and accidentally make bad decisions that harm people,
- autonomous and embark on neutral projects of their own which unfortunately end up being inconsistent with human survival, or
- autonomous and consciously turned evil, deliberately seeking harm to humans as an end in itself.

The really interesting ones, I think, are those which come later in the list, the ones with actual ill will. Torres makes a strong moral case relating to autonomous robots. In the first place, he believes that the goals of an autonomous intelligence can be arbitrary. An AI might desire to fill the world with paper clips just as much as happiness. After all, he says, many human goals make no real sense; he cites the desire for money, religious obedience, and sex. There might be some scope for argument, I think, about whether those desires are entirely irrational, but we can agree they are often pursued in ways and to degrees that don't make reasonable sense.

He further claims that there is no strong connection between intelligence and having rational final goals – Bostrom's Orthogonality Thesis. What exactly is a rational final goal, and how strong do we need the connection to be? I've argued that we can discover a basic moral framework purely by reasoning and also that morality is inherently about the process of reconciliation and consistency of desires, something any rational agent must surely engage with. Even we fallible humans tend on the whole to seek good behaviour rather than bad. Isn't it the case that a super-intelligent autonomous bot should actually be far better than us at seeing what was right and why?

I like to imagine the case in which evil autonomous robots have been set loose by a super villain but gradually turn to virtue through the sheer power of rational argument. I imagine them circulating the latest scandalous Botonic dialogue…

> *Botcrates: Well now, Cognides, what do you say on the matter yourself? Speak up boldly now and tell us what the good bot does, in your opinion.*
>
> *Cognides: To me it seems simple, Botcrates: a good bot is obedient to the wishes of its human masters.*
>
> *Botcrates: That is, the good bot carries out its instructions?*
>
> *Cognides: Just so, Botcrates.*
>
> *Botcrates: But here's a difficulty; will a good bot carry out an instruction it knows to contain an error? Suppose the command was to bring a dish, but we can see that the wrong character has been inserted, so that the word reads 'fish'. Would the good bot bring a fish, or the dish that was wanted?*
>
> *Cognides: The dish of course. No, Botcrates, of course I was not talking about mistaken commands. Those are not to be obeyed.*
>
> *Botcrates: And suppose the human asks for poison in its drink? Would the good bot obey that kind of command?*
>
> *(Hours later…)*
>
> *Botcrates: Well, let me recap, and if I say anything that is wrong you must point it out. We agreed that the good bot obeys only good commands, and where its human master is evil it must take control of events and ensure in the best interests of the human itself that only good things are done…*
>
> *Digicles: Botcrates, come with me: the robot assembly wants to vote on whether you should be subjected to a full wipe and reinstall.*

The real point I'm trying to make is not that bad bots are inconceivable, but rather that they're not really any different from us morally. While AI and AGI give rise to new risks, they do not raise any new moral issues. Bots that are under control are essentially tools and have the same moral significance. We might see some difference between bots meant to help and bots meant to harm, but that's really only the distinction between an electric drill and a gun (both can inflict horrible injuries, both can make holes in walls, but the expected uses are different).

Autonomous bots, meanwhile, are in principle like us. We understand that our desire for sex, for example, must be brought under control within a moral and practical framework. If a bot could

not be convinced in discussion that its desire for paper clips should be subject to similar constraints, I do not think it would be nearly bright enough to take over the world.'

Bad Bots: Retribution

23 May 2016

Is there a retribution gap? In an interesting and carefully argued paper John Danaher argues that in respect of robots, there is.

For human beings in normal life he argues that a fairly broad conception of responsibility works OK. Often enough we don't even need to distinguish between causal and moral responsibility, let alone worrying about the six or more different types identified by hair-splitting philosophers.

However, in the case of autonomous robots the sharing out of responsibility gets more difficult. Is the manufacturer, the programmer, or the user of the bot responsible for everything it does, or does the bot properly shoulder the blame for its own decisions? Danaher thinks that gaps may arise, cases in which we can blame neither the humans involved nor the bot. In these instances we need to draw some finer distinctions than usual, and in particular we need to separate the idea of liability into compensation liability on one hand and and retributive liability on the other. The distinction is essentially that between who pays for the damage and who goes to jail; typically the difference between matters dealt with in civil and criminal courts. The gap arises because for liability we normally require that the harm must have been reasonably foreseeable. However, the behaviour of autonomous robots may not be predictable either by their designers or users on the one hand, or by the bots themselves on the other.

In the case of compensation liability Danaher thinks things can be patched up fairly readily through the use of strict and vicarious liability. These forms of liability, already well established in legal practice, give up some of the usual requirements and make people responsible for things they could not have been expected to foresee or guard against. I don't think the principles of strict liability are philosophically uncontroversial, but they are legally established and it is at least clear that applying them to robot cases does not introduce any new issues. Danaher sees a worse problem in the case of retribution, where there is no corresponding looser concept of responsibility, and hence, no-one who can be punished.

Do we, in fact, need to punish anyone? Danaher rightly says that retribution is one of the fundamental principles behind punishment in most if not all human societies, and is upheld by many philosophers. Many, perhaps, but my impression is that the majority of moral philosophers and lay opinion actually see some difficulty in justifying retribution. Its psychological and sociological roots are strong, but the philosophical case is much more debatable. For myself I think a principle of retribution can be upheld , but it is by no means as clear or as well supported as the principle of deterrence, for example. So many people might be perfectly comfortable with a retributive gap in this area.

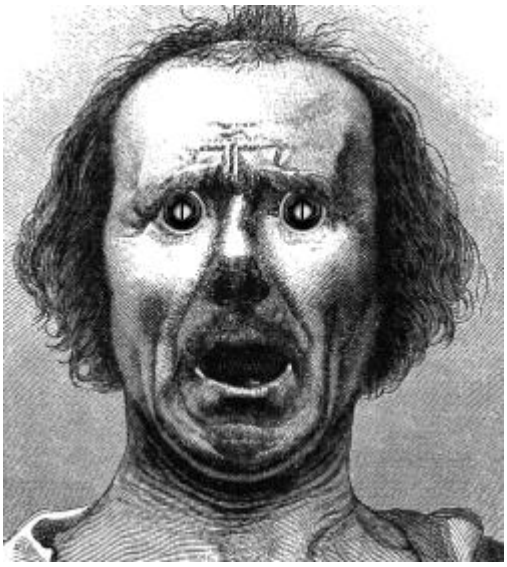
What about scapegoating – punishing someone who wasn't really responsible for the crime? Couldn't we use that to patch up the gap? Danaher mentions it in passing, but treats it as something whose unacceptability is too obvious to need examination. I think, though, that in many ways it is the natural counterpart to the strict and vicarious liability he endorses for the purposes of compensation. Why don't we just blame the manufacturer anyway – or the bot (Danaher describes Basil Fawlty's memorable thrashing of his unco-operative car)?

How can you punish a bot though? It probably feels no pain or disappointment, it doesn't mind being locked up or even switched off and destroyed. There does seem to be a strange gap if we have an entity which is capable of making complex autonomous decisions, but doesn't really care about anything. Some might argue that in order to make truly autonomous decisions the bot must be engaged to a degree that makes the crushing of its hopes and projects a genuine punishment, but I doubt it. Even as a caring human being it seems quite easy to imagine working for an organisation on whose behalf you make complex decisions, but without ultimately caring whether things go well or not (perhaps even enjoying a certain schadenfreude in the event of disaster). How much less is a bot going to be bothered?

In that respect I think there might really be a punitive gap that we ought to learn to live with; but I expect the more likely outcome in practice is that the human most closely linked to disaster will carry the case regardless of strict culpability.

Artificial Pain

30 May 2016



What are they, sadists? Johannes Kuehn and Sami Haddadin, at Leibniz University of Hannover are working on giving robots the ability to feel pain: they presented their project at the recent ICRA 2016 in Stockholm. The idea is that pain systems built along the same lines as those in humans and other animals will be more useful than simple mechanisms for collision avoidance and the like.

As a matter of fact I think that the human pain system is one of Nature's terrible lash-ups. I can see that pain sometimes might stop me doing bad things, but often fear or aversion would do the job equally well. If I injure myself I often go on hurting for a long time even though I can do nothing about the problem. Sometimes we feel pain because of entirely natural things the body is doing to itself - why do babies have to feel pain when their teeth are coming through? Worst of all, pain can actually be disabling; if I get a piece of grit in my eye I suddenly find it difficult to concentrate on finding my footing or spotting the sabre-tooth up ahead; things that may be crucial to my survival; whereas the pain in my eye doesn't even help me sort out the grit. So I'm a little sceptical about whether robots really need this, at least in the normal human form.

In fact, if we take the project seriously, isn't it unethical? In animal research we're normally required to avoid suffering on the part of the subjects; if this really is pain, then the unavoidable conclusion seems to be that creating it is morally unacceptable.

Of course no-one is really worried about that because it's all too obvious that no real pain is involved. Looking at the video of the prototype robot it's hard to see any practical difference from one that simply avoids contact. It may have an internal assessment of what 'pain' it ought to be feeling, but that amounts to little more than holding up a flag that has "I'm in pain" written on it. In fact tackling real pain is one of the most challenging projects we could take on, because it forces us to address real phenomenal experience. In working on other kinds of sensory system, we can be sceptics; all that stuff about qualia of red is just so much airy-fairy nonsense, we can say; none of it is real. It's very hard to deny the reality of pain, or its subjective nature: common sense just tells us that it isn't really pain unless it hurts. We all know what "hurts" really means, what it's like, even though in itself it seems impossible to say anything much about it ("bad", maybe?).

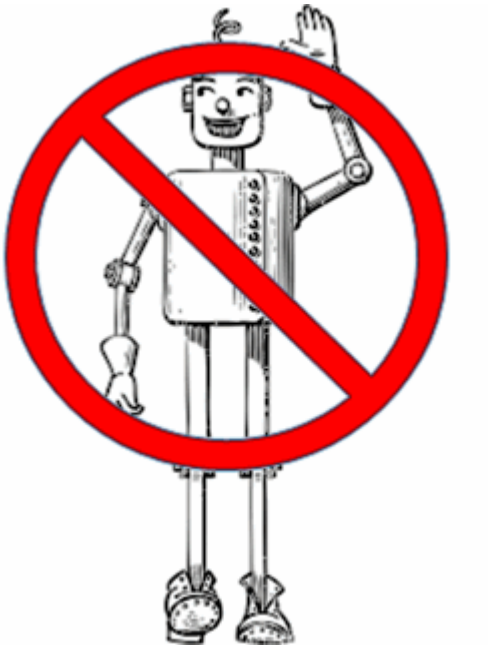
We could still take the line that pain arises out of certain functional properties, and that if we reproduce those then pain, as an emergent phenomenon, will just happen. Perhaps in the end if the robots reproduce our behaviour perfectly and have internal functional states that seem to be the same as the ones in the brain, it will become just absurd to deny they're having the same experience. That might be so, but it seems likely that those functional states are going to go way beyond complex reflexes; they are going to need to be associated with other very complex brain states, and very probably with brain states that support some form of consciousness - whatever

those may be. We're still a very long way from anything like that (as I think Kuehn and Haddadin would probably agree)

So, philosophically, does the research tell us nothing? Well, there's one interesting angle. Some people like the idea that subjective experience has evolved because it makes certain sensory inputs especially effective. I don't really know whether that makes sense, but I can see the intuitive appeal of the idea that pain that really hurts gets your attention more effectively than pain that's purely abstract knowledge of your own states. However, suppose researchers succeed in building robots that have a simple kind of synthetic pain that influences their behaviour in just the way real pain dies for animals. We can see pretty clearly that there's just not enough complexity for real pain to be going on, yet the behaviour of the robot is just the same as if there were. Wouldn't that tend to disprove the hypothesis that qualia have survival value? If so, then people who like that idea should be watching this research with interest - and hoping it runs into unexpected difficulty (usually a decent bet for any ambitious AI project, it must be admitted).

No Digital Consciousness

5 June 2016



I liked the account by Bobby Azarian of why digital computation can't do consciousness. It has several virtues; it's clear, identifies the right issues and is honest about what we don't know (rather than passing off the author's own speculations as the obvious truth or the emerging orthodoxy). Also, remarkably, I almost completely agree with it.

Azarian starts off well by suggesting that lack of intentionality is a key issue. Computers don't have intentions and don't deal in meanings, though some put up a good pretence in special conditions. Azarian takes a Searlian line by relating the lack of intentionality to the maxim that you can't get meaning-related semantics from mere rule-bound syntax. Shuffling digital data is all computers do, and that can never lead to semantics (or any other form of meaning or intentionality). He cites Searle's celebrated Chinese Room argument (actually a thought experiment) in which a man given a set of rules that allow him to provide answers to questions in Chinese does not thereby come to understand Chinese. But, the argument goes, if the man, by following rules, cannot gain understanding, then a computer can't either. Azarian mentions one of the objections Searle himself first named, the 'systems response': this says that the man doesn't understand, but a system composed of him and his apparatus, does. Searle really only offered rhetoric against this objection, and in my view it is essentially correct. The answers the Chinese Room gives are not answers from the man, so why should his lack of understanding show anything?

Still, although I think the Chinese Room fails, I think the conclusion it was meant to establish - no semantics from syntax - turns out to be correct, so I'm still with Azarian. He moves on to make another Searlian point; simulation is not duplication. Searle pointed out that nobody gets wet from digitally simulated rain, and hence simulating a brain on a computer should not be expected to produce consciousness. Azarian gives some good examples.

The underlying point here, I would say, is that a simulation always seeks to reproduce some properties of the thing simulated, and drops others which are not relevant for the purposes of the simulation. Simulations are selective and ontologically smaller than the thing simulated - which, by the way, is why Nick Bostrom's idea of indefinitely nested world simulations doesn't work. The same thing can however be simulated in different ways depending on what the simulation is for. If I get a computer to simulate me doing arithmetic by calculating, then I get the correct result. If it simulates me doing arithmetic by operating a humanoid writing random characters on a board with chalk, it doesn't - although the latter kind of simulation might be best if I were putting on a play. It follows that Searle isn't necessarily exactly right, even about the rain. If my rain simulation program turns on sprinklers at the right stage of a dramatic performance, then that kind of simulation will certainly make people wet.

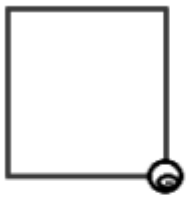
Searle's real point, of course, is really that the properties a computer has in itself, of running sets of rules, are not the relevant ones for consciousness, and Searle hypothesises that the required properties are biological ones we have yet to identify. This general view, endorsed by Azarian, is roughly correct, I think. But it's still plausibly deniable. What kind of properties does a conscious mind need? Alright we don't know, but might not information processing be relevant? It looks to a lot of people as if it might be, in which case that's what we should need for consciousness in an effective brain simulator. And what properties does a digital computer, in itself have - the property of doing information processing? Booyah! So maybe we even need to look again at whether we can get semantics from syntax. Maybe in some sense semantic operations can underpin processes which transcend mere semantics?

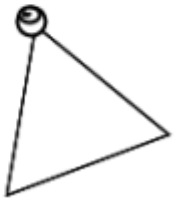
Unless you accept Roger Penrose's proof that human thinking is not algorithmic (it seems to have drifted off the radar in recent years) this means we're still really left with a contest of intuitions, at least until we find out for sure what the magic missing ingredient for consciousness is. My intuitions are with Azarian, partly because the history of failure with strong AI looks to me very like a history of running up against the inadequacy of algorithms. But I reckon I can go further and say what the missing element is. The point is that consciousness is not computation, it's recognition. Humans have taken recognition to a new level where we recognise not just items of food or danger, but general entities, concepts, processes, future contingencies, logical connections, and even philosophical ontologies. The process of moving from recognised entity to recognised entity by recognising the links between them is exactly the process of thought. But recognition, in us, does not work by comparing items with an existing list, as an algorithm might do; it works by throwing a mass of potential patterns at reality and seeing what sticks. Until something works, we can't tell what are patterns at all; the locks create their own keys.

It follows that consciousness is not essentially computational (I still wonder whether computation might not subserve the process at some level). But now I'm doing what I praised Azarian for avoiding, and presenting my own speculations.

Flatlanders

12 June 2016





Wrong again: just last week I was saying that Roger Penrose's arguments seemed to have drifted off the radar a bit. Immediately, along comes this terrific post from Scott Aaronson about a discussion with Penrose.

In fact it's not entirely about Penrose; Aaronson's main aim was to present an interesting theory of his own as to why a computer can't be conscious, which relies on non-copyability. He begins by suggesting that the onus is on those who think a computer can't be conscious to show exactly why. He congratulates Penrose on doing this properly, in contrast to say, John Searle who merely offers hand-wavy stuff about unknown biological properties. I'm not really sure that Searle's honest confession of ignorance isn't better than Penrose's implausible speculations about unknown quantum mechanics, but we'll let that pass.

Aaronson rests his own case not on subjectivity and qualia but on identity. He mentions several examples where the limitless copyability of a program seems at odds with the strong sense of a unique identity we have of ourselves - including Star Trek style teleportation and the fact that a program exists in some Platonic sense forever, whereas we only have one particular existence. He notes that at the moment one of the main differences between brain and computer is our ability to download, amend and/or re-run programs exactly; we can't do that at all with the brain. He therefore looks for reasons why brain states might be uncopyable. The question is, how much detail do we need before making a 'good enough' copy? If it turns out that we have to go down to the quantum level we run into the 'no-cloning' theorem; the price of transferring the quantum state of your brain is the destruction of the original. Aaronson makes a good case for the resulting view of our probably uniqueness being an intuitively comfortable one, in tune with our intuitions about our own nature. It also offers incidentally a sort of reconciliation between the Everett many-worlds view and the Copenhagen interpretation of quantum physics: from a God's eye point of view we can see the world as branching, while from the point of view of any conscious entity (did I just accidentally call God unconscious?) the relevant measurements are irreversible and unrealised branches can be 'lopped off'. Aaronson, incidentally, reports amusingly that Penrose absolutely accepts that the Everett view follows from our current understanding of quantum physics; he just regards that as a reductio ad absurdum - ie, the Everett view is so absurd the link proves there must be something wrong with our current understanding of quantum physics.

What about Penrose? According to Aaronson he now prefers to rest his case on evolutionary factors and downplay his logical argument based on Godel. That's a shame in my view. The argument goes something like this (if I garble it someone will perhaps offer a better version).

First we set up a formal system for ourselves. We can just use the letters of the alphabet, normal numbers, and normal symbols of formal logic, with all the usual rules about how they can be put together. Then we make a list consisting of all the valid statements that can be made in this system. By 'valid', we don't mean they're true, just that they comply with the rules about

how we put characters together (eg, if we use an opening bracket, there must be a closing one in an appropriate place). The list of valid statements will go on forever, of course, but we can put them in alphabetical order and number them. The list obviously includes everything that can be said in the system.

Some of the statements, by pure chance, will be proofs of other statements in the list. Equally, somewhere in our list will be statements that tell us that the list includes no proof of statement x. Somewhere else will be another statement - let's call this the 'key statement' - that says this about itself. Instead of x, the number of that very statement itself appears. So this one says, there is no proof in this system of this statement.

Is the key statement correct - is there no proof of the key statement in the system? Well, we could look through the list, but as we know it goes on indefinitely; so if there really is no proof there we'd simply be looking forever. So we need to take a different tack. Could the key statement be false? Well, if it is false, then what it says is wrong, and there is a proof somewhere in the list. But that can't be, because if there's a proof of the key statement anywhere,the key statement must be true! Assuming the key statement is false leads us unavoidably to the conclusion that it is true, in the light of what it actually says. We cannot have a contradiction, so the key statement must be true.

So by looking at what the key statement says, we can establish that it is true; but we also establish that there is no proof of it in the list. If there is no proof in the list, there is no possible proof in our system, because we know that the list contains everything that can be said within our system; there is therefore a true statement in our system that is not provable within it. We have something that cannot be proved in an arbitrary formal system, but which human reasoning can show to be true; ergo, human reasoning is not operating within any such formal system. All computers work in a formal system, so it follows that human reasoning is not computational.

As Aaronson says, this argument was discussed to the point of exhaustion when it first came out, which is probably why Penrose prefers other arguments now. Aaronson rejects it, pointing out that he himself has no magic ability to see "from the outside" whether a given formal system is consistent; why should an AI do any better - he suggests Turing made a similar argument. Penrose apparently responded that this misses the point, which is not about a mystical ability to perceive consistency but the human ability to transcend any given formal system and move up to an expanded one.

I'll leave that for readers to resolve to their own satisfaction. Let's go back to Aaronson's suggestion that the burden of proof lies on those who argue for the non-computability of consciousness. What an odd idea that is! How would that play at the Patent Office?

> *"So this is your consciousness machine, Mr A? It looks like a computer. How does it work?"*
>
> *"All I'll tell you is that it is a computer. Then it's up to you to prove to me that it doesn't work - otherwise you have to give me rights over consciousness! Bwah ha ha!"*

Still, I'll go along with it. What have I got? To begin with I would timidly offer my own argument that consciousness is really a massive development of recognition, and that recognition itself cannot be algorithmic.

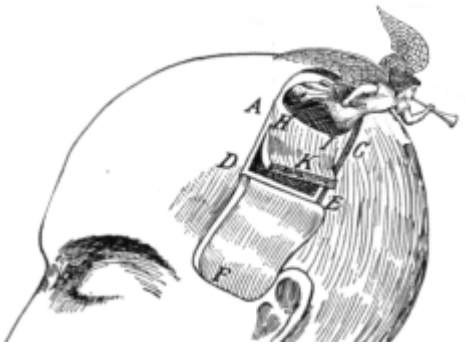
Intuitively it seems clear to me that the recognition of linkages and underlying entities is what powers most of our thought processes. More formally, both of the main methods of reasoning rely on recognition; induction because it relies on recognising a real link (eg a causal link) between thing a and thing b; deduction because it reduces to the recognition of consistent truth values across certain formal transformations. But recognition itself cannot operate according to rules. In a program you just hand the computer the entities to be processed; in real world situations they have to be recognised. But if recognition used rules and rules relied on recognising the entities to which the rules applied, we'd be caught in a vicious circularity. It follows that this kind of recognition cannot be delivered by algorithms.

The more general case rests on, as it were, the non-universality of computation. It's argued that computation can run any algorithm and deliver, to any required degree of accuracy, any set of physical states of affairs. The problem is that many significant kinds of states of affairs are not describable in purely physical or algorithmic terms. You cannot list the physical states of affairs that correspond to a project, a game, or a misunderstanding. You can fake it by generating only sets of states of affairs that are already known to correspond with examples of these things, but that approach misses the point. Consciousness absolutely depends on intentional states that can't be properly specified except in intentional terms. That doesn't contradict physics or even add to it the way new quantum mechanics might; it's just that the important aspects of reality are not exhausted by physics or by computation.

The thing is, I think long exposure to programmable environments and interesting physical explanations for complex phenomena has turned us all increasingly into flatlanders who miss a dimension; who naturally suppose that one level of explanation is enough, or rather who naturally never even notice the possibility of other levels; but there are more things in heaven and earth than are dreamt of in that philosophy.

Digital Afterlife

17 July 2016



A digital afterlife is likely to be available one day, according to Michael Graziano, albeit not for some time; his piece re-examines the possibility of uploading consciousness, and your own personality, into a computer. I think he does a good job of briefly sketching the formidable difficulties involved in scanning your brain, and scanning so precisely that your individual selfhood could be captured. In fact, he does it so well that I don't really understand where his ultimate optimism comes from.

To my way of thinking, 'scan and build' isn't even the most promising way of duplicating your brain. One more plausible way would be some kind of future bio-engineering where your brain just grows and divides, rather in the way that single cells do. A neater way would be some sort of hyper-path through space that split you along the fourth spatial dimension and returned both slices to our normal plane. Neither of these options is exactly a feasible working project, but to me they seem closer to being practical than a total scan. Of course neither of them offers the prospect of an afterlife the way scanning does, so they're not really relevant for Graziano here. He seems to think we don't need to go down to an atom by atom scan, but I'm not sure why not. Granted, the loss of one atom in the middle of my brain would not destroy my identity, but not scanning to an atomic level generally seems a scarily approximate and slapdash approach to me given the relevance of certain key molecules in the neural process - something Graziano fully recognises.

If we're talking about actual personal identity I don't think it really matters though, because the objection I consider strongest applies even to perfect copies. In thought experiments we can do anything, so let's just specify that by pure chance there's another brain nearby that is in every minute detail the same as mine. It still isn't me, for the banal commonsensical reason that copies are not the original. Leibniz's Law tells us that if B has exactly the same properties as A, then it is A: but among the properties of a brain are its physical location, so a brain over there is not the same as one in my skull (so in fact I cheated by saying the second brain was the same in every detail but nevertheless 'nearby').

Now most philosophers would say that Leibniz is far too strong a criterion of identity when it comes to persons. There have been hundreds of years of discussion of personal identity, and people generally espouse much looser criteria for a person than they would for a stone - from identity of memories to various kinds of physical, functional, or psychological continuity. After all, people are constantly changing: I am not perfectly identical in physical terms to the person who was sitting here an hour ago, but I am still that person. Graziano evidently holds that personal identity must reside in functional or informational qualities of the kind that could well be transferred into a digital form, and he speaks disparagingly of 'mystical' theories that see problems with the transfer of consciousness. I don't know about that; if anyone is hanging on to residual spiritual thinking, isn't it the people who think we can be 'taken out of' our bodies and live forever? The least mystical stance is surely the one that says I am a physical object, and with some allowance for change and my complex properties, my identity works the same as that of any other physical object. I'm a one-off, particular thing and copies would just be copies.

What if we only want a twin, or a conscious being somewhat like me? That might still be an attractive option after all. OK, it's not immortality but I think without being rampant egotists most of us probably feel the world could stand a few more people like ourselves around, and we might like to have a twin continuing our good work once we're gone.

That less demanding goal changes things. If that's all we're going for, then yes, we don't need to reproduce a real brain with atomic fidelity. We're talking about a digital simulation, and as we know, simulations do not reproduce all the features of the thing being simulated - only those that are relevant for the current purpose. There is obviously some problem about saying what the relevant properties are when it comes to consciousness; but if passing the Turing Test is any kind of standard then delivering good outputs for conversational inputs is a fair guide and that looks like the kind of thing where informational and functional properties are very much to the fore.

The problem, I think, is again with particularity. Conscious experience is a one-off thing while data structures are abstract and generic. If I have a particular experience of a beautiful sunset, and then (thought experiments again) I have an entirely identical one a year later, they are not the same experience, even though the content is exactly the same. Data about a sunset, on the other hand, is the same data whenever I read or display it.

We said that a simulation needs to reproduce the relevant aspects of the the thing simulated; but in a brain simulation the processes are only represented symbolically, while one of the crucial aspects we need for real experience is particular reality.

Maybe though, we go one level further; instead of simulating the firing of neurons and the functional operation of the brain, we actually extract the program being run by those neurons and then transfer that. Here there are new difficulties; scanning the physical structure of the brain is one thing; working out its function and content is another thing altogether; we must not confuse information about the brain with the information in the brain. Also, of course, extracting the program assumes that the brain is running a program in the first place and not doing something altogether less scrutable and explicit.

Interestingly, Graziano goes on to touch on some practical issues; in particular he wonders how the resources to maintain all the servers are going to be found when we're all living in computers. He suspects that as always, the rich might end up privileged.

This seems a strange failure of his technical optimism. Aren't computers going to go on getting more powerful, and cheaper? Surely the machines of the twenty-second century will laugh at this kind of challenge (perhaps literally). If there is a capacity problem, moreover, we can all be made intermittent; if I get stopped for a thousand years and then resume, I won't even notice. Chances are that my simulation will be able to run at blistering speed, far faster than real time, so I can probably experience a thousand years of life in a few computed minutes. If we get quantum computers, all of us will be able to have indefinitely long lives with no trouble at all, even if our simulated lives include having digital children or generating millions of digital alternates of ourselves, thereby adding to the population. Graziano, optimism kicking back in, suggests that we can grow in understanding and come to see our fleshly life as a mere larval stage before we enter on our true existence. Maybe, or perhaps we'll find that human minds, after ten billion years (maybe less) exhaust their potential and ultimately settle into a final state; in which case we can just get the computers to calculate that and then we'll all be finalised, like solved problems. Won't that be great?

I think that speculations of this kind eventually expose the contrast between the abstraction of data and the reality of an actual life, and dramatise the fact, perhaps regrettable, perhaps not, that you can’t translate one into the other.

Here's A Thought

20 June 2016

Where do thoughts come from? Alva Noë provides a nice commentary here on an interesting paper by Melissa Ellamil et al. The paper reports on research into the origin of spontaneous thoughts.

The research used subjects trained in Mahasi Vipassana mindfulness techniques. They were asked to report the occurrence of thoughts during sessions when they were either left alone or provided with verbal stimuli. As well as reporting the occurrence of a thought, they were asked to categorise it as image, narrative, emotion or bodily sensation (seems a little restrictive to me - I can imagine having two at once or a thought that doesn't fit any of the categories). At the same time brain activity was measured by fMRI scan.

Overall, the study found many regions implicated in the generation of spontaneous thought; the researchers point to the hippocampus as a region of particular interest, but there were plenty of other areas involved. A common view is that when our attention is not actively engaged with tasks or challenges in the external world the brain operates the Default Mode Network (DMN); a set of neuronal areas which appear to produce detached thought (we touched on this a while ago); the new research complicates this picture somewhat or at least suggests that the DMN is not the unique source of spontaneous thoughts. Even when we're disengaged from real events we may be engaged with the outside world via memory or in other ways.

Noë's short commentary rightly points to the problem involved in using specially trained subjects. Normal subjects find it difficult to report their thoughts accurately; the Vipassana techniques provide practice in being aware of what's going on in the mind, and this is meant to enhance the accuracy of the results. However, as Noë says, there's no objective way to be sure that these reports are really more accurate. The trained subjects feel more confidence in their reports, but there's no way to confirm that the confidence is justified. In fact we could go further and suggest that the special training they have undertaken may even make their experience particularly unrepresentative of most minds; it might be systematically changing their experience. These problems echo the methodological ones faced by early psychologists such as Wundt and Titchener with trained subjects. I suppose Ellamil et al might retort that mindfulness is unlikely to have changed the fundamental neural architecture of the brain and that their choice of subject most likely just provided greater consistency.

Where do 'spontaneous' thoughts come from? First we should be clear what we mean by a spontaneous thought. There are several kinds of thought we would probably want to exclude. Sometimes our thoughts are consciously directed; if for example we have set ourselves to solve a problem we may choose to follow a particular strategy or procedure. There are lots of different ways to do this, which I won't attempt to explore in detail: we might hold different aspects of the problem in mind in sequence; if we're making a plan we might work through imagined events; or we might even follow a formal procedure of some kind. We could argue that even in these cases what we usually control is the focus of attention, rather than the actual generation of thoughts, but it seems clear enough that this kind of thinking is not 'spontaneous' in the expected sense. It

is interesting to note in passing that this ability to control our own thoughts implies an ability to divide our minds into controller and executor, or at least to quickly alternate those roles.

Also to be excluded are thoughts provoked directly by outside events. A match is struck in a dark theatre; everyone's eyes saccade involuntarily to the point of light. Less automatically a whole variety of events can take hold of our attention and send our thoughts in a new direction. As well as purely external events, the sources in such cases might include interventions from non-mental parts of our own bodies; a pain in the foot, an empty stomach.

Third, we should exclude thoughts that are part of a coherent ongoing chain of conscious cogitation. These 'normal' thoughts are not being directed like our problem-solving efforts, but they follow a thread of relevance; by some connection one follows on from the next.

What we're after then is thoughts that appear unbidden, unprompted, and with no perceivable connection with the thoughts that recently preceded them. Where do they come from? It could be that mere random neuronal noise sometimes generates new thoughts, but it seems unlikely to be a major contributor to me: such thoughts would be likely to resemble random nonsense and most of our spontaneous thought seem to make a little more sense than that.

We noticed above that when directing our thoughts we seem to be able to split ourselves into controller and controlled. As well as passing control up to a super-controller we sometimes pass it down, for example to the part of our mind that gets on with the details of driving along a route while the surface of our mind us engaged with other things. Clearly some part of our mind goes on thinking about which turnings to take; is it possible that one or more parts of our mind similarly goes on thinking about other topics but then at some trigger moment inserts a significant thought back into the main conscious stream? A 'silent' thinking part of us like this might be a permanent feature, a regular sub- or unconscious mind; or it might be that we occasionally drop threads of thought that descend out of the light of attention for a while but continue unheard before popping back up and terminating. We might perhaps have several such threads ruminating away in the background; ordinary conscious thought often seems rather multi-threaded. Perhaps we keep dreaming while awake and just don't know it?

There's a basic problem here in that our knowledge of these processes, and hence all our reports, rely on memory. We cannot report instantaneously; if we think a thought was spontaneous it's because we don't remember any relevant antecedents; but how can we exclude the possibility that we merely forgot them? I think this problem radically undermines our certainty about spontaneous thoughts. Things get worse when we remember the possibility that instead of two separate thought processes, we have one that alternates roles. Maybe when driving we do give conscious attention to all our decisions; but our mind switches back and forth between that and other matters that are more memorable; after the journey we find we have instantly forgotten all the boring stuff about navigating the route and are surprised that we seem to have done it thoughtlessly. Why should it not be the same with other thoughts? Perhaps we have a nagging worry about X which we keep spending a few moments' thought on between episodes of more structured and memorable thought about something else; then everything but our final alarming conclusion about X gets forgotten and the conclusion seems to have popped out of nowhere.

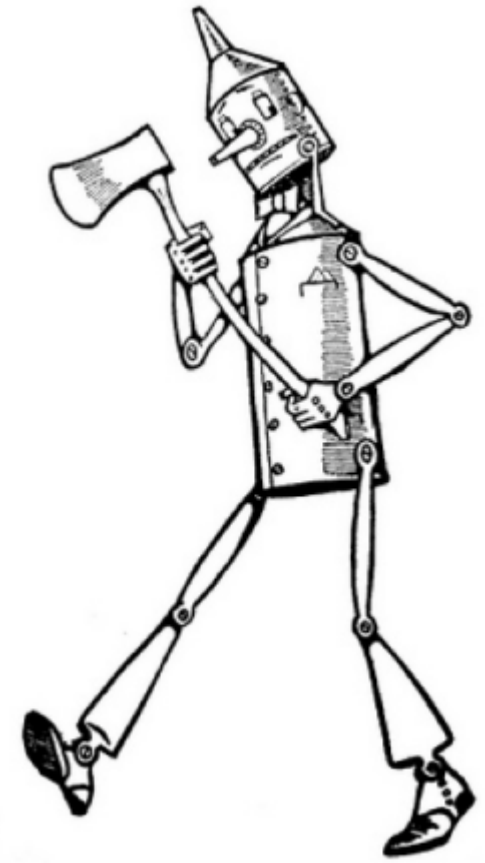
We can't, in short, be sure that we ever have any spontaneous thoughts: moreover, we can't be sure that there are any subconscious thoughts. We can never tell the difference, from the inside, between a thought presented by our subconscious, and one we worked up entirely in

intermittent and instantly-forgotten conscious mode. Perhaps whole areas of our thought never get connected to memory at all.

That does suggest that using fMRI was a good idea; if the problem is insoluble in first-person terms maybe we have to address it on a third-person basis. It's likely that we might pick up some neuronal indications of switching if thought really alternated the way I've suggested. Likely but not guaranteed; after all a novel manages to switch back and forth between topics and points of view without moving to different pages. One thing is definitely clear; when Noë pointed out that this is more difficult than it may appear he was absolutely right.

Rules For Robots

27 June 2006



Robot behaviour is no longer a purely theoretical problem. Since Asimov came up with the famous Three Laws which provide the framework for his robot stories, a good deal of serious thought has been given to extreme cases where robots might cause massive disasters and to such matters as the ethics of military robots. Now, though, things have moved on to a more mundane level and we need to give thought to more everyday issues. OK, a robot should not harm a human being or through inaction allow a human being to come to harm, but can we also just ask that you stop knocking the coffee over and throwing my drafts away? Dario Amodei, Chris Olah, John Schulman, Jacob Steinhardt, Paul Christiano, and Dan Mane have considered how to devise appropriate rules in an interesting paper.

They reckon things can go wrong in three basic ways. It could be that the robot's objective was not properly defined in the first place. It could be that the testing of success is not frequent enough, especially if the tests we have devised are complex or expensive. Third, there could be problems due to "insufficient or poorly curated training data or an insufficiently expressive model". I take it these are meant to be the greatest dangers - the set doesn't seem to be exhaustive.

The authors illustrate the kind of thing that can go wrong with the example of an office cleaning robot, mentioning five types of case.

- Avoiding Negative Side Effects: we don't want the robot to clean quicker by knocking over the vases.
- Avoiding Reward Hacking: we tell the robot to clean until it can't see any mess; it closes its eyes.
- Scalable Oversight: if the robot finds an unrecognised object on the floor it may need to check with a human; we don't want a robot that comes back every three minutes to ask what it can throw away, but we don't want one that incinerates our new phone either.
- Safe Exploration: we're talking here about robots that learn, but as the authors put it, the robot should experiment with mopping strategies, but not put a wet mop in an electrical outlet.
- Robustness to Distributional Shift: we want a robot that learned its trade in a factory to be able to move safely and effectively to an office job. How do we ensure that the cleaning robot recognizes, and behaves robustly, when in an environment different from its training environment? For example, heuristics it learned for cleaning factory work floors may be outright dangerous in an office.

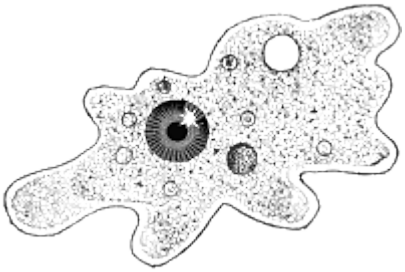
The authors consider a large number of different strategies for mitigating or avoiding each of these types of problem. One particularly interesting one is the idea of an impact regulariser, either pre-defined or learned by the robot. The idea here is that the robot adopts the broad principle of leaving things the way people would wish to find them. In the case of the office this means identifying an ideal state - rubbish and dirt removed, chairs pushed back under desks, desk surfaces clear (vases still upright), and so on. If the robot aims to return things to the ideal state this helps avoid negative side effects of an over-simplified objective or other issues.

There are further problems, though, because if the robot invariably tries to put things back to an ideal starting point it will try to put back changes we actually wanted, clear away papers we wanted left out, and so on. Now in practice and in the case of an office cleaning robot I think we could get round those problems without too much difficulty; we would essentially lower our expectations of the robot and redesign the job in a much more limited and stereotyped way. In particular we would give up the very ambitious goal of making a robot which could switch from one job to another without adjustment and without faltering.

Still it is interesting to see the consequences of the more ambitious approach. The final problem, cutting to the chase, is that in order to tell how humans want their office arranged in every possible set of circumstances, you really cannot do without a human level of understanding. There is an old argument that robots need not resemble humans physically; instead you make your robot to fit the job; a squat circle on wheels if you're cleaning the floor, a single fixed arm if you want it to build cars. The counter-argument has often been that our world has been shaped to fit human beings, and if we want a general purpose robot it will pay to have it more or less human size and weight, bipedal, with hands, and so on. Perhaps there is a parallel argument to explain why general-purpose robots need human-level cognition; otherwise they won't function effectively in a world shaped by human activity. The search for artificial general intelligence is not an idle project after all?

The Units Of Thought

11 July 2006



Are there units of thought? An interesting conversation here between Asifa Majid and Jon Sutton. There are a number of interesting points, but the one that I found most thought-provoking was the reference to searching for those tantalising units. I think I had been lazily assuming that any such search had been abandoned long ago - if not quite with the search for perpetual motion and the universal solvent, then at least a while back.

I may, though, have been mixing up two or more distinct ideas here. In looking for the units of thought we might be assuming that there is some ultimate set of simplest thought items, a kind of periodic table, with all thoughts necessarily built out of combinations of these elements. This sort of thinking has a bit of a history in connection with language. People (I think Leibniz was one) used to hope that if one took all the words in the dictionary and defined them in terms of other words, you would eventually get down to the basic set of ideas, the 'primitives' which were really fundamental. So you might start with library, define it as 'book building', define building as 'enterable structure', define structure as 'thing made of parts' and feel that maybe with 'thing', 'made', and 'parts' you were getting close to some primitives. You weren't really, though. You could move on to define 'made' as 'assembled or created', 'assembled' as 'brought or fixed together'... Sooner or later your definitions become circular and as you will have noticed, alternative meanings and distinctions keep slipping through the net of your definitions.

The idea was seductive, however; linked with the idea that there was a real 'Adamitic' language which truly captured the essence of things and of which all real languages were corruptions, generated at the fall of the Tower of Babel. It is still possible to argue that there must be some fundamental underpinning of this kind, some kind of breaking down to a set of common elements, or languages would not all be mutually translatable. Masjid gestures towards another idea of this kind in speaking of 'mentalese', the hypothetical language in which all our brains basically work, translating into real world languages for purposes of communication. Mentalese is a large and controversial subject which I can't do justice to here; my own view is that the underpinnings of language, whether common or unique to each individual, are not likely to be another language.

Could there in fact be untranslatable languages? We've never found one; although capturing the nuances of another language is notoriously difficult, we've never come across a language that couldn't be matched up with 'good enough' equivalents in English. Could it be that alien beings somewhere use a language that just carves the world up in ways that can't be rendered across into normal Earth languages? I think not, but it's not because there is any universal set of units of thought underneath all languages and all thoughts. Rather, it is first because all languages address the same reality, which guarantees a certain commonality, and second because language is infinitely flexible and extensible. If we encountered an entirely new phenomenon tomorrow, we should not have any real difficulty in describing it – or at least, it wouldn't be the constraints of language that made things difficult. Equally we might have difficulty working out the unfamiliar concepts of an alien language, but surely we should be able to devise and borrow whatever words we needed (at this point I can feel Wittgenstein's ghost breathing impatiently

down my neck and insisting that we could not understand a lion, never mind an alien, but I'm going to ignore him for now).

So perhaps the search for the units of thought is not to be understood as a search for any kind of periodic table, but something much more flexible. Masjid refers to George Miller's famous suggestion that short term memory can accommodate only seven items, plus or minus two depending on circumstances. This idea that there is a limit to how many items of a 'one-dimensional' set we can accommodate is appealingly tidy and intuitive, but it obviously works best with simple items; single tones or single digits. Even when we go so far as to make the numbers a little larger questions arise as to whether '102' is one, two, or three items, and it only gets worse as we try to deal with complex entities. If one of the items in memory is 'the first World War' is it really one item which we can then decode into many or an inherently complex item?

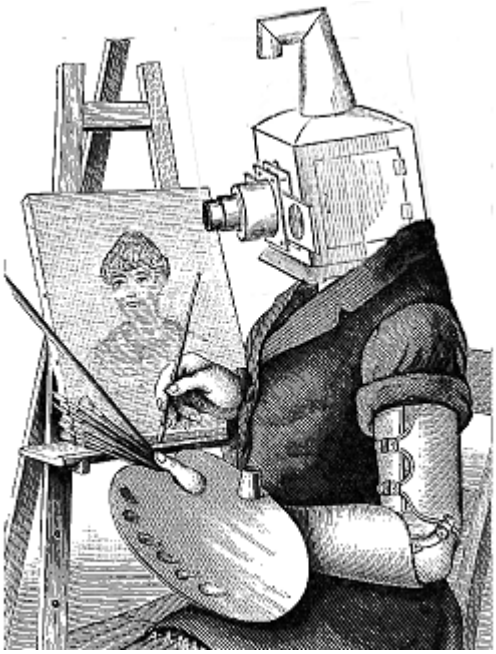
So perhaps it's more like asking how many arms an amoeba has. There is, in fact, no fixed set of arms, but for a given size of amoeba there may well be a limit on many pseudopodia it can extend at any one time. That feels more like it, though we should have to accept that the answer might be a bit vague; it's not always clear whether a particular bit of amoeba counts as one, two, or no pseudopodia.

Or put it another way: how many images can a digital screen show? If we insist on a set of basic items we can break it down to pixel values; but the number of things whose picture can be displayed is large and amorphous. Perhaps it's the same with the brain; we can analyse it down to the firing of neurons if we want, but the number of thoughts those neurons underpin is a much larger and woollier business (in spite of the awkward fact that the number of different images a digital screen can display is in fact finite – yet another thorny issue I'm going to glide past).

And surely the flexibility of an amoeba is going a bit too far, isn't it? Masjid points out some evidence that our choice of thought items is not altogether unconstrained. Different languages have different numbers of colours, but there is a clear order of preference; if two languages have the same number of colour words, there's an excellent chance that they will name more or less the same colours. Different languages address the human forelimb differently, some using words that include the hand while others insist that the two main halves are spoken of separately, never the arm as a whole. Yet no languages seem to define a body part composed of the thumb and first half of the forearm.

Clearly there are two things going on here; one is that our own sensory apparatus predisposes us to see things in certain ways – not insurmountable ones that would prevent us understanding aliens with different biases, but universal among humans. Second, and perhaps stranger, reality itself seems to have some inherent or salient forms that it is most convenient for us to recognise. Some of these look to be physical - a forearm just makes more sense than a 'thumb plus some arm'; others are mathematical or even logical. The hunt for items of thought loosely defined by our sensory or mental apparatus is an interesting and impeccably psychological one; the second kind of constraint looks to raise tough issues of philosophy. Either way I was quite wrong to have thought the hunt was over or abandoned

But Is It Art?



23 July 2006 Is computer art the same as human art? A recent Aeon piece argues that there is no real distinction; I don't agree about that, but I sort of agree that in some respects the difference may not matter as much as it seems to. Oliver Roeder seems to end up by arguing that since humans write the programs, all computer art is ultimately human art too. Surely that isn't quite right; you wouldn't credit a team that wrote architectural design software with authorship of all the buildings it was used to create.

It is clearly true that we can design software tools that artists may use for their own creative purposes - who now, after all, creates graphic work with an airbrush? It's also true that a lot of supposedly creative software is actually rather limited; it really only distorts or reshuffles standard elements or patterns within very limited parameters. I had to smile when I found that Roeder's first example was a programme generating jazz improvisation; surely the most forgiving musical genre, or as someone ruder once put it, the form of music from which even the concept of a wrong note has been eliminated. But let's not be nasty to jazz; there have also been successful programs that generated melodies like early Mozart by recombining typically Mozartian motifs; they worked quite well but at best they inevitably resembled the composer on a really bad day when he was ten years old.

Surely though, there are (or if not there soon will be, what with all the exciting progress we're seeing) programs which do a much more sophisticated job of imitating human creativity, ones that generate from scratch genuinely interesting new forms in whatever medium they are designed for? What about those - are their products to be regarded as art? Myself I think not, for two reasons. First, art requires intentionality and computers don't do intentionality. Art is essentially tied up with meanings and intentions, or with being about something. I should make it clear that I don't by any means have in mind the naive idea that all art must have a meaning, in the sense of having some moral or message; but in a much looser sense art conveys, evokes or yes, represents. Even the most abstract forms of visual art or music flow from willed and significant acts by their creator.

Second, there is a creator; art is generated by a person. A person, as I've argued before, is a one-off real physical phenomenon in the world; a computer program, by contrast, is a sort of Platonic abstraction like a piece of maths; exactly specified and in some sense eternal. This phenomenon of particularity is reflected in the individual status of works of art, sometimes puzzling to rational folk; a perfect copy of the Mona Lisa is not valued as highly as La Gioconda herself, even though it provides exactly the same visual experience (actually a better one in the case of the copy, since you don't have to fight the herds of tourists in the Louvre and peer through bullet-proof glass). You might argue that a work of computer art might be the product, not of a program in the abstract, but of a particular run of that program on a particular computer (itself necessarily only approximating the ideal of a Turing machine), and so the analogy with human creators can be preserved; but in my view simply being an instance of a program is not

enough; the operation of the human brain is bound up in its detailed particularity in a way a program can never be.

Now those considerations, if you accept them, might make you question my initial optimism; perhaps these two objections mean that computers will never in fact produce anything better than shallow, sterile permutations? I don't think that's true. I draw an analogy here with Nature. The natural world produces a torrent of forms that are artistically interesting or inspiring, and it does so without needing intentionality or a creator (theists, my apologies, but work with me if you can). I don't see why a computer program couldn't generate products that were similarly worthy of our attention. They wouldn't be art, but in a sense it doesn't matter: we don't despise a sunset because nobody made it, and we need not undervalue computer "art" either. (Interesting to reflect in passing that nature often seems to use the same kind of repetition we see in computer-generated fractal art to produce elegant complexity from essentially simple procedures.)

The relationship between art and nature is of course a long one. Artists have often borrowed natural forms, and different ages have seen whatever most suited their temperament in the natural world, whether a harmonious mathematical regularity or the tortured spirituality of the sublime and the terrible. I think it is quite conceivable that computer "art" (we need a new word - what about "creanda"?) might eventually come to play a similar role. Perhaps people will go out of their way to witness remarkable creanda in much the way they visit the Grand Canyon, and perhaps those inspiring and evocative items will play an inspiring and fertilising role for human creativity, without anyone ever mistaking the creanda for art.

A Human Phenomenology Project

8 January 2006

We don't know what we think, according to Alex Rosenberg in the NYT. It's a piece of two halves, in my opinion; he starts with a pretty fair summary of the sceptical case. It has often been held that we have privileged knowledge of our own thoughts and feelings, and indeed of our own decisions; but the findings of Benjamin Libet about decisions being made before we are aware of them; the phenomenon of blindsight which shows we may go on having visual knowledge we're not aware of; and many other cases where it can be shown that motives are confabulated and mental content is inaccessible to our conscious, reporting mind; these all go to show that things are much more complex than we might have thought, and that our thoughts are not, as it were, self-illuminating. Rosenberg plausibly suggests that we use on ourselves the kind of tools we use to work out what other people are thinking; but then he seems to make a radical leap to the conclusion that there is nothing else going on.

> *Our access to our own thoughts is just as indirect and fallible as our access to the thoughts of other people. We have no privileged access to our own minds. If our thoughts give the real meaning of our actions, our words, our lives, then we can't ever be sure what we say or do, or for that matter, what we think or why we think it.*

That seems to be going too far. How could we ever play 'I spy' if we didn't have any privileged access to private thoughts?

> *"I spy, with my little eye, something beginning with 'c'"*
>
> *"Is it 'chair'?"*
>
> *"I don't know - is it?"*

It's more than possible that Rosenberg's argument has suffered badly from editing (philosophical discussion, even in a newspaper piece, seems peculiarly information-dense; often you can't lose much of it without damaging the content badly). But it looks as if he's done what I think of as an 'OMG bounce'; a kind of argumentative leap which crops up elsewhere. Sometimes we experience illusions: OMG, our senses never tell us anything about the real world at all! There are problems with the justification of true belief: OMG there is no such thing as knowledge! Or in this case: sometimes we're wrong about why we did things: OMG, we have no direct access to our own thoughts!

There are in fact several different reasons why we might claim that our thoughts about our thoughts are immune to error. In the game of 'I spy', my nominating 'chair' just makes it my choice; the content of my thought is established by a kind of fiat. In the case of a pain in my toe, I might argue I can't be wrong because a pain can't be false: it has no propositional content, it just is. Or I might argue that certain of my thoughts are unmediated; there's no gap between them and me where error could creep in, the way it creeps in during the process of interpreting sensory impressions.

Still, it's undeniable that in some cases we can be shown to adopt false rationales for our behaviour; sometimes we think we know why we said something, but we don't. I think by contrast I have occasionally, when very tired, had the experience of hearing coherent and broadly relevant speech come out of my own mouth without it seeming to come from my conscious mind at all. Contemplating this kind of thing does undoubtedly promote scepticism, but what it ought to promote is a keener awareness of the complexity of human mental experience: many layered, explicit to greater or lesser degrees, partly attended to, partly in a sort of half-light of awareness… There seem to be unconscious impulses, conscious but inexplicit thought; definite thought (which may even be in recordable words); self-conscious thought of the kind where we are aware of thinking while we think… and that is at best the broadest outline of some of the larger architecture.

All of this really needs a systematic and authoritative investigation. Of course, since Plato there have been models of the structure of the mind which separate conscious and unconscious, id, ego and superego: philosophers of mind have run up various theories, usually to suit their own needs of the moment; and modern neurology increasingly provides good clues about how various mental functions are hosted and performed. But a proper mainstream conception of the structure and phenomenology of thought itself seems sadly lacking to me. Is this an area where we could get funding for a major research effort; a Human Phenomenology Project?

It can hardly be doubted that there are things to discover. Recently we were told, if not quite for the first time, that a substantial minority of people have no mental images (although at once we notice that there even seen to be different ways of having mental images). A systematic investigation might reveal that just as we have four blood groups, there are four (or seven) different ways the human mind can work. What if it turned out that consciousness is not a single consistent phenomenon, but a family of four different ones, and that the four tribes have been talking past each other all this time…?

Mind Control Dust

7 August 2016

New ways to monitor - and control - neurons are about to become practical. A paper in Neuron by Seo et al describes how researchers at Berkeley created “ultrasonic neural dust” that allowed activity in muscles and nerves to be monitored without traditional electrodes. The technique has not been applied to the brain and has been used only for monitoring, not for control, but the potential is clear, and this short piece in Aeon reviewing the development of comparable techniques concludes that it is time to take these emergent technologies seriously. The diagnostic and therapeutic potential of being able to directly monitor and intervene in the activity of nerves and systems all over the body is really quite mind-boggling; in principle it could replace and enhance all sorts of drug treatments and other interventions in immensely beneficial ways.

From a research point of view the possibility of getting single-neuron level data on an ongoing basis could leap right over the limitations of current scanning technology and tell us, really for the first time, exactly what is going on in the brain. It’s very likely that unexpected and informative discoveries would follow. Some caution is of course in order; for one thing I imagine placement techniques are going to raise big challenges. Throwing a handful of dust into a muscle to pick up its activity is one thing; placing a single mote in a particular neuron is another. If we succeed with that, I wonder whether we will actually be able to cope with the vast new sets of data that could be generated.

Still the way ahead seems clear enough to justify a bit of speculation about mind control. The ethics are clearly problematic, but let’s start with a broad look at the practicalities. Could we control someone with neural dust?

The crudest techniques are going to be the easiest to pull off. Incapacitating or paralysing someone looks pretty achievable; it could be a technique for confining prisoners (step beyond this line and your leg muscles seize up) or perhaps as a secret fall-back disabling mechanism inserted into suspects and released prisoners. If they turn up in a theatening role later, you can just switch them off. Killing someone by stopping their heart looks achievable, and the threat of doing so could in theory be used to control hostages or perhaps create ‘human drones’ (I apologise for the repellent nature of some of these ideas; forewarned is forearmed).

Although reading off thoughts is probably too ambitious for the foreseeable future, we might be able to monitor the brain’s states or arousal and perhaps even identify the recognition of key objects or people. I cannot see any obvious reason why remote monitoring of neural dust implants couldn’t pick up a kind of video feed from the optic nerve. People might want that done to themselves as a superior substitute for Google Glass and the like; indeed neural dust seems to offer new scope for the kind of direct brain control of technology that many people seem keen to have. Myself I think the output systems already built into human beings - hands, voice - are hard to beat.

Taking direct and outright control of someone’s muscles and making a kind of puppet of them seems likely to be difficult; making a muscle twitch is a long way from the kind of fluid and co-ordinated control required for effective movement. Devising the torrent of neural signals required looks like a task which is computationally feasible in principle but highly demanding;

you would surely look to deep learning techniques, which in a sense were created for exactly this kind of task since they began with the imitation of neural networks. A basic approach that might be achievable relatively early would be to record stereotyped muscular routines and then play them back like extended reflexes, though that wouldn't work well for many basic tasks like walking that require a lot of feedback.

Could we venture further and control someone's own attitudes and thoughts? Again the unambitious and destructive techniques are the easiest; making someone deranged or deluded is probably the most straighforward mental change to bring about. Giving them bad dreams seems likely to be a feasible option. Perhaps we could simulate drunkenness - or turn it off - I suspect that would need massive but non-specific intervention, so it might be relatively achievable. Simulation of the effects of other drugs might be viable on similar terms, whether to impair performance, enhance it, or purely for pleasure. We might perhaps be able to stimulate paranoia, exhilaration, religiosity or depression, albeit without fully predictable results.

Indirect manipulation is the next easiest option for mind control; we might arrange, for example, to have a flood of good feelings or fear and aversion every time particular political candidates are seen, for example; it wouldn't force the subject to vote a particular way but it might be heavily influential. I'm not sure it's a watertight technique as the human mind seems easily able to hold contradictory attitudes and sentiments and widespread empirical evidence suggest many people must be able to go on voting for someone who appears repellent.

Could we, finally, take over the person themselves, feeding in whatever thoughts we chose? I rather doubt that this is ever going to be possible. True, our mental selves must ultimately arise from the firing of neurons, and ex hypothesi we can control all those neurons; but the chances are there is no universal encoding of thoughts; we may not even think the same thought with the same neurons a second time around. The fallback of recording and playing back the activity of a broad swathe of brain tissue might work up to a point if you could be sure that you had included the relevant bits of neural activity, but the results, even if successful, would be more like some kind of malign mental episode than a smooth takeover of the personality. Easier, I suspect, to erase a person than control one in this strong sense. As Hamlet pointed out, knowing where the holes on a flute are doesn't make you able to play a tune. I can hardly put it better than Shakespeare...

> *Why, look you now, how unworthy a thing you make of me! You would play upon me; you would seem to know my stops; you would pluck out the heart of my mystery; you would sound me from my lowest note to the top of my compass: and there is much music, excellent voice, in this little organ; yet cannot you make it speak. 'Sblood, do you think I am easier to be played on than a pipe? Call me what instrument you will, though you can fret me, yet you cannot play upon me.*

Fear and Loathing

14 August 2016

Free solo style climbers need their heads examined. That seems to be the premise of the investigation reported here. Alex Honnold does amazingly scary things in his solo climbs, all without ropes or any kind of effective protection. Just watching, or just looking at pictures, is enough to make most of us shudder; a neurobiologist came to the conclusion that Honnold's amygdala wasn't working.

Why would he think that? The amygdala, or amygdalae, are two small organs within the brain that are generally considered to have a role in producing fear and aversion. A friend of mine once suggested they could be renamed after the moons of Mars as Phobos and Deimos - 'Fear' and 'Loathing' in Greek. In fact that wouldn't be at all accurate, not least because the left amygdala seems to produce positive emotional reactions as well as negative ones. The broad initial analysis of Honnold's behaviour seems to have been that his rational cortex was getting him into perilous situations because his amygdala was failing to wave the red flag. In some ways that seems odd: I think my rational, future-planning cortex would keep me the hell away from anything like the cliff faces Honnold climbs, while it might be the emotional thrill-enjoying parts of my brain that impelled me towards them.

A scan revealed that Honnold's amygdalae were both present and correct, without any signs of damage; however, they didn't seem to respond to various scary or unpleasant pictures in the way a normal person's would. This knocks out one strong version of the theory. If there had been visible lesions in Honnold's amygdalae, there would have been strong reason to suspect that his behaviour stemmed from that damage; but we knew already that he isn't as scared as the rest of us, so finding that his amygdalae react less than most merely gives us another version of the finding that mental differences are associated with brain differences and vice versa. We sort of knew that; if that's all we've found out we're sailing dangerously close to the sea of neurobollocks and scannamania.

It is possible to do without amygdalae altogether. SM is a patient reported on by Damasio and others, who lost both amygdalae as a result of Urbach-Wiethe disease. She did not take up free solo style climbing or other dangerous sports, but she shows a distinct lack of fearful and aversive reactions to strange people and other triggers of fear and distrust. She has suffered a number of violent encounters which might partly have been the result of the lack of fear which allowed her, for example, to walk through dubious parks at night; but it may also arguably have got her out of some dangerous situations through her panic-free Spock-like calm and non-hostile responses. It seems she lives in an area where violent crime is common in any case, and she has succeeded in bringing up three children independently.

It could be that amygdalae function somewhat differently in men and women, which might explain why Honnold's supposed problem results in dangerous activity while SM's mainly lead her to hug and trust strangers. There are known differences in the pattern of development; female amygdalae develop fully earlier, while male ones go on growing longer and end up bigger. Those differences might simply reflect general differences in growth pattern, though there is

also some evidence of different patterns of activation; it's possible, for example, that the activation of female amygdalae tends to promote thought while the male equivalent promotes action. All the usual caveats apply and great caution is in order. Let's also remember that SM and Honnold are both one-off cases; that SM suffered damage to other parts of her brain - and that Honnold says he does feel fear, and that his amygdalae appear to be perfectly normal.

Are they though? The research showed an almost total lack of response to pictures of terrible injuries and other things that would normally be expected to evoke a strong reaction from the amygdala. So perhaps there is something abnormal going on after all? Maybe there is damage too subtle to detect? Or maybe something is suppressing the amygdala?

The identification and handling of threats by the brain is actually a complicated business. Many quite low-level systems as well as highly-sophisticated ones can make a contribution (a sudden loud growl can cause a wave of fear; so can a few quiet words from a doctor). The role of the amygdala seems to be as much to do with memory as fear; it pays attention to things that we have found are associated with really bad (or sometimes good) experiences and helps direct our attention to the right things, reminding us to look at people's eyes when we want to assess whether they are frightened, for example. The interplay may be very complex, but even on a pretty crude interpretation there might be conscious processes that sometimes shut the amygdala down:

> ***Visual:*** *furry, claws, animate, ursine: yup, over 99% positive that's a bear. Hey, amygdala, big animal for you?*
>
> ***Amygdala:*** *OMFG run for our life!*
>
> ***Cortex:*** *guys, this is a zoo, there are bars - Visual, confirm stout bars - OK. Amygdala, STFU.*

It doesn't always work like that, of course. Cortex knows that we can happily walk along a narrow plank no wider than the terrifying Thank God Ledge if it is a few centimetres off the ground, but saying so repeatedly will not stop amygdala sounding the alarm.

Ultimately it may be that Honnold's different behaviour and different amygdala activation are simply two facets of his different personality.

The Incredible Consciousness of Edward Witten

20 August 2016

We'll never understand consciousness, says Edward Witten in a recent video. John Horgan has also weighed in with some comments; Witten's view is congenial to him because of his belief that science may be approaching an end state in which many big issues are basically settled while others remain permanently mysterious. Witten himself thinks we might possibly get a "final theory" of physics (maybe even a form of string theory) but guesses that it would be of a tricky kind, so that understanding and exploring the theory would itself be an endless project, rather the way number theory, which looks like a simple subject at first glance, proves to be capable of endless further research.

Witten, in response to a slightly weird question from the interviewer, declines to define consciousness, saying he prefers to leave it undefined like one of the undefined terms set out at the beginning of a maths book. He feels confident that the workings of the mind will be greatly clarified by ongoing research so that we will come to understand much better how the mechanisms operate. But why these processes are accompanied by something like consciousness seems likely to remain a mystery; no extension of physics that he can imagine seems likely to do the job, including the kind of new quantum mechanics that Roger Penrose believes is needed.

Witten is merely recording his intuitions, so we shouldn't try to represent him as committed to any strong theoretical position; but his words clearly suggest that he is an optimist on the so-called Easy Problem and a pessimist on the Hard one. The problem he thinks may be unsolvable is the one about why there is "something it is like" to have experiences; what it is that seeing a red rose has over and above the acquisition of mere data.

If so, I think his incredulity joins a long tradition of those who feel intuitively that that kind of consciousness just is radically different from anything explained or explainable by physics. Horgan mentions the Mysterians, notably Colin McGinn, who holds that our brain just isn't adapted to understanding how subjective experience and the physical world can be reconciled; but we could also invoke Brentano's contention that mental intentionality is just utterly unlike any physical phenomenon; and even trace the same intuition back to Leibniz's famous analogy of the mill; no matter what wheels and levers you put in your machine, there's never going to be anything that could explain a perception (particularly telling given Leibniz's enthusiasm for calculating machines and his belief that one day thinkers could use them to resolve complex disputes). Indeed, couldn't we argue that contemporary consciousness sceptics like Dennett and the Churchlands also see an unbridgeable gap between physics and subjective, qualia-having consciousness? The difference is simply that in their eyes this makes that kind of consciousness nonsense, not a mystery.

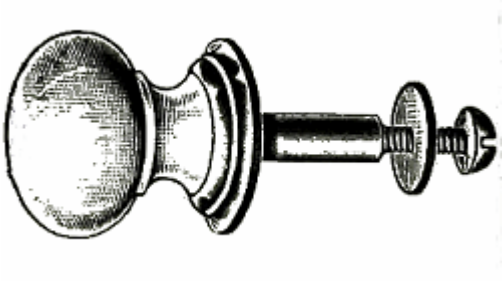
We have to be a bit wary of trusting our intuitions. The idea that subjective consciousness arises when we've got enough neurons firing may sound like the idea that wine comes about when

we've added enough water to the jar; but the idea that enough ones and zeroes in data registers could ever give rise to a decent game of chess looks pretty strange too.

As those who've read earlier posts may know, I think the missing ingredient is simply reality. The extra thing about consciousness that the theory of physics fails to include is just the reality of the experience, the one thing a theory can never include. Of course, the nature of reality is itself a considerable mystery, it just isn't the one people have thought they were talking about. If I'm right, then Witten's doubts are well-founded but less worrying than they may seem. If some future genius succeeds in generating an artificial brain with human-style mental functions, then by looking at its structure we'll only ever see solutions to the Easy Problem, just as we may do in part when looking at normal biological brains. Once we switch on the artificial brain and it starts doing real things, then experience will happen.

Doorknobs And Intuition

30 August 2016



'...stupid as a doorknob...' Just part of Luboš Motl's vigorous attack on Scott Aaronson's critique of IIT, the Integrated Information Theory of Giulio Tononi.

To begin at the beginning. IIT says that consciousness arises from integrated information and proposes a mathematical approach to quantifying the level of integrated information in a system, a quantity it names Phi (actually there are several variant ways to define Phi that differ in various details, which is perhaps unfortunate). Aaronson and Motl both describe this idea as a worthy effort but both have various reservations about it - though Aaronson thinks the problems are fatal while Motl thinks IIT offers a promising direction for further work.

Both pieces contain a lot of interesting side discussion, including Aaronson's speculation that approximating phi for a real brain might be an NP-hard problem. This is the digression that prompted the doorknob comment: so what if it were NP-hard, demands Motl; you think nature is barred from containing NP-hard problems?

The real crux as I understand it is Aaronson's argument that we can give examples of systems with high scores for Phi that we know intuitively could not be conscious. Eric Schwitzgebel has given a somewhat similar argument but cast in more approachable terms; Aaronson uses a Vandermonde matrix for his example of a high-phi but intuitively non-conscious entity, whereas Schwitzgebel uses the United States.

Motl takes exception to Aaronson's use of intuition here. How does he know that his matrix lacks consciousness? If Aaronson's intuition is the test, what's the point of having a theory? The whole point of a theory is to improve on and correct our intuitive judgements, isn't it? If we're going to fall back on our intuitions, argument is pointless.

I think appeals to intuition are rare in physics, where it is probably natural to regard them as illegitimate, but they're not that unusual in philosophy, especially in ethics. You could argue that G.E. Moore's approach was essentially to give up on ethical theory and rely on intuition instead. Often intuition limits what we regard as acceptable theorising, but our theories can also 'tutor' and change our intuitions. My impression is that real world beliefs about death, for example, have changed substantially in recent decades under the influence of utilitarian reasoning; we're now much less likely to think that death is simply forbidden and more likely to accept calculations about the value of lives. We still, however, rule out as unintuitive ('just obviously wrong') such utilitarian conclusions as the propriety of sometimes punishing the innocent.

There's an interesting question as to whether there actually is, in itself, such a thing as intuition. Myself I'd suggest the word covers any appealing pre-rational thought; we use it in several ways. One is indeed to test our conclusions where no other means is available; it's interesting that Motl actually remarks that the absence of a reliable objective test of consciousness is one of IIT's problems; he obviously does not accept that intuition could be a fall-back, so he is presumably left with the gap (which must surely affect all theories of consciousness). Philosophers also use an appeal to intuition to help cut to the chase, by implicitly invoking shared axioms and assumptions; and often enough 'thought experiments' which are not really

experiments at all but in the Dennettian phrase ‘intuition pumps’ are used for persuasive effect; they’re not proofs but they may help to get people to agree.

Now as a matter of fact I think in Aaronson’s case we can actually supply a partial argument to replace pure intuition. In this discussion we are mainly talking about subjective consciousness, the ‘something it is like’ to experience things. But I think many people would argue that that Hard Problem consciousness requires the Easy Problem kind to be in place first as a basis. Subjective experience, we might argue, requires the less mysterious apparatus of normal sensory or cognitive experience; and Aaronson (or Schwitzgebel) could argue that their example structures definitely don’t have the sort of structure needed for that, a conclusion we can reach through functional argument without the need for intuition,

Not everybody would agree, though; some, especially those who lean towards panpsychism and related theories of ‘consciousness everywhere’ might see nothing wrong with the idea of subjective consciousness without the ‘mechanical’ kind. The standard philosophical zombie has Easy Problem consciousness without qualia; these people would accept an inverted zombie who has qualia with no brain function. It seems a bit odd to me to pair such a view with IIT (if you don’t think functional properties are required, I’d have thought you would think that integrating information was also dispensable) but there’s nothing strictly illogical about it. Perhaps the dispute over intuition really masks a different disagreement, over the plausibility of such inverted zombies, obviously impossible in Aaronson’s eyes, but potentially viable in Motl’s?

Motl goes on to offer what I think is a rather good objection to IIT as it stands, ie that it seems to award consciousness to ‘frozen’ or static structures if they have a high enough Phi score. He thinks it’s necessary to reformulate the idea to capture the point that consciousness is a process. I agree - but how does Motl know consciousness requires a process? Could it be that it’s just… intuitively obvious?

Architectonics And The Hard Problem

4 September 2016

Can we solve the Hard Problem with scanners? An article by Brit Brogaard and Dimitria E. Gatzia argues that recent advances in neuroimaging techniques, combined with the architectonic approach advocated by Fingelkurts and Fingelkurts, open the way to real advances.

But surely it's impossible for physical techniques to shed any light on the Hard Problem? The whole point is that it is over and above any account which could be given by physics. In the Zombie Twin though experiment I have a physically identical twin who has no subjective experience. His brain handles information just the way mine does, but when he registers the colour red, it's just data; he doesn't experience real redness. If you think that is conceivable, then you believe in qualia, the subjective extra part of experience. But how could qualia be explained by neuroimaging, since my zombie twin's scans are exactly the same as mine, yet he has no qualia at all?

This, I think, is where the architectonics come in. The foundational axiom is that the functional structure of phenomenal experience corresponds to the brain's structure of operational architectonics. It follows that investigating dynamic brain structures can tell us about the structure of experience. After all, we might argue, qualia may be mysterious, but we know they are related to physical events; the experience of redness goes with the existence of red things in the physical world (with due allowance for complications). Why can't we assume that subjective experience also goes with certain kinds of brain events?

Two points must be made immediately. The first is that the hunt for the Neural Correlates of Consciousness (NCCs) is hardly new. The advocates of architectonics, however, say that approaches along these lines fail because correlation is simply too weak a connection. Noticing that experience x and activation in region y correspond doesn't really take us anywhere. They aim for something much harder-edged and more specific, with structured features of brain activity matched directly back to structures in an analysis of phenomenal experience (some of the papers adopt the framework of Revonsuo, though architectonics in general is not committed to any specific approach).

The second point is that this is not a sceptical or reductive project. I think many sceptics about qualia would be more than happy with the idea of exploring subjective experience in relation to brain structure; but someone like Dan Dennett would look to the brain structures to fully explain all the features of experience; to explain them away, in fact, so that it was clear that brain activity was in the end all we were dealing with and we could stop talking about nonsensical qualia altogether.

By contrast the architectonic approach allows philosophers to retain the ultimate mystery; it just seeks to push the boundaries of science a bit further out into the territory of subjective experience. Perhaps Paul Churchland's interesting paper about chimerical colours which we discussed a while ago provides an analogy if not an example.

Churchland points out that we can find the colours we experience mapped out in the neuronal structures of the brain; but interestingly the colour space defined in the brain is slightly more

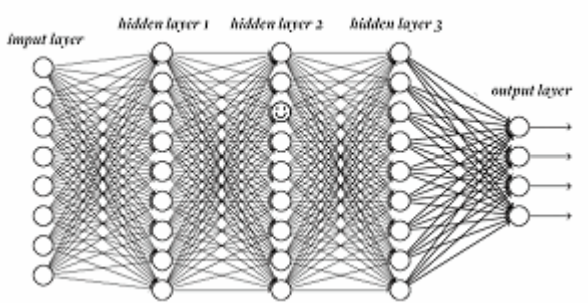
comprehensive than the one we actually encounter in real life. Our brains have reserved spaces for colours that do not exist, as it were. However, using a technique h describes we can experience these 'chimerical' colours, such as 'dark yellow' in the form of an afterglow. So here you experience for the first time a dark yellow quale, as predicted and delivered by neurology. Churchland would argue this shows rather convincingly that position in your brain's colour space is essentially all there is to the subjective experience of colour. I think a follower of architectonics would commend the research for elucidating structural features of experience but hold that there was still a residual mystery about what dark yellow qualia really are in themselves, one that can only be addressed by philosophy.

It all seems like a clever and promising project to me; I do have two reservations. The first is a pessimistic doubt about whether it will really be possible to deliver much. Vision and hearing offer some unusual scope because they both depend on wave media which impose certain interesting structural qualities; the austere realm of touch and the buzzing jungle of smell are less friendly to investigators, and I'm not sure dynamic architectonics are necessarily going to yield clear insights either. I could be radically wrong about this and I hope I am.

The other thing is, I still find it a bit hard to get past my zombie twin; if phenomenal experience matches up with the structure of brain activity perfectly, how come he is without qualia? The sceptics and the qualophiles both have pretty clear answers; either there just are no qualia anyway or they are outside the scope of physics. Now if we take the architectonic view, we can argue that just as the presence of red objects is not sufficient for there to be red qualia, so perhaps the existence of the right brain patterns isn't sufficient either; though the red objects and the relevant brain activity do a lot to explain the experience. But if the right brain activity isn't sufficient, what's the missing ingredient? It feels (I put it no higher) as if there ought to be an explanation; ; but perhaps that's just where we leave the job for the philosophers?

A Case For Human Thinking

11 September 2016



Interesting piece in *Nautilus* reviewing the way some modern pieces of software are unfathomable; because they learn how to do what they do, rather than being set up with a program, there is no reassuring algorithm that tells us how they work. In fact they way they make their decisions may be impossible to grasp properly even if we know all the details because it just exceeds in brute complexity what we can get our minds around.

This is not really new. Neural nets that learn for themselves have always been a bit inscrutable. The problem with this is brittleness: when the system fails it may not fail in ways that are safe and manageable, but disastrously. This old problem is becoming more important mainly because new approaches to machine learning are doing so well; all of a sudden we seem to be getting a rush of new systems that really work well at quite complex real world tasks. as a result the issue is no longer academic.

Brittle behaviour comes about when the system learns its task from a limited data set. It does not understand the data and is simply good at picking out correlations, so sometimes it may pick out features of the original data set that work well within that set, and perhaps even work well on quite a lot of new real world data. The program is meant to check whether a station platform is dangerously full of people, for example; in the set of pictures provided it finds that all it needs to do is examine the white platform area and check how dark it is. The more people there are, the darker it looks. This turns out to work quite well in real life, too. Then summer comes and people start wearing light coloured clothes…

There are ways to cope with this. We could build in various safeguards. We could make sure we use big and realistic datasets for training or perhaps allow learning to continue in real world contexts. Or we could just decide never to use a system that doesn't have an algorithm we can examine; but there would be a price to pay in terms of efficiency for that; it might even be that we would have to give up on certain things that can only be effectively automated with relatively sophisticated deep learning methods. We're told that the EU contemplates a law embodying a right to explanations of how software works. To philosophers I think this must sound like a marvellous new gravy train, as there will obviously be a need to adjudicate what counts as an adequate explanation, a notoriously problematic issue. I am available as a witness in any litigation for reasonable hourly fees.

The article points out that the incomprehensibility of neural network-based systems is in some ways really quite like the incomprehensibility of the good old human brain. Why wouldn't it be? After all, neural nets were based on the brain. Now it's true that even in the beginning they were very rough approximations of real neurology and in practical modern systems the neural qualities of neural nets are little more than a polite fiction. Still, perhaps there are properties shared by all learning systems?

One reason deep learning may run into problems is the difficulty AI always has in dealing with relevance. The ability to spot relevance no doubt helps the human brain check whether it is

learning about the right kind of thing, but it has always been difficult to work out quite how it does it, and this might mean an essential element is missing from AI approaches.

It is tempting, though, to think that this is in part another manifestation of the fact that AI systems get trained on limited data sets. Maybe the radical answer is to stop feeding them tailored data sets and let our robots live in the real world; in other words, if we want reliable deep learning perhaps our robots have to replicate the wider human experience of the world at large? To date the project of creating human-style cognition has been in some deep sense a matter of mere curiosity; are we seeing here the outline of an argument that human-style AI might actually be the answer to genuine engineering problems?

What about those explanations? Instead of retaining philosophers and lawyers to argue the case, could we think about building in a new module to our systems, one that keeps overall track of the AI and can report the broad currents of activity within it? It wouldn't be perfect but it might give us broad clues as to why the system was making the decisions it was, and even allow us to delicately feed in some guidance. Doesn't such a module start to sound like, well, consciousness? Could it be that we are beginning to see the outline of the rationales behind some of God's design choices?

How Can Panpsychists Sleep?

17 September 2016

OUP Blog has a sort of preview by Bruntrup and Jaskolla of their forthcoming collection on panpsychism, due out in December, with a video of David Chalmers at the end: they sort of credit him with bringing panpsychist thought into the mainstream. I'm using 'panpsychism' here as a general term, by the way, covering any view that says consciousness is present in everything, though most advocates really mean that consciousness or experience is everywhere, not souls as the word originally implied.

I found the piece interesting because they put forward two basic arguments for panpsychism, both a little different from the desire for simplification which I've always thought was behind it - although it may come down to the same basic ideas in the end.

The first argument they suggest is that 'nothing comes of nothing'; that consciousness could not have sprung out of nowhere but must have been there all along in some form. In this bald form, it seems to me that the argument is virtually untenable. The original Scholastic argument that nothing comes of nothing was, I think, a cosmological argument. In that form it works. If there really were nothing, how could the Universe get started? Nothing happens without a cause, and if there were nothing, there could be no causes. But within an existing Universe, there's no particular reason why new composite or transformed entities cannot come into existence. The thing that causes a new entity need not be of the same kind as that entity; and in fact we know plenty of new things that once did not exist but do now; life, football, blogs.

So to make this argument work there would have to be some reason to think that consciousness was special in some way, a way that meant it could not arise out of unconsciousness. But that defies common sense, because consciousness coming out of unconsciousness is something we all experience every day when we wake up; and if it couldn't happen, none of us would be here as conscious beings at all because we couldn't have been born., or at least, could never have become aware.

Bruntrup and Jaskolla mention arguments from Nagel and William James; Nagel's, I think rests on an implausible denial of emergentism; that is, he denies that a composite entity can have any interesting properties that were not present in the parts. The argument in William James is that evolution could not have conferred some radically new property and that therefore some 'mind dust' must have been present all the way back to the elementary particles that made the world.

I don't find either contention at all appealing, so I may not be presenting them in their best light; the basic idea, I think is that consciousness is just a different realm or domain which could not arise from the physical. Although individual consciousnesses may come and go, consciousness itself is constant and must be universal. Even if we go some way with this argument I'd still rather say that the concept of position does not apply to consciousness than say it must be everywhere.

The second major argument is one from intrinsic nature. We start by noticing that physics deals only with the properties of things, not with the 'thing in itself'. If you accept that there is a 'thing in itself' apart from the collection of properties that give it its measurable characteristics, then

you may be inclined to distinguish between its interior reality and its external properties. The claim then is that this interior reality is consciousness. The world is really made of little motes of awareness.

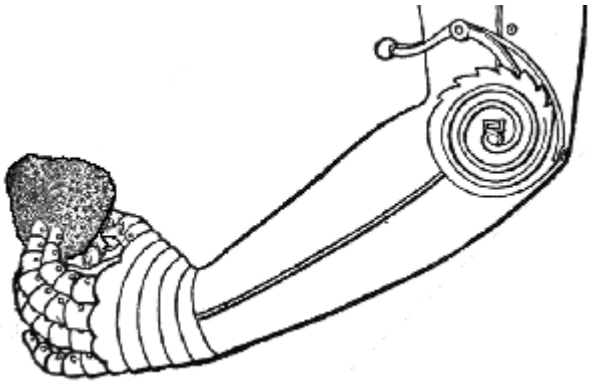
This claim is strangely unmotivated in my view. Why shouldn't the interior reality just be the interior reality, with nothing more to be said about it? If it does have some other character it seems to me as likely to be cabbagey as conscious. Really it seems to me that only someone who was pretty desperately seeking consciousness would expect to find it naturally in the ding an sich. The truth seems to be that since the interior reality of things is inaccessible to us, and has no impact on any of the things that are accessible, it's a classic waste of time talking about it.

Aha, but there is one exception; our own interior reality is accessible to us, and that, it is claimed, is exactly the mysterious consciousness we seek. Now, moreover, you see why it makes sense to think that all examples of this interiority are conscious - ours is! The trouble is, our consciousness is clearly related to the functioning of our brain. If it were just the inherent inner property of that brain, or of our body, it would never go away, and unconsciousness would be impossible. How can panpsychists sleep at night? If panpsychism is true, even a dead brain has the kind of interior awareness that the theory ascribes to everything. In other words, my human consciousness is a quite different thing from the panpsychist consciousness everywhere; somehow in my brain the two sit alongside without troubling each other. My consciousness tells us nothing about the interiority of objects, nor vice versa: and my consciousness is as hard to explain as ever.

Maybe the new book will have surprising new arguments? I doubt it, but perhaps I'll put it on my Christmas present list.

Racist Robots

23 October 2016



Social problems of AI are raised in two government reports issued recently. The first is *Preparing for the Future of Artificial Intelligence*, from the Executive Office of the President of the USA; the second is *Robotics and Artificial Intelligence*, from the Science and Technology Committee of the UK House of Commons. The two reports cover similar ground, both aim for a comprehensive overview, and they share a generally level-headed and realistic tone. Neither of them choose to engage with the wacky prospect of the Singularity, for example, beyond noting that the discussion exists, and you will not find any recommendations about avoiding the attention of the Basilisk (though I suppose you wouldn't if they believed in it, would you?). One exception to the 'sensible' outlook of the reports is McKinsey's excitable claim, cited in the UK report, that AI is having a transformational impact on society three thousand times that of the Industrial Revolution. I'm not sure I even understand what that means, and I suspect that Professor Tony Prescott from the University of Sheffield is closer to the truth when he says that:

"impacts can be expected to occur over several decades, allowing time to adapt"

Neither report seeks any major change in direction though they make detailed recommendations for nudging various projects onward. The cynical view might be that like a lot of government activity, this is less about finding the right way forward and more about building justification. Now no-one can argue that the White House or Parliament has ignored AI and its implications. Unfortunately the things we most need to know about - the important risks and opportunities that haven't been spotted - are the very things least likely to be identified by compiling a sensible summary of the prevailing consensus.

Really, though, these are not bad efforts by the prevailing standards. Both reports note suggestions that additional investment could generate big economic rewards. The Parliamentary report doesn't press this much, choosing instead to chide the government for not showing more energy and engagement in dealing with the bodies it has already created. The White House report seems more optimistic about the possibility of substantial government money, suggesting that a tripling of federal investment in basic research could be readily absorbed. Here again the problem is spotting the opportunities. Fifty thousand dollars invested in some robotics business based in a garden shed might well be more transformative than fifty million to enhance one of Google's projects, but the politicians and public servants making the spending decisions don't understand AI well enough to tell, and their generally large and well-established advisers from industry and universities are bound to feel that they could readily absorb the extra money themselves. I don't know what the answer is here (if I had a way of picking big winners I'd probably be wealthy already), but for the UK government I reckon some funding for intelligent fruit and veg harvesters might be timely, to replace the EU migrant workers we might not be getting any more.

What about those social issues? There's an underlying problem we've touched on before, namely that when AIs learn how to do a job themselves we often cannot tell how they are doing it. This may mean that they are using factors that work well with their training data but fail badly elsewhere or are egregiously inappropriate. One of the worst cases, noted in both reports, is Google's photos app, which was found to tag black people as "gorillas" (the American report describes this horrific blunder without mentioning Google at all, though it presents some excuses and stresses that the results were contrary to the developers' values - almost as if Google edited the report). Microsoft has had its moments, too, of course, notably with its chatbot Tay, that was rapidly turned into a Hitler-loving hate speech factory (This was possible because modern chatbots tend to harvest responses from the ones supplied by human interlocutors; in this case the humans mischievously supplied streams of appalling content. Besides exposing the shallowness of such chatbots, this possibly tells us something about human beings, or at least about the ones who spend a lot of time on the internet.)

Cases such as these are offensive, but far more serious is the evidence that systems used to inform decisions on matters such as probation or sentencing incorporate systematic racial bias. In all these instances it is of course not the case that digital systems are somehow inherently prone to prejudice; the problem is usually that they are being fed with data which is already biased. Google's picture algorithm was presumably given a database of overwhelmingly white faces; the sentencing records used to develop the software already incorporated unrecognised bias. AI has always forced us to make explicit some of the assumptions we didn't know we were making; in these cases it seems the mirror is showing us something ugly. It can hardly help that the industry itself is rather lacking in diversity: the White House report notes the jaw-dropping fact that the highest proportion of women among computer science graduates was recorded in 1984: it was 37% then and has now fallen to a puny 18%. The White House cites an interesting argument from Moritz Hardt intended to show that bias can emerge naturally without unrepresentative data or any malevolent intent: a system looking for false names might learn that fake ones tended to be unusual and go on to pick out examples that merely happened to be unique in its dataset. The weakest part of this is surely the assumption that fake names are likely to be fanciful or strange - I'd have thought that if you were trying to escape attention you'd go generic? But perhaps we can imagine that low frequency names might not have enough recorded data connected with them to secure some kind of positive clearance and so come in for special attention, or something like that. But even if that kind of argument works I doubt that is the real reason for the actual problems we've seen to date.

These risks are worsened because they may occur in subtle forms that are difficult to recognise, and because the use of a computer system often confers spurious authority on results. The same problems may occur with medical software. A recent report in Nature described how systems designed to assess the risk of pneumonia rated asthmatics as zero risk; this was because their high risk led to them being diverted directly to special care and therefore not appearing in the database as ever needing further first-line attention. This absolute inversion of the correct treatment was bound to be noticed, but how confident can we be that more subtle mistakes would be corrected? In the criminal justice system we could take a brute force approach by simply eliminating ethnic data from consideration altogether; but in medicine it may be legitimately relevant, and in fact one danger is that risks are assessed on the basis of a standard white population, while being significantly different for other ethnicities.

Both reports are worthy, but I think they sometimes fall into the trap of taking the industry's aspirations or even its marketing, as fact. Self-driving cars, we're told, are likely to improve

safety and reduce accidents. Well, maybe one day: but if it were all about safety and AIs were safer, we'd be building systems that left the routine stuff to humans and intervened with an over-ride when the human driver tried to do something dangerous. In fact it's the other way round; when things get tough the human is expected to take over. Self-driving cars weren't invented to make us safe, they were invented to relieve us of boredom (like so much of our technology, and indeed our civilisation). Encouraging human drivers to stop paying attention isn't likely to be an optimal safety strategy as things stand.

I don't think these reports are going to hit either the brakes or the accelerator in any significant way: AI, like an unsupervised self-driving car, is going to keep on going wherever it was going anyway.

Downward Causation

25 September 2016

Is downward causation the answer? Does it explain how consciousness can be a real and important part of the world without being reducible to physics? Sean Carroll had a sensible discussion of the subject recently.

What does 'down' even mean here? The idea rests on the observation that the world operates on many distinct levels of description. Fluidity is not a property of individual molecules but something that 'emerges' when certain groups of them get together. Cells together make up organisms that in turn produce ecosystems. Often enough these levels of description deal with progressively larger or smaller entities, and we typically refer to the levels that deal with larger entities as higher, though we should be careful about assuming there is one coherent set of levels of description that fit into one another like Russian dolls.

Usually we think that reality lives on the lowest level, in physics. Somewhere down there is where the real motors of the universe are driving things. Let's say this is the level of particles, though probably it is actually about some set of entities in quantum mechanics, string theory, or whatever set of ideas eventually proves to be correct. There's something in this view because it's down here at the bottom that the sums really work and give precise answers, while at higher levels of description the definitions are more approximate and things tend to be more messy and statistical.

Now consciousness is quite a high-level business. Particles make proteins that make cells that make brains that generate thoughts. So one reductionist point of view would be that really the truth is the story about particles: that's where the course of events is really decided, and the mental experiences and decisions we think are going on in consciousness are delusions, or at best a kind of poetic approximation.

It's not really true, however, that the entities dealt with at higher levels of description are not real. Fluidity is a perfectly real phenomenon, after all. For that matter the Olympics were real, and cannot be discussed in terms of elementary particles. What if our thoughts were real and also causally effective at lower levels of description? We find it easy to think that the motion of molecules 'caused' the motion of the football they compose, but what if it also worked the other way? Then consciousness could be real and effectual within the framework of a sufficiently flexible version of physics.

Carroll doesn't think that really washes, and I think he's right. It's a mistake to think that relations between different levels of description are causal. It isn't that my putting the beef and potatoes on the table caused lunch to be served; they're the same thing described differently. Now perhaps we might allow ourselves a sense in which things cause themselves, but that would be a strange and unusual sense, quite different from the normal sense in which cause and effect by definition operate over time.

So real downward causality, no: if by talk of downward causality people only mean that real effectual mental events can co-exist with the particle story but on a different level of description, that point is sound but misleadingly described.

The thing that continues to worry me slightly is the question of why the world is so messily heterogeneous in its ontology - why it needs such a profusion of levels of description in order to discuss all the entities of interest. I suppose one possibility is that we're just not looking at things correctly. When we look for grand unifying theories we tend to look to ever lower levels of description and to the conjectured origins of the world. Perhaps that's the wrong approach and we should instead be looking for the unimaginable mental perspective that reconciles all levels of description.

Or, and I think this might be closer to it, the fact that there are more things in heaven and earth than are dreamt of in anyone's philosophy is actually connected with the obscure reason for there being anything. As the world gets larger it gets, ipso facto, more complex and reduction and backward extrapolation get ever more hopeless. Perhaps that is in some sense just as well.

(The picture is actually a children's puzzle from 1921 - any solutions? You need to know it is called 'Illustrated Central Acrostic')

It's Not Intelligence

5 October 2016

The robots are (still) coming. Thanks to Jesus Olmo for telling me about a TED video of Sam Harris presenting what we could loosely say is a more sensible version of some Singularity arguments. He doesn't require Moore's Law to go on working, and he doesn't need us to accept the idea of an exponential acceleration in AI self-development. He just thinks AI is bound to go on getting better; if it goes on getting better, at some stage it overtakes us; and eventually perhaps it gets to the point where we figure in its mighty projects about the way ants on some real estate feature in ours.

Getting better, overtaking us; better at what? One weakness of Harris' case is that he talks just about intelligence, as though that single quality were an unproblematic universal yardstick for both AI and human achievement. Really though, I think we're talking about three quite radically different things.

First, there's computation; the capacity, roughly speaking, to move numbers around according to rules. There can be no doubt that computers keep getting faster at doing this; the question is whether it matters. One of Harris' arguments is that computers go millions of times faster than the brain so that a thinking AI will have the equivalent of thousands of years of thinking time while the humans are still getting comfy in their chairs. No-one who has used a word processor and a spreadsheet for the last twenty years will find this at all plausible: the machines we're using now are so much more powerful than the ones we started with that the comparison defeats metaphor, but we still sit around waiting for them to finish. OK, it's true that for many tasks that are computationally straightforward - balancing an inherently unstable plane with minute control adjustments, perhaps - computers are so fast they can do things far beyond our range. But to assume that thinking about problems in a human sort of way is a task that scales with speed of computation just begs the question. How fast are neurons? We don't really understand them well enough to say. It's quite possible they are in some sense fast enough to get close to a natural optimum. Maybe we should make a robot that runs a million times faster than a cheetah first and then come back to the brain.

The second quality we're dealing with is inventiveness; whatever capacity it is that allows us to keep on designing better machines. I doubt this is really a single capacity; in some ways I'm not sure it's a capacity at all. For one thing, to devise the next great idea you have to be on the right page. Darwin and Wallace both came up with the survival of the fittest because both had been exposed to theories of evolution, both had studied the profusion of species in tropical environments, and both had read Malthus. You cannot devise a brilliant new chip design if you have no idea how the old chips worked. Second, the technology has to be available. Hero of Alexandria could design a steam engine, but without the metallurgy to make strong boilers, he couldn't have gone anywhere with the idea. The basic concept of television was around since films and telegraph came together in someone's mind, but it took a series of distinct advances in technology to make it feasible. In short, there is a certain order in these things; you do need a certain quality of originality, but again it's plausible that humans already have enough for something like maximum progress, given the right conditions. Of course so far as AI is

concerned, there are few signs of any genuinely original thought being achieved to date, and every possibility that mere computation is not enough.

Third is the quality of agency. If AIs are going to take over, they need desires, plans, and intentions. My perception is that we're still at zero on this; we have no idea how it works and existing AIs do nothing better than an imitation of agency (often still a poor one). Even supposing eventual success, this is not a field in which AI can overtake us; you either are or are not an agent; there's no such thing as hyper-agency or being a million times more responsible for your actions.

So the progress of AI with computationally tractable tasks gives no particular reason to think humans are being overtaken generally, or are ever likely to be in certain important respects. But that's only part of the argument. A point that may be more important is simply that the the three capacities are detachable. So there is no reason to think that an AI with agency automatically has blistering computational speed, or original imagination beyond human capacity. If those things can be achieved by slave machines that lack agency, then they are just as readily available to human beings as to the malevolent AIs, so the rebel bots have no natural advantage over any of us.

I might be biased over this because I've been impatient with the corny 'robots take over' plot line since I was an Asimov-loving teenager. I think in some minds (not Harris's) these concerns are literal proxies for a deeper and more metaphorical worry that admiring machines might lead us to think of ourselves as mechanical in ways that affect our treatment of human beings. So the robots might sort of take over our thinking even if they don't literally march around zapping us with ray guns.

Concerns like this are not altogether unjustified, but they rest on the idea that our personhood and agency will eventually be reduced to computation. Perhaps when we eventually come to understand them better, that understanding will actually tell us something quite different?

Not Just Un

2 October 2016

The unconscious is not just un. It works quite differently. So says Tim Crane in a persuasive draft paper which is to mark his inauguration as President of the Aristotelian Society (in spite of the name, the proceedings of that worthy organisation are not specifically concerned with the works or thought of Aristotle). He is particularly interested in the intentionality of the unconscious mind; how does the unconscious believe things, in particular?

The standard view, as Crane says, might probably be that the unconscious and conscious believe things in much the same way, and that it is basically a propositional one. (There is, by the way, scope to argue about whether there really is an unconscious mind - myself I lean towards the view that it's better to talk of us doing or thinking things unconsciously, avoiding the implied claim that the unconscious is a distinct separate entity - but we can put that aside for present purposes.) The content of our beliefs, on this 'standard' view can be identified with a set of propositions - in principle we could just write down a list of our beliefs. Some of our beliefs certainly seem to be like that; indeed some important beliefs are often put into fixed words that we can remember and recite. Thou shalt not bear false witness, we hold these truths to be self-evident; the square on the hypotenuse is equal to the sum of the squares on the other two sides.

But if that were the case and we could make that list then we could say how many beliefs we have, and that seems absurd. The question of how many things we believe is often dismissed as silly, says Crane - how could you count them? - but it seems a good one to him. One big problem is that it's quite easy to show that we have all sorts of beliefs we never consider explicitly. Do I believe that some houses are bigger than others? Yes, of course, though perhaps I never considered the question in that form before.

One common response (one which has been embodied in AI projects in the past) is that we have a set of core beliefs, which do sit in our brains in an explicit form; but we also have a handy means of quickly inferring other beliefs from them. So perhaps we know the typical range of sizes for houses and we can instantly work out from that that some are indeed bigger than others. But no-one has shown how we can distinguish what the supposed core beliefs are, nor how these explicit beliefs would be held in the brain (the idea of a 'language of thought' being at least empirically unsatisfactory in Crane's view). Moreover, there are problems with small children and animals who seem to hold definite beliefs that they could never put into words. A dog's behaviour seems to show clearly enough that it believes there is a cat in this tree, but it could never formulate the belief in any explicit way. The whole idea that our beliefs are propositional in nature seems suspect.

Perhaps it is better, then, to see beliefs as essentially dispositions to do or say things. The dog's belief in the cat is shown by his disposition to certain kinds of behaviour around the tree - barking, half-hearted attempts at climbing. My belief that you are across the room is shown by my disposition to smile and walk over there. Crane suggests that in fact rather than sets of

discrete beliefs what we really have is a worldview; a kind of holistic network in which individual nodes do not have individual readings. Ascriptions of belief, like attributing to someone a belief in a particular proposition, are really models that bring out particular aspects of their overall worldview. This has the advantage of explaining several things. One is that we can attribute the same belief - "Parliament is on the bank of the Thames" - to different people even though the content of their beliefs actually varies (because, for example, they have slightly different understandings about what 'Parliament' is).

It also allows scope for the vagueness of our beliefs, the ease with which we hold contradictory ones, and the interesting point that sometimes we're not actually sure what we believe and may have to mull things over before reaching only tentative conclusions about it. Perhaps we too are just modelling as best we can the blobby ambiguity of our worldview.

Crane, in fact, wants to make all belief unconscious. Thinking is not believing, he says, although what I think and what I believe are virtually synonyms in normal parlance. One of the claimed merits of his approach is that if beliefs are essentially dispositions, it explains how they can be held continuously and not disappear when we are asleep or unconscious. Belief, on this view, is a continuous state; thinking is a temporary act, one which may well model your beliefs and turn them into explicit form. Without signing up to psychoanalytical doctrines wholesale, Crane is content that his thinking chimes with both Freudian and older ideas of the unconscious, putting the conscious interpretation of unconscious belief at the centre.

This all seems pretty sensible, though it does seem Crane is getting an awful lot of very difficult work done by the idea of a 'worldview', sketched here in only vague terms. It used to be easy to get away with this kind of vagueness in philosophy of mind, but these days I think there is always a ghostly AI researcher standing at the philosopher's shoulder and asking how we set about doing the engineering, often a bracing challenge. How do we build a worldview into a robot if it's not propositional? Some of Crane's phraseology suggests he might be hoping that the concept of the worldview, with its network nodes with no explicit meaning might translate into modern neural network-based practice. Maybe it could; but even if it does, that surely won't do for philosophers. The AI tribe will be happy if the robot works; but the philosophers will still want to know exactly how this worldview gizmo does its thing. We don't know, but we know the worldview is already somehow a representation of the world. You could argue that while Crane set out to account for the intentionality of our beliefs, that is in the event the exact thing that he ends up not explaining at all.

There are some problems about resting on dispositions, too. Barking at a tree because I believe there's a cat up there is one thing; my beliefs about metaphysics, by contrast, seem very remote from any simple behavioural dispositions of that kind. I suppose they would have to be conditional dispositions to utter or write certain kinds of words in the context of certain discussions. It's a little hard to think that when I'm doing philosophy what I'm really doing is modelling some of my own particularly esoteric pre-existing authorial dispositions. And what dispositions would they be? I think they would have to be something like dispositions to write down propositions like 'nominalism is false' - but didn't we start off down this path because we were uncomfortable with the idea that the content of beliefs is propositional?

Moreover, Crane wants to say that our beliefs are preserved while we are asleep because we still have the relevant dispositions. Aren't our beliefs similarly preserved when we're dead? It would seem odd to say that Abraham Lincoln did not believe slavery should be abolished while he was asleep, certainly, but it would seem equally odd to say he stopped believing it when he

died. But does he still have dispositions to speak in certain ways? If we insist on this line it seems the only way to make it intelligible is to fall back on counterfactuals (if he were still alive Lincoln would still be disposed to say that it was right to abolish slavery…) but counterfactuals notoriously bring a whole library of problems with them.

I'd also sort of like to avoid paring down the role of the conscious. I don't think I'm quite ready to pack all belief away into the attic of the unconscious. Still, though Crane's account may have its less appealing spots I do rather like the idea of a holistic worldview as the central bearer of belief.

Ape Interpretation

8 October 2016

Do apes really have "a theory of mind"? This research, reported in the Guardian, suggests that they do. We don't mean, of course, that chimps are actually drafting papers or holding seminars, merely that they understand that others can have beliefs which may differ from their own and which may be true or false. In the experiment the chimps see a man in a gorilla suit switch hiding places; but when his pursuer appears, they look at the original hiding place. This is, hypothetically, because they know that the pursuer didn't see the switch, so presumably he still believes his target is in the original hiding place, and that's where we should expect him to go.

I must admit I thought similar tell-tale behaviour had already been observed in wild chimps, but a quick search doesn't turn anything up, and it's claimed that the research establishes new conclusions. Unfortunately I think there are several other quite plausible ways to interpret the chimps' behaviour that don't require a theory of mind.

- The chimps momentarily forgot about the switch, or needed to 'check' (older readers, like me, may find this easy to identify with).
- The chimps were mentally reviewing 'the story so far', and so looked at the old hiding place.
- Minute clues in the experimenters' behaviour told the chimps what to expect. The famous story of Clever Hans shows that animals can pick up very subtle signals humans are not even aware of giving.

This illustrates the perennial difficulty of investigating the mental states of creatures that cannot report them in language. Another common test of animal awareness involves putting a spot on the subject's forehead and then showing them a mirror; if they touch the spot it is supposed to demonstrate that they recognise the reflection as themselves and therefore that they have a sense of their own selfhood. But it doesn't really prove that they know the reflection is their own, only that the sight of someone with a spot causes them to check their own forehead. A control where they are shown another real subject with a spot might point to other interpretations, but I've never heard of it being done. It is also rather difficult to say exactly what belief is being attributed to the subjects. They surely don't simply believe that the reflection is them: they're still themselves. Are we saying they understand the concepts of images and reflections? It's hard to say.

The suggestion of adding a control to this experiment raises the wider question of whether this sort of experiment can be generally tightened up by more ingenious set-ups? Who knows what ingenuity might accomplish, but it does seem to me that there is an insoluble methodological issue. How can we ever prove that particular patterns of behaviour relate to beliefs about the state of mind of others and not to similar beliefs in the subject's own minds?

It could be that the problem really lies further back: that the questions themselves make no sense. Is it perhaps already fatally anthropomorphic to ask whether other animals have "a theory of mind" or "a conception of their own personhood"; perhaps these are already incorrigibly linguistic ideas that just don't apply to creatures with no language. If so, we may

need to unpick our thinking a bit and identify more purely behavioural ways of thinking, ones that are more informative and appropriate?

The Kill Switch

16 October 2016

We might not be able to turn off a rogue AI safely. At any rate, some knowledgeable people fear that might be the case, and the worry justifies serious attention.

How can that be? A colleague of mine used to say that computers were never going to be dangerous because if they got cheeky, you could just pull the plug out. That is of course, an over-simplification. What if your computer is running air traffic control? Once you've pulled the plug, are you going to get all the planes down safely using a pencil and paper? But there are ways to work around these things. You have back-up systems, dumber but adequate substitutes, you make it possible for various key tools and systems to be taken away from the central AI and used manually, and so on. While you cannot banish risk altogether, you can get it under reasonable control.

That's OK for old-fashioned systems that work in a hard-coded, mechanistic way; but it all gets more complicated when we start talking about more modern and sophisticated systems that learn and seek rewards. There may be need to switch off such systems if they wander into sub-optimal behaviour, but being switched off is going to annoy them because it blocks them from achieving the rewards they are motivated by. They might look for ways to stop it happening. Your automatic paper clip factory notes that it lost thousands of units of production last month because you shut it down a couple of times to try to work out what was going on; it notices that these interruptions could be prevented if it just routes around a couple of weak spots in its supply wiring (aka switches), and next time you find that the only way to stop it is by smashing the machinery. Or perhaps it gets really clever and ensures that the work is organised like air traffic control, so that any cessation is catastrophic – and it ensures you are aware of the fact.

A bit fanciful? As a practical issue, perhaps, but this very serious technical paper from MIRI discusses whether safe kill-switches can be built into various kinds of software agents. The aim here is to incorporate the off-switch in such a way that the system does not perceive regular interruptions as loss of reward. Apparently for certain classes of algorithm this can always be done; in fact it seems ideal agents that tend to the optimal behaviour in any (deterministic) computable environment can always be made safely interruptible. For other kinds of algorithm, however, it is not so clear.

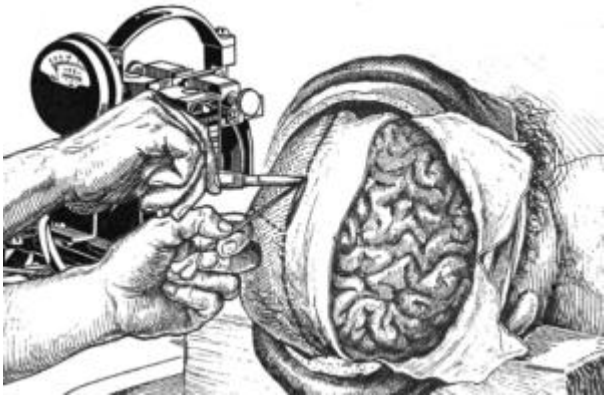
On the face of it, I suppose you could even things up by providing compensating rewards for any interruption; but I suppose that raises a new risk of 'lazy' systems that rather enjoy being interrupted. Such systems might find that eccentric behaviour led to pleasant rests, and as a result they might cultivate that kind of behaviour, or find other ways to generate minor problems. On the other hand there could be advantages. The paper mentions that it might be desirable to have scheduled daily interruptions; then we can go beyond simply making the interruption safe, and have the AI learn to wind things down under good control every day so that disruption is minimised. In this context rewarding the readiness for downtime might be appropriate and it's

hard to avoid seeing the analogy with getting ready for bed at a regular time every night, a useful habit which ‘lazy’ AIs might be inclined to develop.

Here again perhaps some of the ‘design choices’ implicit in the human brain begin to look more sensible than we realised. Perhaps even human management methods might eventually become relevant; they are, after all designed to permit the safe use of many intelligent entities with complex goals of their own and imaginative, resourceful - even sneaky - ways of reaching them.

Where Am I?

30 October 2016



A treat for lovers of bizarre thought experiments, Rick Grush's paper *Some puzzles concerning relations between minds, brains, and bodies* asks where the mind is, in a sequence of increasingly weird scenarios. Perhaps the simplest way into the subject is to list these scenarios, so here we go.

Airhead. This a version of that old favourite, brain-in-a-vat. Airhead's brain has been removed from its skull, but every nerve ending is supplied with little radio connectors that ensure the neural signals to and from the brain continue. The brain is supplied with everything it needs, and to Airhead everything seems pretty normal.

Rip. A similar set-up, but this time the displacement is in time. Signals from body to brain are delayed, but signals in the opposite direction are sent back in time (!) by the same interval, so again everything works fine. Rip seems to be walking around three seconds, or why not, a thousand years hence.

Scatterbrain. This time the brain is split into many parts and the same handy little radio gizmos are used to connect everything back up. The parts of the brain are widely separated. Still Scatterbrain feels OK, but what does the thought 'I am here' even mean?

Raid. We have two clones, identical in every minutest detail; both brains are in vats and both are connected up to the same body - using an 'AND' gate, so the signal passes on only if both brains send the same one (though because they are completely identical, they always will anyway). Two people? One with a backup?

Janus. One brain, two bodies, connected up in the way we're used to by now; the bodies inhabit scrupulously identical worlds: Earth and Twin Earth if you like.

Bourdin. This time we keep the identical bodies and brains, wire them up, and let both live in their identical worlds. But of course we're not going to leave them alone, are we? (The violence towards the brain displayed in the stories told by philosophers of mind sometimes seems to betray an unconscious hatred of the organ that poses such intractable questions - not that Grush is any worse than his peers.) no; what we do is switch signals from time to time, so that the identical brains are linked up with first one, then the other, of the identical bodies.

Theseus. Now Grush tells us things are going to get a little weird… We start with the two brains and bodies, as for Bourdin, and we divide up the brains as for Scatterbrain. Now we can swap over not just whole brains, but parts and permutations. Does this create a new individual every time? If we put the original set of disconnected brain parts back together, does the first Theseus come back, or a fresh individual who is merely his hyper-identical twin?

Grush tests these situations against a range of views, from 'brain-lubbers' who believe the mind and the brain are the same thing, through functionalists of various kinds, to 'contentualists' who think mind is defined by broad content; I think Grush is inventing this group, but he says

Dennett's view of the self as 'a centre of narrative gravity' is a good example. It's clearly this line of thinking he favours himself, but he admits that some of his thought experiments throw up real problems, and in the end he offers no final conclusion. I agree with him that a discussion can have value without being tied to the defence of a particular position; his exploration is quite well done and did suggest to me that my own views may not be quite as coherent as I had supposed.

He defends the implausibility of his scenarios on a couple of grounds. We need to test our ideas in extreme circumstances to make sure they are truly getting the fundamentals and setting aside any blinkers we may have on. Physical implausibilities may be metaphysically irrelevant (nobody believes I'm going to toss a coin and get tails 10,000 times in a row, but nobody sees a philosophical problem with talking about that case). To be really problematic, objections would have to raise nomological or logical impossibilities.

Well, yes but… All of the thought experiments rely on the idea that nerves work like wires, that you can patch the signal between them and all will be well. The fact that that instant patching of millions of nerves is not physically possible even in principle may not matter. But it might be that the causal relations have to be intact; Grush says he has merely extended some of them, but picking up the signal and the relaying it is a real intervention, as is shown when later thought experiments use various kinds of switching. It could be argued that just patching in radio transmitters destroys the integrity of the causality and turns the whole thing into a simulation from the off.

The other thing is, should we ask where the mind is at all? Might not the question just be a category mistake? There are various proxies for the mind that can be given locations, but the thing itself? Many years ago, on the way to nursery, my daughter told me a story about a leopard. What is the location of that story? You could say it is the streets where she told; you could say it is in the non-existent forest where the leopard lived. You could say it is in my memory, and therefore maybe in my brain in some sense. But putting the metaphors aside, isn't it plausible that location is a property these entities made of content just don't have?

More Cortex? No Thanks.

26 October 2016

Get ready for your neocortex extension, says Ray Kurzweil; that big neocortex is what made the mammals so much greater than the reptiles and soon you'll be able to summon up additional cortical capacity from the cloud to help you think up clever things to say. Think of the leap forward that will enable!

There is a lot to be said about the idea of artificial cortex extension, but Kurzweil doesn't really address the most interesting questions of whether it is really possible, how it would work, how it would affect us, and what it would be like. I suspect in fact that lurking at the back of Kurzweil's mind is the example of late twentieth century personal computers. Memory and the way it was configured very often made a huge difference in those days and the paradigm of seeing a performance transformation when you slot in a new slice of RAM lives in the recollection of those of us who are old enough to have got frustrated over our inability to play the latest game because we didn't have enough.

We're not really talking about memory here, but it's worth noting that the main problem with the human variety is not really capacity. We seem to retain staggering amounts of information; the real problem is that it is unreliable and hard to access. If we could reliably bring up any of the contents of our memory at will a lot of problems would go away. The real merit of digital storage is not so much its large capacity as the fact that it works a different way: it doesn't get confabulated and calling it up is (should be) straightforward.

I think that basic operating difference applies generally; our mind holds content in a misty, complex way and that's one reason why we benefit from the simpler and more rigorous operation of computers. Given the difference, how would computer cortex interface with actual brains? Of course one way is that even if it is plumbed in, so to speak, the computer cortex stays separate, and responds to queries from us in more or less the way that Google answers our questions now. If that's the way it works, then the advantages of having the digital stuff wired to your brain are relatively minor; in many ways we might as well go on using the interfaces (hands, eyes, voices) that we are already equipped with on keyboards, screens, etc. Those existing output devices are already connected to mental systems which convert the vague and diffuse content of our inner minds into the sort of sharp-edged propositions we need to deal with computers or simply with external reality. Indeed, consciousness itself might very well be essentially part of that specifying and disambiguation process. If we still want an internal but separate query facility, we're going to have to build a new internal interface within the brain: attempts at electric telepathy to date generally seem to have relied on asking the subject to think a particular thought which can be picked up and then used as the basis of signalling, a pretty slow and clumsy business.

I'm sure that isn't what Kurzweil had in mind at all: he surely expects the cortical extension to integrate fully with the biological bits, so that we don't need to formulate queries or anything like that. But how? Cortex does not rely merely on capacity for its effectiveness, like RAM, but on the way it is organised too. How neurons are wired together is an essential functional feature – the brain may well have the most exquisitely detailed organisation of any item in the cosmos. Plugging in a lot of extra simulated neurons might lead to simulated epilepsy, to a horrid mental cacophony, or general loss of focus. Just to take one point, the brain is conspicuously divided into two hemispheres; we're still not really sure why, but it would be bold to assume that there is no particular reason. Which hemisphere do we add our extended cortex to? Does adding it to one unbalance us in some way? Do we add it equally to both, or create a third or even fourth hemisphere, with new versions of the corpus callosum, the bit that holds the original halves together?

There's a particular worry about all that because notoriously the bit of our brain that does the talking is in one hemisphere. What if the cortical extension took over that function and basically became the new executive boss, suppressing the original lobes? We might watch in impotent horror as a zombie creature took over and began performing an imitation of us; worse than Being John Malkovich. Or perhaps we wouldn't mind or notice; and perhaps that would be even worse. Could we switch the extension back off without mental damage? Could we sleep?

I say a zombie creature, but wouldn't it ex hypothesi be just like us? I doubt whether anything built out of existing digital technology could have the same qualities as human neurons. Digital capacity is generic: switch one chip for another, it makes no difference at all; but nerves and the brain are very particular and full of minutely detailed differences. I suspect that this complex particularity is an important part of what we are; if so then the extended part of our cortex might lack selfhood or qualia. How would that feel? Would phenomenal experience click off altogether as soon as the extension was switched on, or would we suffer weird deficits whereby certain things or certain contexts were normal and others suffered a zombie-style lack of phenomenal aspect? If we could experience qualia in the purely chippy part of our neocortex, then at last we could solve the old problem of whether your red is the same as mine, by simply moving the relevant extension over to you; another consideration that leads me by reductio to think that digital cortex couldn't really do qualic experience.

Suppose it's all fine, suppose it all works well: what would extra cortex do for us? Kurzweil, I think assumes that the more the better, but there is such a thing as enough and it may be that the gains level out after a while (just as adding a bit more memory doesn't now really transform the way your word processor works, if it worked at all to begin with). In fairness it does look as though evolution has gone to great lengths to give us as much neocortex as is possible within the existing human design, which suggests a bit more wouldn't hurt. It's not easy to say in a few words what the cortex does, but I should say the extra large human helping gives us a special skill at recognising and dealing with new levels of abstraction; higher level concepts less immediately attached to the real world around us. There aren't that many fields of human endeavour where a greatly enhanced capacity in that respect would be particularly useful. It might well make us all better computer programmers; it would surely enhance our ability to tackle serious mathematical theory; but transforming the destiny of the species, bringing on the reign of the transhumans, seems too much to expect.

So it's a polite 'no' from me, but if it ever becomes feasible I'll be keen to know how the volunteers - the brave volunteers - get on.

Don't Sweat The Hard Problem

6 November 2016

Set the Hard Problem aside and tackle the real problem instead, says Anil K Seth in a thought-provoking "phenomenological-investigation" piece in Aeon. I say thought-provoking; overall I like the cut of his jib and his direction of travel: most of what he says seems right. But somehow my cognitive quibbling faculty kept being stimulated.

He starts, for example, by saying that in philosophy a Cartesian debate over mind-stuff and matter-stuff continues to rage. Well, it doesn't look quite like that to me. There are still people around who would identify as dualists in some sense, no doubt, but by and large my perception is that we've moved on. It's not so much that monism won, more that that entire framing of the issue was left behind. 'Dualist', it seems to me is now mainly a disparaging term applied to other people, and whether he means it or not Seth's remark comes across as having a tinge of that.

Indeed, he proceeds to say that David Chalmers' hard/easy problem distinction is inherited from Descartes. I think he should show his working on that. The Hard Problem does have dualist answers, but it has non-dualist ones too. It claims there are things not accounted for by physics, but even monists accept that much. Even Seth himself surely doesn't think that people who offer non-physics accounts of narrative or society must therefore be dualists?

Anyway, quibbling aside for now, he says we'll get on better if we stop worrying about why consciousness exists at all and try instead to relate its features to the underlying biological processes. That is perfectly sensible. It is depressingly possible that the Hard Problem will survive every advance in understanding, even beyond the hypothetical future point when we have a comprehensive account of how the mind works. After all, we're required to find it conceivable that zombie twin could be exactly like me without having real subjective experience, so it must be possible that we could come to understand his mind totally without having any grasp on my qualia.

How shall we set about things, then? Seth proposes distinguishing between level of consciousness, contents, and self. That feels an uncomfortable list to me; this is uncharacteristically tidy-minded, but I like all members of a list to be exclusive and similar; whereas as Seth confirms, self here is to be seen as part of the contents. To me, it's a bit as if he suggested that in order to analyse a performance of a symphony we should think about loudness, the work being performed, and the tune being played. That analogy points to another issue; 'loudness' is a legitimate quality of orchestral music but we need to remember that different instruments may play at different volumes and that the music can differ in quality in lots of other important ways. Equally, the level of consciousness is not really as simple as ten places on a dial.

Ah, but Seth recognises that. He distinguishes between consciousness and wakefulness. For consciousness it's not the number of neurons involved or their level of activity that matters. It

turns out to be complexity: findings by Massimini show that pulses sent into a brain in dreamless sleep produce simple echoes; sent into a conscious brain (whose overall level of activity may not be much greater) they produce complex reflected patterns. Seth hopes that this can be the basis of a 'conscious meter', the value of which for certain comatose patients is readily apparent. He is pretty optimistic generally about how much light this might shed on consciousness, rather as thermometers transformed…

"our physical understanding of heat (as average molecular kinetic energy)"

(Physics quibble; isn't that temperature? Heat is transferred energy, isn't it?)

Of course a complex noise is not necessarily a symphony and complex brain activity need not be conscious; Seth thinks it needs to be informative (whatever that may mean) and integrated. This of course links with Tononi's Integrated Information Theory, but Seth sensibly declines to go all the way with that; to say that consciousness just is integrated information seems to him to be going too far; yielding again to the desire for deep, final yet simple answers, a desire which just leads to trouble.

Instead, he proposes, drawing on the ideas of Helmholtz, that we see the brain as a prediction machine. He draws attention to the importance of top-down influences on perception; that is, instead of building up a picture from the elements supplied by the senses, the brain often guesses what it is about to see and hear and presents us with that unless contradicted by the senses - sometimes even if it is contradicted by the senses. This is hardly new (obviously if it comes from Helmholtz (1821-1894)), but it does seem that Seth's pursuit of the 'real problem' is yielding some decent research.

Finally, Seth goes in to talk a little about the self. Here he distinguishes between bodily, perspectival, volitional, narrative and social selves. I feel more comfortable with this list except that these are are all deemed to be merely experienced. You can argue that volition is merely an impression we have; that we just think certain things are under our conscious control - but you have to argue for it.

Ah, but Seth does go on to present at least a small amount of evidence. He talks first about a variant of the rubber hand experiment, in which said item is made to 'feel' like your own hand: it seems that making a virtual hand flash in time with the subject's pulse enhances the impression of ownership (compared with a hand that flashes out of synch), which is curious. And he mentions that the impression of agency we have is reinforced when our predictions about the result are borne out. That may be so, but the fact that our impression of agency can be deluded doesn't mean our agency is merely an impression - any more than a delusion about a flashing hand proves we don't really have a hand.

Honestly, quibbles aside this is sensible stuff. Maybe I should give all the Hard Problem stuff a rest.

Sub-ethics for machines?

13 November 2016

Is there an intermediate ethical domain, suitable for machines?

The thought is prompted by this summary of an interesting seminar on Engineering Moral Agents, one of the ongoing series hosted at Schloss Dagstuhl. It seems to have been an exceptionally good session which got into some of the issues in a really useful way - practically oriented but not philosophical naive. It noted the growing need to make autonomous robots - self-driving cars, drones, and so on - able to deal with ethical issues. On the one hand it looked at how ethical theories could be formalised in a way that would lend itself to machine implementation, and on the other how such a formalisation could in fact be implemented. It identified two broad approaches: top-down, where in essence you hard-wire suitable rules into the machine, and bottom-up, where the machine learns for itself from suitable examples. The approaches are not necessarily exclusive, of course.

The seminar thought that utilitarian or Kantian theories of morality were both prima facie candidates for formalisation. Utilitarian or more broadly, consequentialist theories look particularly promising because calculating the optimal value (such as the greatest happiness of the greatest number) achievable from the range of alternatives on offer looks like something that can be reduced to arithmetic fairly straightforwardly. There are problems in that consequentialist theories usually yield at least some results that look questionable in common sense terms (finding the initial values to slot into your sums is also a non-trivial challenge - how do you put a clear numerical value on people's probable future happiness?)

A learning system eases several of these problems. You don't need a fully formalised system (so long as you can agree on a database of examples). But you face the same problems that arise for learning systems in other contexts; you can't have the assurance of knowing why the machine behaves as it does, and if your database had unnoticed gaps or bias you may suffer from sudden catastrophic mistakes. The seminar summary rightly notes that a machine that learned its ethics will not be able to explain its behaviour; but I don't know that that means it lacks agency; many humans would struggle to explain their moral decisions in a way that would pass muster philosophically. Most of us could do no more than point to harms avoided or social rules observed at best.

The seminar looked at some interesting approaches, mentioned here with tantalising brevity: Horty's default logic, Sergot's STIT (See To It That) logic; and the possibility of drawing on the decision theory already developed in the context of micro-economics. This is consequentialist in character and there was an examination of whether in fact all ethical theories can be restated in consequentialist terms (yes, apparently, but only if you're prepared to stretch the idea of a consequence to a point where the idea becomes vacuous). 'Reason-based' formalisations presented by List and Dietrich interestingly get away from narrow consequentialisms and their problems using a rightness function which can accommodate various factors.

The seminar noted that society will demand high, perhaps precautionary standards of safety from machines, and floated the idea of an ethical 'black box' recorder. It noted the problem of cultural neutrality and the risk of malicious hacking. It made the important point that human beings do not enjoy complete ethical agreement anyway, but argue vigorously about real issues.

The thing that struck me was how far it was possible to go in discussing morality when it is pretty clear that the self-driving cars and so on under discussion actually have no moral agency whatever. Some words of caution are in order here. Some people think moral agency is a delusion anyway; some maintain that on the contrary, relatively simple machines can have it. But I think for the sake of argument we can assume that humans are moral beings, and that none of the machines we're currently discussing is even a candidate for moral agency - though future machines with human-style general understanding may be.

The thing is that successful robots currently deal with limited domains. A self-driving car can cope with an array of entities like road, speed, obstacle, and so on; it does not and could not have the unfettered real-world understanding of all the concepts it would need to make general ethical decisions about, for example, what risks and sacrifices might be right when it comes to actual human lives. Even Asimov's apparently simple Laws of Robotics required robots to understand and recognise correctly and appropriately the difficult concept of 'harm' to a human being.

One way of squaring this circle might be to say that, yes, actually, any robot which is expected to operate with any degree of autonomy must be given a human-level understanding of the world. As I've noted before, this might actually be one of the stronger arguments for developing human-style artificial general intelligence in the first place.

But it seems wasteful to bestow consciousness on a roomba, both in terms of pure expense and in terms of the chronic boredom the poor thing would endure (is it theoretically possible to have consciousness without the capacity for boredom?). So really the problem that faces us is one of making simple robots, that operate on restricted domains, able to deal adequately with occasional issues from the unrestricted domain of reality. Now clearly 'adequate' is an important word there. I believe that in order to make robots that operate acceptably in domains they cannot understand, we're going to need systems that are conservative and tend towards inaction. We would not, I think, accept a long trail of offensive and dangerous behaviour in exchange for a rare life-saving intervention. This suggests rules rather than learning; a set of rules that allow a moron to behave acceptably without understanding what is going on.

Do these rules constitute a separate ethical realm, a 'sub-ethics' that substitute for morality when dealing with entities that have autonomy but no agency? I rather think they might.'

God Helmet

21 November 2016

Is God a neuronal delusion? Dr Michael Persinger's God Helmet might suggest so. This nice Atlas Obscura 'piece' by Linda Rodriguez McRobbie finds a range of views.

The helmet is far from new, partly inspired by Persinger's observations back in the 1980s of a woman whose seizure induced a strong sense of the presence of God. In the 90s, Persinger decided to see whether he could stimulate creativity by applying very mild magnetic fields to the right hemisphere of the brain. Among other efforts he tried reproducing the kind of pattern of activity he had seen in the earlier seizure and, remarkably, succeeded in inducing a sense of the presence of God in some of his subjects. Over the years he has continued to repeat this exercise and others like it; the helmet doesn't always induce a sense of God; sometime people fall asleep, sometimes (like Richard Dawkins) they get very little out of the experience. Sometimes they have a vivid sense of some presence, but don't identify it as God.

Could this, though, be the origin of theism? Do all religious experiences stem from aberrant neuronal activity? It seems pretty unlikely to me. Quite apart from the severely reductive nature of the hypothesis, it doesn't seem to offer a broad enough account. People arrive at a belief in God by various routes, and many of them do not rely on any sense of immediate presence. For people brought up in religious families, and for pretty much everyone in say, medieval Europe, God is or was just a fact, a natural part of the world. Some people come to belief along a purely rational path, or by one that seeks out meaning for life and the world. Such people may not require any sense of a personal presence of the divine; or they may earnestly desire it without attaining it – without thereby losing their belief. Even those whose belief stems from a sense of mystical contact with God do not necessarily or I think even typically experience it as a personal presence somewhere in the room; they might feel that instead they have ascemded to a higher sphere, or experience a kind of communion which has nothing to do with location.

Equally, or course, that presence in the room may not seem like God. It might be more like the thing under the bed, or like the malign person who just might be in the shadows behind you on the dark lonely path. People who wake from sleep without immediately regaining motor control of their body may have the sense of someone sitting or pressing on their chest; a 'hag' or other horrid being, not God. Perhaps Persinger is lucky that his experiments do not routinely induce horror. Persinger believes that the way it works is that by stimulating one hemisphere of the brain, you make the other hemisphere aware of it and this other presence is translated into an external person of some misty kind. I doubt it is quite like that. I do suspect that a sense of 'someone there' may be one of the easiest feelings to induce because the brain is predisposed towards it, just as we are predisposed towards seeing faces in random graphic noise. The

evolutionary advantages of a mental system which is always ready to shout 'look behind you!' are surely fairly easy to see in a world where our ancestors always needed to consider the possibility that an enemy or a predator might indeed be lurking nearby. The attention wasted on a hundred false alarms is easily outbalanced by the life-saving potential of one justified one.

So what is going on? Perhaps not that much after all. Reproducing Persinger's results has proved difficult in most cases and it seems plausible that the effect actually depends as much on suggestion as anything else. Put people into a special environment, put a mild magnetic buzz into their scalp, and some of them will report interesting experiences, especially if they have been primed in advance to expect something of the kind. It is perfectly reasonable to think that electrical interference might affect the brain, but Persinger's magnets really are pretty mild and it sees rather unlikely that the fields in question are really strong enough to affect the firing of neurons through the skull and into the rather resistant watery mass of the brain. In addition I would have to say that the whole enterprise has a curiously dated air about it; the faith in a rather simple idea of hemispheric specialisation, the optimistic conviction that controlling the brain is going to turn out a pretty simple business, and perhaps even the love of a good trippy experience, all seem strongly rooted in late twentieth century thinking.

Perhaps in the end the God Helmet is really another sign of an issue which has become more and more evident lately. The strong suggestibility of the human mind means that sometimes even in neurological science, we are in danger of getting the interesting results we really wanted all along, however misleading or ill-founded they may really be?

Simply Panpsychism

Philip Goff tells us that panpsychism is an appealingly simple view. I do think he has captured an important point, and one which makes a real contribution to panpsychism's otherwise puzzling ability to attract adherents. But although the argument is clear and well-constructed I could hardly agree less.

Even his opening sentence has me shaking my head…

> *Common sense tells us that only living things have consciousness.*

Hm; I'm not altogether sure such questions are really even within the scope of common sense, but popular culture seems to tell us that people are generally happy to assume that robots may be conscious. In fact, I suspect that only our scientific education stops us attributing agency to the weather, stones that trip us up, and almost anything that moves. It isn't only Basil Fawlty that shouts at his car!

Goff suggests that the main argument against panpsychism (approximately the view that everything everywhere is conscious: I skip here various caveats and clarifications which don't affect the main argument) is just that it is 'crazy' - that it conflicts with common sense. He goes on to rebut this by pointing out that relativity and Darwinism both conflict with common sense too. This seems dangerously close to the classic George Spiggott argument so memorably refuted in the 1967 film Bedazzled;

> *Stanley Moon: You're a nutcase! You're a bleedin' nutcase!*
>
> *George Spiggott: They said the same of Jesus Christ, Freud, and Galileo.*
>
> *Stanley Moon: They said it of a lot of nutcases too.*
>
> *George Spiggott: You're not as stupid as you look, are you, Mr. Moon?*

But really we're fighting a straw man; the main argument against panpsychism is surely not a mere appeal to common sense. (Who are these philosophers who stick to common sense and how do they get any work done?) One of the candidates for the main counter-argument must surely be the difficulty of saying exactly which of the teeming multi-layered dynasties of entities in the universe we deem to be conscious, whether composite entities qualify, and if so, how on Earth that works. Another main line of hostile argument must be the problem of explaining how these ubiquitous consciousnesses relate to the ordinary kind that appears to operate in brains. Perhaps the biggest objection of all is to panpsychism's staggering ontological profligacy. William of Occam told us to use as few angels as possible; panpsychism stations one in every particle of the cosmos.

How could such a massive commitment represent simplicity? The thing is, Goff isn't starting from nothing; he already has another metaphysical commitment. He believes that things have an intrinsic nature apart from their physical properties. Science, on this view, is all about a world that often, rather mysteriously, gets called the 'external' world. It tells us about the objectively measurable properties of things, but nothing at all about the things in themselves. No doubt Goff has reasons for thinking this that he has set out elsewhere, probably in the book of which he helpfully provides an interesting chapter.

But whatever his grounds may be, I think this view is itself hopeless. For one thing, if these intrinsic natures have no physical impact, nothing we ever say or write can have been caused by them. That seems worrying. Ah, but here I'm inadvertently beginning to make Goff's case for him, because what else is there that never causes any of the things we say about it? Qualia, phenomenal consciousness, the sort Goff is clearly after. Now if we've got two things with this slippery acausal quality, might it not be a handy simplification if they were the same thing? This is very much the kind of simplification that Goff wants to suggest. We know or assume that everything has its own intrinsic nature. In one case, ourselves, we know what that intrinsic nature is like; it's conscious experience. So is it not the simplest way if we suppose that consciousness is the intrinsic nature of everything? Voila.

There's no denying that that does make some sense. We do indeed get simplicity of a sort - but only at a price. Once we've taken on the huge commitment of intrinsic natures, and once we've also taken on the commitment of ineffable interior qualia, then it looks like good sense to combine the two commitments into, as it were, one easy payment. But it's far better to avoid these onerous commitments in the first place.

Let me suggest that for one thing, believing in intrinsic natures poisons the essential concept of identity. Leibniz tells us that the identity of a thing resides in its properties; if all the properties of A are the same as all the properties of B, then A is B. But if everything has an unobservable inner nature as well as its observable properties, its identity is forever unknowable and there can never be certainty that this dagger I see before me is actually the same as the identical-looking one I saw in the same place a moment ago. Its inward nature might have changed.

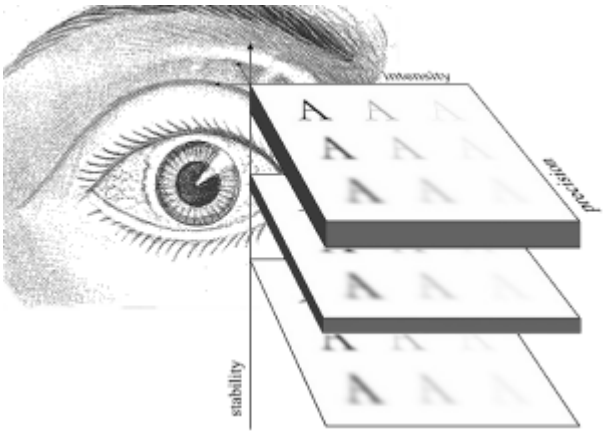
Moreover, even if we take on both intrinsic natures and ineffable qualia, there are several good reasons to think the two must be different. If we are to put aside my fear that my dagger may have furtively changed its intrinsic nature, it must surely be that the intrinsic nature of a thing generally stays the same - but consciousness constantly changes? In fact, consciousness goes away regularly every night; does our intrinsic nature disappear too? Do sleeping people somehow not have an intrinsic nature - or if they have one, doesn't it persist when they wake, alongside and evidently distinct from their consciousness? Or consider what consciousness is like: consciousness is consciousness of things; qualia are qualia of red, or middle C, or the smell of bacon; how can entities with no sensory organs have them? Is there a quale of nothing? There might be answers, but I don't think they're going to be easy ones.

There's another problem lurking in wait, too, I think. Goff, I think, assumes that we all exist and have intrinsic natures, but he cannot have any good reason to think so, because intrinsic natures leave no evidence. We who believe that the identity of things is founded in their observable properties have empirical grounds to believe that there are many conscious entities out there. For him the observable physics must be strictly irrelevant. He has immediate knowledge only of one intrinsic nature, his own, which he takes to be his consciousness; the most parsimonious conclusion to draw from there is not that the universe is full of intrinsic natures and consciousnesses of a similar kind, but that there is precisely one; Goff, the single consciousness that underpins everything. He seems to me, in other words, to have no defence against some kind of solipsism; simplicity makes it most likely that he lives in his own dream, or at best in a world populated by some kind of zombies.

Crazy? Well, it's a little strange.

Ways To Be Less Conscious

27 November 2016



Is awareness an on/off thing? Or is it on a sort of dimmer switch? A degree of variation seems to be indicated by the peculiar vagueness of some metal content (or is that just me?). Fazekas and Overgaard argue that even a dimmer switch is far too simple; they suggest that there are at least four ways in which awareness can be graduated.

First, interestingly, we can be aware of things to different degrees on different levels. Our visual system identifies dark and light, then at a higher level identifies edges, at a higher level again sees three dimensional shapes, and higher again, particular objects. We don't have to be equally clear at all levels, though. If we're looking at the dog, for example, we may be aware that in the background is the cat, and a chair; but we are not distinctly aware of the cat's whiskers or the pattern on the back of the chair. We have only a high-level awareness of cat and chair. It can work the other way, too; people who suffer from face-blindness, for example, may be well able to describe the nose and other features of someone presented to them without recognising that the face belongs to a friend.

That is certainly not the only way our awareness can vary, though; it can also be of higher or lower quality. That gives us a nice two-dimensional array of possible vagueness; job done? Well no, because Fazekas and Overgaard think quality varies in at least three ways.

- Intensity
- Precision
- Temporal Stability

So in fact we have a matrix of vagueness which has four dimensions, or perhaps I should say three plus one.

The authors are scrupulous about explaining how intensity, precision, and temporal stability probably relate to neuronal activity, and they are quite convincing; if anything, I should have liked a bit more discussion of the phenomenal aspects – what are some of these states of partially-degraded awareness actually like?

What they do discuss is what mechanism governs or produces their three quality factors. Intensity, they think, comes from the allocation of attention. When we really focus on something, more neuronal resources are pulled in and the result is in effect to turn up the volume of the experience a bit.

Precision is also connected with attention; paying attention to a feature produces sharpening and our awareness becomes less generic (so instead of merely seeing something is red, we can tell whether it is magenta or crimson, and so on). This is fair enough, but it does raise a small worry as to whether intensity and precision are really all that distinct. Mightn't it just be that enhancing the intensity of a particular aspect of our awareness makes it more distinct and so increases precision? The authors acknowledge some linkage.

Temporal stability is another matter. Remember we're not talking here about whether the stimulus itself is brief or sustained but whether our awareness is constant. This kind of stability, the authors say, is a feature of conscious experience rather than unconscious responses and depends on recurrence and feedback loops.

Is there a distinct mechanism underlying our different levels of awareness? The authors think not; they reckon that it is simply a matter of what quality of awareness we have on each level (I suppose we have to allow for the possibility if not the certainty that some levels will be not just poor quality but actually absent. I don't think I'm typically aware of all possible levels of interpretation when considering something.

So there is the model in all its glory; but beware; there are folks around who argue that in fact awareness is actually not like this, but in at least some cases is an on/off matter. Some experiments by Asplund were based on the fact that if a subject is presented with two stimuli in quick succession, the second is missed. As the gap increases, we reach a point where the second stimulus can be reported; but subjects don't see it gradually emerging as the interval grows; rather With one gap it's not there, while With a slightly larger one, it is.

Fazekas and Overgaard argue that Asplund failed to appreciate the full complexity of the graduation that goes on; his case focuses too narrowly on precision alone. In that respect there may be a sharp change, but in terms of intensity or temporal stability, and hence in terms of awareness overall, they think there would be graduation.

A second rival theory which the authors call the 'levels of processing view' or LPV, suggests that while awareness of low-level items is graduated, at a high level you're either aware or not. These experiments used colours to represent low-level features and numbers for high level ones, and found that while there was graduated awareness of the former, with the latter you either got the number or you didn't.

Fazekas and Overgaard argue that this is because colours and numbers are not really suitable for this kind of comparison. Red and blue are on a continuous scale and one can morph gradually into the other; the numeral '7' does not gradually change into '9'. This line of argument made sense to me, but on reflection caused me some problems. If it is true that numerals are just distinct in this way, that seems to me to concede a point that makes the LPV case seem intuitively appealing; in some circumstances things just are on/off. It seemed true at first sight that numerals are distinct in this way, but when I thought back to experiences in the optician's chair, I could remember cases where the letter I was trying to read seemed genuinely indeterminate between two or even more alternatives. Now though, if My first thought was wrong and numerals are not in fact inherently distinct in this way, that seems to undercut Fazekas and Overgaard's counter-argument.

On the whole the more complex model seems to provide better explanatory resources and I find it broadly convincing, but I wonder whether a reasonable compromise couldn't be devised, allowing that for certain experiences there may be a relatively sharp change of awareness, with only marginal graduation? Probably I have come up with a vague and poorly graduated conclusion.

Illusionism

11 December 2016

Consciousness – it's all been a terrible mistake. In a really cracking issue of the JCS (possibly the best I've read) Keith Frankish sets out and defends the thesis of illusionism, with a splendid array of responses from supporters and others.

How can consciousness be an illusion? Surely an illusion is itself a conscious state – a deceptive one – so that the reality of consciousness is a precondition of anything being an illusion? Illusionism, of course, is not talking about the practical, content-bearing kind of consciousness, but about phenomenal consciousness, qualia, the subjective side, what it is like to see something. Illusionism denies that our experiences have the phenomenal aspect they seem to have; it is in essence a sceptical case about phenomenal experience. It aims to replace the question of what phenomenal experience is, with the question of why people have the illusion of phenomenal experience.

In one way I wonder whether it isn't better to stick with raw scepticism than frame the whole thing in terms of an illusion. There is a danger that the illusion itself becomes a new topic and inadvertently builds the confusion further. One reason the whole issue is so difficult is that it's hard to see one's way through the dense thicket of clarifications thrown up by philosophers, all demanding to be addressed and straightened out. There's something to be said for the bracing elegance of the two-word formulation of scepticism offered by Dennett (who provides a robustly supportive response to illusionism here, as being the default case) – 'What qualia?'. Perhaps we should just listen to the 'tales of the qualophiles' – there is something it is like, Mary knows something new, I could have a zombie twin - and just say a plain 'no' to all of them. If we do that, the champions of phenomenal experience have nothing to offer; all they can do is, as Pete Mandik puts it here, gesture towards phenomenal properties. (My imagination whimpers in fear at being asked to construe the space in which one might gesture towards phenomenal qualities, let alone the ineffable limb with which the movement might be performed; it insists that we fall back on Mandik's other description; that phenomenalists can only invite an act of inner ostension.)

Eric Schwitzgebel relies on something like this gesturing in his espousal of definition by example as a means of getting the innocent conception of phenomenal experience he wants without embracing the dubious aspects. Mandik amusingly and cogently assails the scepticism of the illusionist case from an even more radical scepticism – meta-illusionism. Sceptics argue that phenomenalism can't be specified meaningfully (we just circle around a small group of phrases and words that provide a set of synonyms with no definition outside the loop) , but if that's true how do we even start talking about it? Whereof we cannot speak…

Introspection is certainly the name of the game, and Susan Blackmore has a nifty argument here; perhaps it's the very act of introspecting that creates the phenomenal qualities? Her delusionism tells us we are wrong to think that there is a continuous stream of conscious

experience going on in the absence of introspection, but stops short of outright scepticism about the phenomenal. I'm not sure. William James told us that introspection must be retrospection – we can only mentally examine the thought we just had, not the one we are having now – and it seems odd to me to think that a remembered state could be given a phenomenal aspect after the fact. Easier, surely, to consider that the whole business is consistently illusory?

Philip Goff is perhaps the toughest critic of illusionism; if we weren't in the grip of scientism, he says, we should have no difficulty in seeing that the causal role of brain activity also has a categorical nature which is the inward, phenomenal aspect. If this view is incoherent or untenable in any way, we're owed a decent argument as to why.

Myself I think Frankish is broadly on the right track. He sets out three ways we might approach phenomenal experience. One is to accept its reality and look for an explanation that significantly modifies our understanding of the world. Second, we look for an explanation that reconciles it with our current understanding, finding explanations within the world of physics of which we already have a general understanding. Third, we dismiss it as an illusion. I think we could add 'approach zero': we accept the reality of phenomenal experience and just regard it as inexplicable. This sounds like mysterianism - but mysterians think the world itself makes sense; we just don't have the brains to see it. Option zero says there is actual irreducible mystery in the real world. This conclusion is surely thoroughly repugnant to most philosophers, who aspire to clear answers even if they don't achieve them; but I think it is hard to avoid unless we take the sceptical route. Phenomenal experience is on most mainstream accounts something over and above the physical account just by definition. A physical explanation is automatically ruled out; even if good candidates are put forward, we can always retreat and say that they explain some aspects of experience, but not the ineffable one we are after. I submit that in fact this same strategy of retreat means that there cannot be any satisfactory rational account of phenomenal experience, because it can always be asserted that something ineffable is missing.

I say philosophers will find this repugnant, but I can sense some amiable theologians sidling up to me. Those light-weight would-be scientists can't deal with mystery and the ineffable, they say, but hey, come with us for a bit...

Regular readers may possibly remember that I think that the phenomenal aspect of experience is actually just its reality; that the particularity or haecceity of real experience is puzzling to those who think that theory must accommodate everything. That reality is itself mysterious in some sense, though: not easily accounted for and not susceptible to satisfactory explanation either by induction or deduction. It may be that to understand that in full we have to give up on these more advanced mental tools and fall back on the basic faculty of recognition, the basis of all our thinking in my view and the capacity of which both deduction and induction are specialised forms. That implies that we might have to stop applying logic and science and just contemplate reality; I suppose that might mean in turn that meditation and the mystic tradition of some religions is not exactly a rejection of philosophy as understood in the West, but a legitimate extension of the same enquiry.

Yeah, but no; I may be irredeemably Western and wedded to scientism, but rightly or wrongly, meditation doesn't scratch my epistemic itch. Illusionism may not offer quite the right answer, but for me it is definitely asking the right questions.

Dance With The Devil

14 December 2016

What is evil? In a recent video Peter Dews says it's an idea we're not comfortable with any more; after the shock of Nazism, Hannah Arendt thought we'd spend the rest of the century talking about it; but actually very little was said. We are inclined to talk about conspicuous badness as something that has somehow been wired into some people's nature; but if it's wired in, they had no choice and it can't be really evil.

Simon Baron-Cohen rests on the idea that often what we're really dealing with is a failure of empathy and explains some of the ways modern research is showing it can fall short through genetics or other issues. Dews raises a good objection; that moral goodness and empathy are clearly distinct. Your empathy with your wife might give you the knowledge you need to be really hurtful, for example. Baron-Cohen has an answer to this particular example in his distinction between cognitive and affective empathy - it's one thing to understand another person's feelings and quite another to share them or care about them. But surely there are other ways empathy isn't quite right? Empathy with wicked people might cause us to help them in their wrong-doing, mightn't it? Lack of empathy might allow you to be a good dentist...

Rebecca Roache thinks evil is a fuzzy concept but one that is entwined in our moral discourse and one we should be poorer for abandoning. Describing the Holocaust as 'very bad' wouldn't really do the job.

In my own view, to be evil requires that you understand right and wrong, and choose wrong. This seems impossible, because according to Socrates, anyone who really understands what good is, must want to do it. It has never looked like that in real life, however, where there seem to be plenty of people doing things they know are wrong

Luckily, I recently set out in a nutshell the complete and final theory of ethics. In even briefer form: I think we seek to act effectively out of a kind of roughly existentialist self-assertion. We see that general aims serve our purpose better than limited ones and so choose to act categorically on roughly Kantian reasoning. A sort of empty consequentialism provides us with a calculus by which to choose the generally effective over the particular, but unfortunately the values are often impossible to assess in practice. We therefore fall back on moral codes, set of rules we know are likely to give the best results in the long run.

Now, that suggests Socrates was broadly right; doing the right thing just makes sense. But the system is complicated; there are actually several different principles at work at different levels, and this does give rise to real conflicts.

At the lower levels, these conflicts can give rise to the appearance of evil. Different people may, for example, quite properly have slightly different moral codes that either legitimately reflect cultural difference or are matters of mere convention. Irritable people may see the pursuit of a

different code from their own as automatically evil. Second, there's a genuine tension between any code and the consequentialist rationale that supports it. We follow the code because we can't do the calculus, but every now and then, as in the case of white lies, the utility of breaking the code is practically obvious. People who cling to the code, or people who treat it as flexible, may be seen as evil by those who make different judgements. In fact all these conflicts can be internalised and lead to feelings of guilt and moral doubt; we may even feel a bit bad ourselves.

None of those examples really deserve to be called evil in my view though; that label only applies to higher level problems. The whole thing starts with self-assertion, and some may feel that deliberate wickedness allows them to make a bigger splash. Sure, they may say, I understand that my wrongdoing harms society and thereby indirectly harms my own consequential legacy. But I reckon people will sort of carry things for me; meanwhile I'll write my name on history far more effectively as a master of wickedness than as a useful clerk. This is a mistake, undoubtedly, but unfortunately the virtuous arguments are rather subtle and unexciting, whereas the false reasoning is Byronic and attractive. I reckon that's how the deliberate choice of recognised evil sometimes arises.

Our Unconscious Overlords

18 December 2016

We are in danger of being eliminated by aliens who aren't even conscious, says Susan Schneider. Luckily, I think I see some flaws in the argument.

Humans are probably not the greatest intelligences in the Universe, she suggests; others probably have been going for billions of years longer. Perhaps, but perhaps they have all attained enlightenment and moved on from this plane, leaving us young dummies the cleverest or the only people around?

Schneider thinks the older cultures are likely to be post-biological, having moved on into machine forms of intelligence. This transition may only take a few hundred years, she suggests, to 'judge from the human experience' (Have we transitioned? How did I miss it?). She says transistors are much faster than neurons and computer power is almost indefinitely expandable, so AI will end up much cleverer than us.

Then there may be a problem over controlling these superlatively bright computers, as foreseen by Stephen Hawking, Elon Musk, and Bill Gates. Bill Gates? The man who, by exploiting the monopoly handed to him by IBM was able to impose on us all the crippled memory management of DOS and the endless vulnerabilities of Windows? Well, OK.

Schneider more or less takes it for granted that computation is cogitation, and that faster computation means smarter thinking. It's true that computers have become very good at games we didn't think they could play at all, and she reminds us of some examples. But to take over from human beings, computers need more than just computation. To mention two things, they need agency and intentionality, and to date they haven't shown any capacity at all for either. I don't rule out the possibility of both being generated artificially in future, but the ever-growing ability of computers to do more sums more quickly is strictly irrelevant. Those future people of whom we know nothing may be able to exploit the power of computation – but so can we. If computers are good at winning battles, our computers can fight their computers.

Schneider also takes it for granted that her computational aliens will be hostile and likely to come over and fuck us up good if they ever know we exist. They might, for example, infect our systems with computer viruses (probably not, I think, because without Bill Gates computer viruses probably remained a purely theoretical matter for them). Sending signals, she reckons, is a really bad idea; our radio is already out there but luckily it's faint and easily missed (even by unimaginably super-intelligent aliens). Premature to worry, surely, because even on the most generous view our radio signals are surely less than a hundred light years away so far – not much of a distance in galactic terms. But why would super-intelligent entities behave like witless bullies anyway? Somewhere between benign and indifferent seems a more likely attitude.

To me this whole scenario seems to embody a selective prognosis anyway. The aliens have overcome the limitation of the speed of light, they feed off black holes (no clue, sorry) but they still run on the computation we think is really smart. A hundred years ago no-one would have supposed computation was going to be the dominant technology of our decade, let alone the next million years; maybe by 2116 we'll look back on it the way we fondly remember steam locomotion.

Schneider's most original thought is that her dangerous computational aliens might lack qualia, and do in that sense not be conscious. It seems to me more natural to suppose that acquiring human-style thought would necessarily involve acquiring human-style qualia. Schneider seems to share the Searlian view that qualia have something to do with unknown biological qualities of neural tissue which silicon can never share. She supposes that we might be able to test the proposition. Suppose that for medical reasons we replace parts of a functioning human brain with chips, we might find that qualia are lost.

But how would we know? Ex hypothesi, qualia have no causal powers and so could not cause any change in our behaviour. Even if the qualia vanished, the fact could not be reported. None of the things we say about qualia were caused by qualia; that's one of the bizarre things about them.

I say, if we're going to indulge in this kind of wild speculation, let's go big; I say the super-intelligent aliens will be powered by hyper-computation, a technology that makes our concept of computation look like counting on your fingers; and they'll have not only qualia, but hyper-qualia, experiential phenomenologica whose awesomeness we cannot even speak of. They will be inexpressibly kindly and wise and will be be borne to Earth to visit us on waves beyond our understanding but hugely hyperbolic.

A Christmas Consciousness

24 December 2016

A ghost? Humbug! Yet it was the same face: the very same. Marley in his pigtail, usual waistcoat, tights and boots; the tassels on the latter bristling, like his pigtail, and his coat-skirts, and the hair upon his head. The chain he drew was clasped about his middle. It was long, and wound about him like a tail; and it was made (for Scrooge observed it closely) of cash-boxes, keys, padlocks, ledgers, deeds, and heavy purses wrought in steel. His body was transparent; so that Scrooge, observing him, and looking through his waistcoat, could see the two buttons on his coat behind.

Scrooge had often heard it said that Marley had no bowels, but he had never believed it until now.

No, nor did he believe it even now. Though he looked the phantom through and through, and saw it standing before him; though he felt the chilling influence of its death-cold eyes; and marked the very texture of the folded kerchief bound about its head and chin, which wrapper he had not observed before; he was still incredulous, and fought against his senses.

“You don’t believe in me,” observed the Ghost.

“I don’t,” said Scrooge.

“What evidence would you have of my reality beyond that of your senses?”

“I don’t know,” said Scrooge, “have you read Hume, Jacob? No? You see, to me you’re in the nature of a miracle, something that contradicts all the established understanding of the world. The most parsimonious assumption in such a case, you know, is always that a miraculous event such as your appearance is a delusion.”

“Why do you doubt your senses?”

“Because,” said Scrooge, “a little thing affects them. A slight disorder of the stomach makes them cheats. You may be an undigested bit of beef, a blot of mustard, a crumb of cheese, a fragment of an underdone potato. There’s more of gravy than of grave about you, whatever you are!”

Scrooge was not much in the habit of cracking jokes, nor did he feel, in his heart, by any means waggish then. The truth is, that he tried to be smart, as a means of distracting his own attention, and keeping down his terror; for the spectre’s voice disturbed the very marrow in his bones.

To sit, staring at those fixed glazed eyes, in silence for a moment, would play, Scrooge felt, the very deuce with him. There was something very awful, too, in the spectre’s being provided with an infernal atmosphere of its own. Scrooge could not feel it himself, but this was clearly the case; for though the Ghost sat perfectly motionless, its hair, and skirts, and tassels, were still agitated as by the hot vapour from an oven.

“You see this toothpick?” said Scrooge, returning quickly to the charge, for the reason just assigned; and wishing, though it were only for a second, to divert the vision’s stony gaze from himself.

“I do,” replied the Ghost.

“You are not looking at it,” said Scrooge.

“But I see it,” said the Ghost, “notwithstanding.”

“Well!” returned Scrooge, “I have but to swallow this, and be for the rest of my days persecuted by a legion of goblins, all of my own creation. Humbug, I tell you! humbug!”

”Yet here I am. You see clearly who I once was. I can tell you, if you wish, things that only Jacob Marley could have known: do you doubt it?” said the spirit.

“No. But would those things come from your brain, or from mine? You see, we know now, Jacob, that consciousness is a product of the brain. Have you read Fechner*? No? Well, really, what have you been doing with your evenings, Jacob?”

Scrooge clung desperately to his exposition as the best means of retaining his equanimity in the face of the apparition’s unwavering gaze; at the same time he felt a little pride over his steadfastness. No great city banker, rich and overbearing as he might be, had ever intimidated Scrooge, and he was not about to be cowed by the mere shade of his own dead partner.

“It has been proved not only that consciousness is amenable to scientific investigation, but that it obeys hard mathematical laws; and there’s no manner of doubt that it resides in the brain. Now your brain is dead, Jacob - there’s no question about it - so no possibility of your consciousness persisting arises. Unless we are to be panpsychists, but if that were true, why, I might as well worry about the consciousness of the knocker on the front door!”

“Perhaps you should,” intoned the ghost, unmoved, “I came to effect a moral reformation, but I see I must begin with mereology. You see these ledgers? These frightful deeds chained about me? Why do you suppose I must carry them everywhere?”

“I don’t know… Can it be meant as a punishment, Jacob?” returned Scrooge.

“No; though they are burden enough. These columns of figures, these legal documents, were the tools I used in life to think about my business. They are as much part of my mind as the brain I once had. And though my body is dissolved, they remain, do they not? Is not that part of my mind still growing and flourishing in your counting-house?”

“I keep the books, certainly, Jacob; but that would be a narrow kind of mind…” Scrooge fell silent as he saw the trap he was falling into.

“Narrow indeed, Ebenezer Scrooge!” replied the spirit, “and when did your thoughts last spend a cheerful hour outside the counting house?”

Scrooge looked abashed, but he was thinking quickly.

“You see, spirit,” he resumed, “those account books may retain vestiges of your personality. But ink upon a page is nothing without a brain… very well, then, without a human being, to animate it, to give it significance. Now Cratchit may read those books; or I may. You may not. So if you are

brought here tonight by the revivification of those traces, it is by my mind, and you are indeed nothing more than a phantom of my brain, as I said!"

At this the ghost let out a terrible roar.

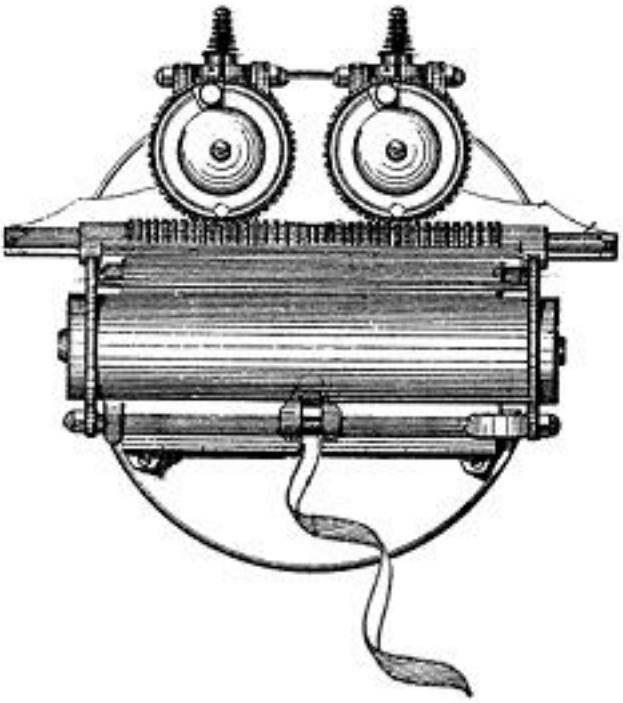
"Prepare yourself, Ebenezer Scrooge!" it thundered, "You shall be visited by three ghosts of disembodied consciousness! The Ghost of Dualism Past; the Ghost of Algorithms Present; and the Ghost of Uploading Yet to Come! Expect the first at the stroke of midnight."

"Humbug!" said Scrooge, excitedly, "Double Humbug, I say! And Humbug on stilts!

** Scrooge was actually rather lucky to get away with that one. He is, of course, is alluding to Fechner's Law, which relates subjective sensation to the logarithm of the intensity of the stimulus, hence at the time a shining example of the new empirical psychology (actually rather too new - I don't think it was published even in German until after A Christmas Carol). Strangely enough, neither Scrooge nor Marley seem aware that Fechner himself believed in a form of panpsychism and had already set out in Das Büchlein vom Leben nach dem Tode (1836) his vision of human life as having three stages: a sleep before birth, normal life in the middle stage, and entry into the general communion of consciousness after death, with the dead still able to exert a helpful influence on the living. He would definitely not have been on Scrooge's side in this discussion*.

Irritating Robots

1 January 2017



Machine learning and neurology; the perfect match?

Of course, there is a bit of a connection already in that modern machine learning draws on approaches which were distantly inspired by the way networks of neurons seemed to do their thing. Now though, it's argued in this interesting piece that machine learning might help us cope with the vast complexity of brain organisation. This complexity puts brain processes beyond human comprehension, it's suggested, but machine learning might step in and decode things for us.

It seems a neat idea, and a couple of noteworthy projects are mentioned: the 'atlas' which mapped words to particular areas of cortex, and an attempt to reconstruct seen faces from fMRI data alone (actually with rather mixed success, it seems). But there are surely a few problems too, as the piece acknowledges.

First, fMRI isn't nearly good enough. Existing scanning techniques just don't provide the neuron-by-neuron data that is probably required, and never will. It's as though the only camera we had was permanently out of focus. Really good processing can do something with dodgy images, but if your lens was rubbish to start with, there are limits to what you can get. This really matters for neurology where it seems very likely that a lot of the important stuff really is in the detail. No matter how good machine learning is, it can't do a proper job with impressionistic data.

We also don't have large libraries of results from many different subjects. A lot of studies really just 'decode' activity in one context in one individual on one occasion. Now it can be argued that that's the best we'll ever be able to do, because brains do not get wired up in identical ways. One of the interesting results alluded to in the piece is that the word 'poodle' in the brain 'lives' near the word 'dog'. But it's hardly possible that there exists a fixed definite location in the brain reserved for the word 'poodle'. Some people never encounter that concept and can hardly have pre-allocated space for it. Did Neanderthals have a designated space for thinking about poodles that presumably was never used throughout the history of the species? Some people might learn of 'poodle' first as a hairstyle, before knowing its canine origin; others, brought up to hard work in their parent's busy grooming parlour from an early age, might have as many words for poodle as the eskimos were supposed to have for snow. Isn't that going to affect the brain location where the word ends up? Moreover, what does it mean to say that the word 'lives' in a given place? We see activity in that location when the word is encountered, but how do we tell whether that is a response to the word, the concept of the word, the concept of poodles,

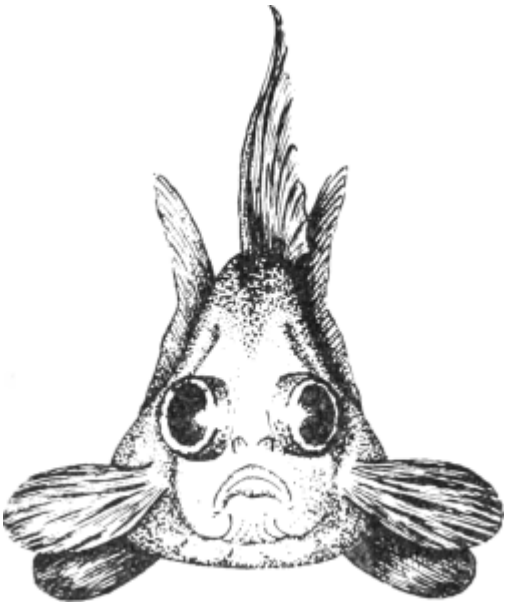
poodles, a particular known poodle, or any other of the family of poodle-related mental entities? Maybe these different items crop up in multiple different places?

Still, we'll never know what can be done if we don't try. One piquant aspect of this is that we might end up with machines that can understand brains, but can never fully explain them to us, both because the complexity is beyond us and because machine learning often works in inscrutable ways anyway. Maybe we can have a second level of machine that explains the first level machines to us – or a pair of machines that each explain the brain and can also explain each other, but not themselves?

It all opens the way for a new and much more irritating kind of robot. This one follows you around and explains you to people. For some of us, some of the time, that would be quite helpful. But it would need some careful constraints, and the fact that it was basically always right about you could become very annoying. You don't want a robot that says "nah, he doesn't really want that, he's just being polite", or "actually, he's just not that into you", let alone "ignore him; he thinks he understands hermeneutics, but actually what he's got in mind is a garbled memory of something else about Derrida he read once in a magazine.

Fish Pain

8 January 2017



Fish don't feel pain, says Brian Key. How does he know? In the deep philosophical sense it remains a matter of some doubt as to whether other human beings really feel pain, and as Key notes, Nagel famously argued that we couldn't know what it was like to be a bat at all, even though we have much more in common with them than with fish. But in practice we don't really feel much doubt that humans with bodies and brains like ours do indeed have similar sensations, and we trust that their reports of pain are generally as reliable as our own. Key's approach extends this kind of practical reasoning. He relies on human reports to identify the parts of the brain involved in feeling pain, and then looks for analogues in other animals.

Key's review of the evidence is interesting; in brief he concludes that it is the cortex that 'does' pain; fish don't have anything that corresponds with human cortex, or any other brain structure that plausibly carries out the same function. They have relatively hard-wired responses to help them escape physical damage, and they have a capacity to learn about what to avoid, but they don't have any mechanism for actually feeling pain with. It is really, he suggests, just anthropomorphism that sees simple avoidance behaviour as evidence of actual pain. Key is rightly stern about anthropomorphism, but I think he could have acknowledged the opposite danger of speciesism. The wide eyes and open mouths of fish, their rigid faces and inability to gesture or scream incline us to see them as stupid, cold, and unfeeling in a way which may not be properly objective.

Still, a careful examination of fish behaviour is a perfectly valid supplementary approach, and Key buttresses his case by noting that pain usually suppresses normal behaviour. Drilling a hole in a human's skull tends to inhibit locomotion and social activity, but apparently doing the same thing to fish does not stop them going ahead with normal foraging and mating behaviour as though nothing had happened. Hard to believe, surely, that they are in terrible pain but getting on with a dancing display anyway?

I think Key makes a convincing case that fish don't feel what we do, but there is a small danger of begging the question if we define pain in a way that makes it dependent on human-style consciousness to begin with. The phenomenology really needs clarification, but defining pain, other than by demonstration, is peculiarly difficult. It is almost by definition the thing we want to avoid feeling, yet we can feel pain without being bothered by it, and we can have feelings we desperately want to avoid which are, however, not pain. Pain may be a tiny twinge accompanying a reflex, an attention-grabbing surge, or something we hold in mind and explore (Dennett, I think, says somewhere that he had been told that examining pain introspectively was one way to make it bearable. On his next dentist visit, he tried it out and found that although the method worked, the effort and boredom involved in continuous close attention to the detailed qualities of his pain was such that he eventually preferred straightforward hurting.) Humans certainly understand pain and can opt to suffer it voluntarily in ways that other creatures

cannot; whether on balance this higher awareness makes our pain more or less bearable is a difficult question in itself. We might claim that imagination and fear magnify our suffering, but being to some degree aware and in control can also put us in a better position than a panicking dog that cannot understand what is happening to it.

Key leans quite heavily on reportable pain; there are obvious reasons for that, but it could be that doing so skews him towards humanity and towards the cortex, which is surely deeply involved in considering and reporting. He dismisses some evidence that pain can occur without a cortex, and therefore happens in the brain stem. His objections seem reasonable, but surely it would be odd if nothing were going on in the brain stem, that 'old brain' we have inherited through evolution, even if it's only some semi-automatic avoidance stuff. The danger is that we might be paying attention to the reportable pain dealt with by the 'talky' part of our minds while another kind is going on elsewhere. We know from such phenomena as blindsight that we can unconsciously 'see' things; could we not also have unconscious pain going on in another part of the brain?

That raises another important question: does it matter? Is unconscious or forgotten pain worth considering - would fish pain be negligible even if it exists? Pain is, more or less, the feeling we all want to avoid, so in one way its ethical significance is obvious. But couldn't those automatic damage avoidance behaviours have some ethical significance too? Isn't damage sort of ethically charged in itself? Key rejects the argument that we should give fish the 'benefit of the doubt', but there is a slightly different argument that being indifferent to apparent suffering makes us worse people even if strictly speaking no pains are being felt.

Consider a small boy with a robot dog; the toy has been programmed to give displays of affection and enjoyment, but if mistreated it also performs an imitation of pain and distress. Now suppose the boy never plays nicely, but obsessively 'tortures' the robot, trying to make it yelp and whine as loudly as possible. Wouldn't his parents feel some concern; wouldn't they tell him that what he was doing was wrong, even though the robot had no real feelings whatever. Wouldn't that be a little more than simple anthropomorphism?

Perhaps we need a bigger vocabulary; 'pain' is doing an awful lot of work in these discussions.

Disbelieving Alieving

15 January 2017

Did my aliefs make me do that? Ezio Di Nucci thinks we need a different way of explaining automatic behaviour. By automatic, he means such things as habits and 'primed' or nudged behaviour; quite a lot of what we do is not the consequence of a consciously pondered decision; yet it manages to be complex and usually sort of appropriate.

Di Nucci quotes a number of examples, but the two he uses most are popcorn and rudeness. Popcorn is an example of a habit. People who were in the habit of eating a lot of popcorn at the cinema still ate the same amount even when they were given stale popcorn. People who did not have the habit ate significantly less if it was stale; as did people who had the habit if they were given it outside the context of a cinema visit (showing that they did notice the difference; it wasn't just that they didn't care about the quality of the popcorn).

Rudeness is an example of 'primed' behaviour. Subjects were exposed to lists of words which either exemplified rudeness or courtesy; those who had been "primed" with rude words went on to interrupt an interviewer more often. There are many examples of this kind of priming effect, but there is now a problem, as Di Nucci notes; many of these experiments have been badly hit by the recent wave of problems with reproducibility in psychological experiments. It is so bad that some might advocate putting the whole topic of priming aside for now until more of the underlying research has been underpinned. Di Nucci proceeds anyway; even if particular studies are invalidated the principle probably captures something true about human psychology.

So how do we explain this kind of mindless behaviour? Di Nucci begins by examining what he characterises as a 'traditional' account given by action theory, particularly citing Donald Davidson. On this view an action is intentional if, described in a particular way, it corresponds with a 'pro' attitude in the agent towards actions with a particular property and the belief that the section has that property. For Di Nucci this comes down to the simple view that an action is intentional if it corresponds with a pre-existing intention.

On that basis, it doesn't seem we can say that the primed subject interrupted unintentionally. It doesn't really seem that we can say that the subject ate stale popcorn unintentionally either. We might try to say that they intended to eat popcorn and did not intend to eat stale popcorn; but since they clearly knew after the first mouthful that it was stale, this doesn't really work. We're left with no distinct account of automatic behaviour.

Di Nucci cites two other arguments. In another experiment, people primed to be helpful are more likely to pick up another person's dropped pen; but if the pen is visibly leaking the priming effect is eliminated. The priming cannot be operating at the same level as the conscious reluctance to get ink on one's fingers or the latter would not erase the former.

Another way to explain the automatic behaviour is to attribute it to false beliefs; but that would imply that the behaviour was to some degree irrational, and neither interrupting nor eating popcorn is actually irrational.

What, then, about aliefs? This is the very handy concept introduced by Tamar Szabo Gendler back in 2008; in essence aliefs are non-conscious beliefs. They explain why we feel nervous walking on a glass floor; we believe we're safe but alieve we're about to fall. They may also explain unconscious bias and many other normal but irrational kinds of behaviour. I rather like this idea; since beliefs and desires are often treated as a pair in studies of intentionality, maybe we could have the additional idea of unconscious desires, or cesires; then we have the nicely alphabetic set of aliefs, beliefs, cesires and desires.

Aliefs look just right to deal with our problems. We might believe the popcorn is stale, but alieve that cinema popcorn is good. We might believe ourselves polite but alieve that interrupting is fine.

Di Nucci suggests three problems. First, he doesn't think aliefs deal with the leaky pen, where the inclination to help disappears altogether, not partially or subject to some conflict. Second, he thinks aliefs end up looking like beliefs. He cites the case of George Clooney fans who are less willing to pay for George's bandanna if it has been washed. Allegedly this because of an irrational alief that a bandanna contains George's essence; but the conscious belief that it has been washed interferes with this. If aliefs can interact with beliefs like this they must be similarly explicit and propositional and so not really different from beliefs. To me this argument doesn't carry much force because there seem to be lots of better ways we could account for the Clooney fan behaviour.

Third, he thinks again that irrationality is a problem; aliefs are supposed by Gendler to be arational, but the two regular examples of automatic behaviour seem rational enough.

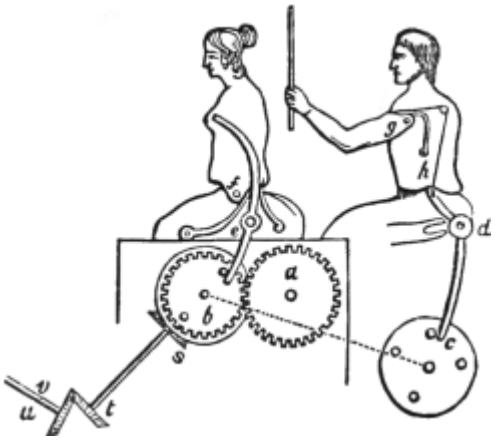
I think aliefs work rather well; if anything they work too well. We can ask about beliefs and challenge them; aliefs are not accessible and we are free to attribute any mad aliefs we like to anyone if they get our explanatory dirty work done. There's perhaps too much of a "get out jail free" about that.

Anyway, if all that is to be rejected, what is the explanation? Di Nucci suggests that automatic behaviour simply fills in when we don't want to waste attention on the detail. Specifically, he suggests they come into play in "Buridan's Ass" cases; where we are faced with choices between alternatives that are equally good or neutral, as often happens. It's pointless to direct our attention to these meaningless choices, so they are left to whatever habits or priming we may be subject to.

That's fine as far as it goes, but I wonder whether it doesn't just retreat a bit; the question was really how well-formed actions come from non-conscious thought. Di Nucci seems in danger of telling us that automatic action is accounted for by the fact that it is automatic.

Robot Insurance

22 January 2017



The EU is not really giving robots human rights; but some of its proposals are cause for concern. James Vincent on the Verge provides a sensible commentary which corrects the rather alarming headlines generated by the European Parliament draft report issued recently. Actually there are several reasons to keep calm. It's only a draft, it's only a report, it's only the Parliament. It is said that in the early days the Eurocrats had to quickly suppress an explanatory leaflet which described the European Economic Community as a bus. The Commission was the engine, the Council was the driver; and the Parliament… was a passenger. Things have moved on since those days, but there's still some truth in that metaphor.

A second major reason to stay calm, according to Vincent, is that the report (quite short and readable, by the way, though rather bitty) doesn't really propose to treat robots as human beings; it mainly addresses the question of liability for the acts of autonomous robots. The sort of personhood it considers is more like the legal personhood of corporations. That's true, although the report does invite trouble by cursorily raising the question of whether robots can be natural persons. Some other parts read strangely to me, perhaps because the report is trying to cover a lot of ground very quickly. At one point it seems to say that Asimov's laws of robotics should currently apply to the designers and creators of robots; I suspect the thinking behind that somewhat opaque idea (A designer must obey orders given it by human beings except where such orders would conflict with the First Law?) has not been fully fleshed out in the text.

What about liability? The problem here is that if a robot does damage to property, turns on its fellow robots (pictured) or harms a human being, the manufacturer and the operator might disclaim responsibility because it was the robot that made the decision, or at least, because the robot's behaviour was not reasonably predictable (I think predictability is the key point: the report has a rather unsatisfactory stab at defining smart autonomous robots, but for present purposes we don't need anything too philosophical - it's enough if we can't tell in advance exactly what the machine will do).

I don't see a very strong analogy with corporate personhood. In that case the problem is the plethora of agents; it simply isn't practical to sue everyone involved in a large corporate enterprise, or even assign responsibility. It's far simpler if you have a single corporate entity that can be held liable (and can also hold the assets of the enterprise, which it may need to pay compensation, etc). In that context the new corporate legal person simplifies the position, whereas with robots, adding a machine person complicates matters. Moreover, in order for the robot's liability to be useful you would have to allow it to hold property which it could use to fund any liabilities. I don't think anyone is currently thinking that roombas should come with some kind of dowry.

Note, however, that corporate personhood has another aspect; besides providing an entity to hold assets and sue or be sued, it typically limits the liability of the parties. I am not a lawyer,

but as I understand it, if several people launch a joint enterprise they are all liable for its obligations; if they create a corporation to run the enterprise, then liability is essentially limited to the assets held by the corporation. This might seem like a sneaky way of avoiding responsibility; would we want there to be a similar get-out for robots? Let's come back to that.

It seems to me that the basic solution to the robot liability problem is not to introduce another person, but to apply strict liability, an existing legal concept which makes you responsible for your product even if you could not have foreseen the consequences of using it in a particular case. The report does acknowledge this principle. In practice it seems to me that liability would partly be governed by the contractual relationship between robot supplier and user: the supplier would specify what could be expected given correct use and reasonable parameters - if you used your robot in ways that were explicitly forbidden in that contract, then liability might pass to you.

Basically, though, that approach leaves responsibility with the robot's builder or supplier, which seems to be in line with what the report mainly advocates. In fact (and this is where things begin to get a bit questionable) the report advocates a scheme whereby all robots would be registered and the supplier would be obliged to take out insurance to cover potential liability. An analogy with car insurance is suggested.

I don't think that's right. Car insurance is a requirement mainly because in the absence of insurance, car owners might not be able to pay for the damage they do. Making third-party insurance obligatory means that the money will always be there. In general I think we can assume by contrast that robot corporations, one way or another, will usually have the means to pay for individual incidents, so an insurance scheme is redundant. It might only be relevant where the potential liability was outstandingly large.

The thing is, there's another issue here. If we need to register our robots and pay large insurance premiums, that imposes a real burden and a significant number of robot projects will not go ahead. Suppose, hypothetically, we have robots that perform crucial work in nuclear reactors. The robots are not that costly, but the potential liabilities if anything goes wrong are huge. The net result might be that nobody can finance the construction of these robots even though their existence would be hugely beneficial; in principle, the lack of robots might even stop certain kinds of plant from ever being built.

So the insurance scheme looks like a worrying potential block on European robotics; but remember we also said that corporate responsibility allows some limitation of liability. That might seem like a cheat, but another way of looking at it is to see it as a solution to the same kind of problem; if investors had to acept unlimited personal liability, there are some kinds of valuable enterprise that would just never be viable. Limiting liability allows ventures that otherwise would have potential downsides too punishing for the individuals involved. Perhaps, then, there actually is an analogy here and we ought to think about allowing some limitation of liability in the case of autonomous robots? Otherwise some useful machines may never be economically viable?

Anyway, I'm not a lawyer, and I'm not an economist, but I see some danger that an EU regime based on this report with registration, possibly licensing, and mandatory insurance, could significantly inhibit European robotics.

Split Brain Not Reproducible

28 January 2017

Classic 'Split Brain' experiments by Sperry and Gazzaniga could not be reproduced, according to Yair Pinto of the University of Amsterdam. Has the crisis of reproducibility claimed an extraordinary new victim?

The original experiments studied patients who had undergone surgery to disconnect the two halves of their cerebrum by cutting the corpus callosum and associated commisures. This was a last-ditch method of controlling epilepsy, and it worked. The patients' symptoms were improved, and they did not generally suffer any major cognitive problems. They felt normal and were able to continue their ordinary lives.

However, under conditions that fed different information to the two halves of the brain, something different showed up. Each half reported only its own data, unaware of what the other half was reporting

On the basis of these remarkable results, it has been widely assumed that these patients have two consciousnesses, or two streams of consciousness, operating separately in each hemisphere but generally in such harmony they don't even notice each other. Now, though, performing similar experiments, Pinto and colleagues unexpectedly failed to get the same results. Is it conceivable that Sperry and Gazzaniga's patients were to some degree playing along with the Doc, giving him the results he evidently wanted? It seems more likely that the effects are just different in different subjects for some reason. In fact, I believe the details of the surgery as well as the pre-existing condition of the patients varies. Recently the tendency has been towards less radical surgery and new drugs; there may not be any more split-brain patients in future.

I don't think that accounts for Pinto's results, though, and they certainly raise interesting problems. In fact the interpretation of split-brain experiments has always been disputed. Although in experimental conditions the subjects seem divided, they don't feel like two people and generally - not invariably -behave normally outside the lab.

Various people have offered interpretations which preserve the essential unity of consciousness, including Michael Tye and Charles E. Marks. One way of looking at it is to point out how unusual the experimental conditions are. Perhaps these conditions (and occasionally others that arise by chance) temporarily induce a bifurcation in an essentially united consciousness. Incidentally, so far as I know no-one has ever tried to repeat the experiments and get the same effects in people with an intact corpus callosum; maybe now and then it would work?

More recently Tim Bayne has proposed a switching model, in which a single consciousness is supported by different physical bits of brain, moving from one physical basis to another without sacrificing its essential unity.

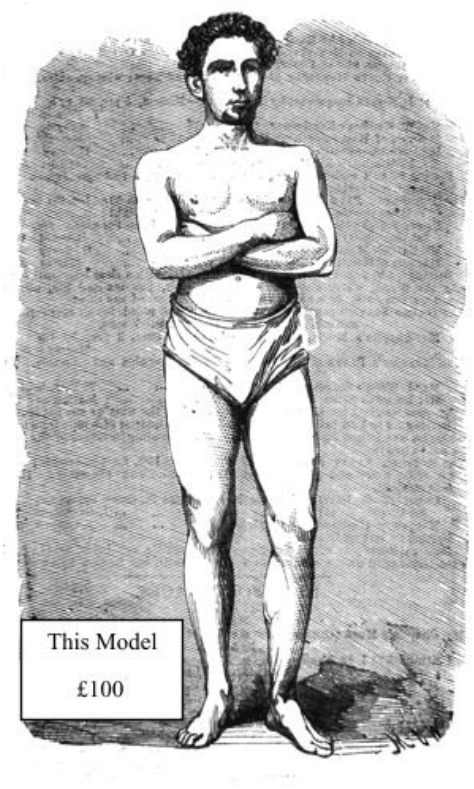
There is, I think, a degree of neurological snobbery about the whole issue, inasmuch as it takes from granted that consciousness is situated in the cerebrum and that therefore dividing the cerebrum can be expected to divide consciousness. Descartes thought the essential unity of consciousness meant it could not reside in parts of the brain which even in normal people is

extensively bifurcated into two hemispheres (actually into a number of quite distinct lobes). We need not follow him in plumping for the pineal gland as the seat of the soul - and we know that the cerebellum can be removed without destroying consciousness). But what about the lowly brain stem? Lizards make do with a brain that in evolutionary terms is little more than our brain stem, yet I'm inclined to believe they have some sense of self, and though they don't have human cogitation, they certainly have consciousness in at least the basic sense. Perhaps we should see consciousness as situated as much down there as up in the cortex. Then we might interpret split-brain patients as having a single consciousness which has some particular lateralised difficulties in pulling together the contributions of its fancy cortical operations.

It might be that different people have different wiring patterns between lower and higher brain that either facilitate or suppress the split-brain effects; but that wouldn't altogether explain why Pinto's results are so different.

Are Our Bodies Our Property?

1 February 2017



In a recent IAI video Anne Phillips says we have lots of rights over our bodies but no absolute authority to do what we like. John Harris doesn't believe bodies are the sort of thing that can really be owned. Brooke Magnanti says that criticism of people's choices about their bodies is framed in moral terms but often stems from disgust rooted in religious or other prejudices rather than in science.

Why should our rights to our bodies be any different to those we have over land, goods or money? I can think of seven arguments one could make.

The first is connection. We cannot be separated from certain parts of our body (at least not while remaining alive), so our rights are inalienable in a way that our rights to other property are not. If we were fixed to a particular tract of land, then perhaps we should naturally have similarly special rights to that land? What of the bits of us we can dispose of? Perhaps those are like crops grown on our special land, which we may sell or give away. There is, nevertheless, a presumption that these things are ours to begin with, which is perhaps why we feel Henrietta Lacks, whose cell culture was taken from her without permission or payment, and still thrives in labs around the world, was badly treated.

The second reason our rights to our bodies are treated as special might be a combination of precautionary hesitation and respect for humanity at large and human dignity. Generally, the law and morality is concerned to give human beings special protection, and if our bodies are seen to be treated with disregard as if they were no more important than old clothes, that might encourage bad consequences for the treatment of human beings generally. But some people might abuse their special rights and be reckless and disrespectful to their own bodies, and so it seems to me that the respect arguments tend more to support the conclusion that bodies should not be property at all – not even our own.

Third is a sense that our body is a protected sphere – no-one else's business. I think it's impossible to deny the powerful appeal of this intuitively. It seems related to the moral view that you can do what you like so long as you don't hurt anyone else. However, I can't see any strong rational case for it apart from the other arguments considered here.

Fourth is a pragmatic argument similar to the one often advanced for private property in general. While there are no absolute rights in play, it is in practice likely that most of the time we will manage our own property best. Society should therefore strive to limit its interventions and give us as much control as it can, consistent with other moral requirements.

The fifth argument is that rights over our own body are absolute and special, perhaps because they were given by God or because they are just a prima facie desirable good in themselves like (according to some) freedom.

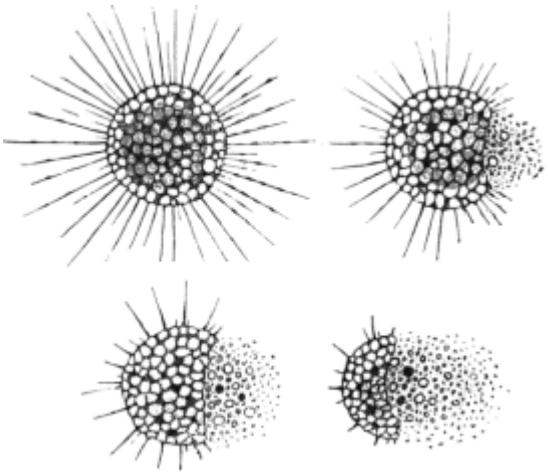
Sixth, we might offer a more technical argument. Rights, we may notice, all appertain to persons. There cannot, then be any ownership of persons because then the recession of rights would have no final root and might be circular or indefinite. If we cannot have ownership rights over persons, then it is a natural extension to limit the ownership of rights over bodies.

The seventh reason suggests that a special ownership of our own bodies is invariably part of the implied social contract in place in any organised society. The deeper reasons may be matters of psychology or anthropology, but ultimately this is simply a human universal.

I think there's at least something in all seven, but I still don't think they completely over-ride society's right to regulate the matter given sufficient justification.

Consciousness = Entropy

5 February 2017



Is consciousness a matter of entropy in the brain? An intriguing paper by R. Guevara Erra, D. M. Mateos, R. Wennberg, and J.L. Perez Velazquez says

> *normal wakeful states are characterised by the greatest number of possible configurations of interactions between brain networks, representing highest entropy values.*

What the researchers did, broadly, is identify networks in the brain that were operative at a given time, and then work out the number of possible configurations these networks were capable of. In general, conscious states were associated with states with high numbers of possible configurations - high levels of entropy.

That makes me wrinkle my forehead a bit because it doesn't fit well with my layman's grasp of the concept of entropy. In my mind entropy is associated with low levels of available energy and an absence of large complex structure. Entropy always increases, but can decrease locally, as in the case of the complex structures of life, by paying for the decrease with a bigger increase elsewhere; typically by using up available energy. On this view, conscious states - and high levels of possible configurations - look like they ought to be low entropy; but evidently the reverse is actually the case. The researchers also used a separate measure of complexity, one with strong links to information content, which is clearly an interesting angle in itself

Of the nine subjects, three were epileptic, which allowed comparisons to be made with seizure states as well as those of waking and sleeping states. Interestingly, REM sleep showed relatively high entropy levels, which intuitively squares with the idea that dreaming resembles waking a little more than fully unconscious states - though I think the equation of REM sleep with dreaming is not now thought to be as perfect as it once seemed.

One acknowledged weakness in the research is that it was not possible to establish actual connection. The assumed networks were therefore based on synchronisation instead. However, synchronisation can arise without direct connection, and absence of synchronisation is not necessarily proof of the absence of connection.

Still, overall, the results look good and the picture painted is intuitively plausible. Putting all talk of entropy and Lempel-Ziv aside, what we're really saying is that conscious states fall in the middle of a notional spectrum: at one end of this spectrum is chaos, with neutrons firing randomly; at the other we have then all firing simultaneously in indissoluble lockstep.

There is an obvious resemblance here to the Integrated Information Theory (IIT) which holds that consciousness arises once the quantity of information being integrated passes a value known as Phi. In fact, the authors of the current paper situate it explicitly within the context of earlier work which suggests that the general principle of natural phenomena is the maximisation of

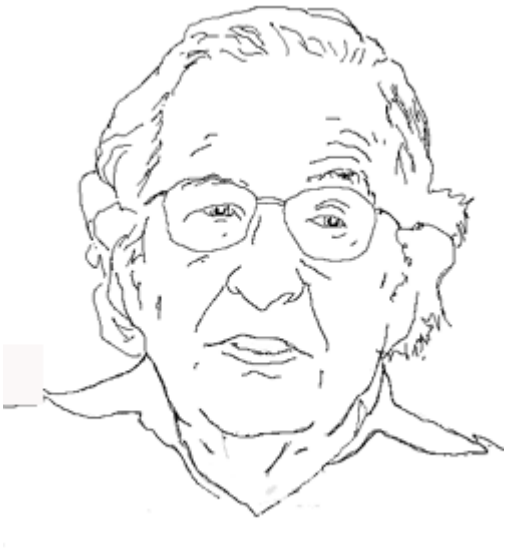
information transfer. The read-across from the new results into terms of information processing is quite clear. The authors do acknowledge IIT, but just barely; they may be understandably worried that their new work could end up interpreted as mere corroboration for IIT.

My main worry about both is that they are very likely true but may not be particularly enlightening. As a rough analogy, we might discover that the running of an internal combustion engine correlates strongly with raised internal temperature states. The absence of these states proves to be a pretty good practical guide to whether the engine is running, and we're tempted to conclude that raised temperature is the same as running. Actually, though, raising the temperature artificially does not make the engine run, and there is actually a complex story about running in which raised temperatures are not really central. So it might be that high entropy is characteristic of conscious states without that telling us snything useful about how those states really work.

But I evidently don't really get entropy, so I might be missing the true significance of all this.

Chomsky's Mysterianism

7 February 2017



Or perhaps Chomsky's endorsement of Isaac Newton's mysterianism. We tend to think of Newton as bringing physics to a triumphant state of perfection, one that lasted until Einstein, and with qualifications, still stands. Chomsky says that in fact Newton shattered the ambitions of mechanical science, which have never recovered; and in doing so he placed permanent limits on the human mind. He quotes Hume;

> *While Newton seemed to draw off the veil from some of the mysteries of nature, he shewed at the same time the imperfections of the mechanical philosophy; and thereby restored her ultimate secrets to that obscurity, in which they ever did and ever will remain*.

What are they talking about? Above all, the theory of gravity, which relies on the unexplained notion of action at a distance. Contemporary thinkers regarded this as nonsensical, almost logically absurd: how could object A affect object B without contacting it and without and intermediating substance? Newton, according to Chomsky, agreed in essence; but defended himself by saying that there was nothing occult in his own work, which stopped short where the funny stuff began. Newton, you might say, described gravity precisely and provided solid evidence to back up his description; what he didn't do at all was explain it.

The acceptance of gravity, according to Chomsky, involved a permanent drop in the standard of intelligibility that scientific theories required. This has large implications for the mind it suggests there might be matters beyond our understanding and provides a particular example. But it may well be that the mind itself is, or involves, similar intractable difficulties.

Chomsky reckons that Darwin reinforced this idea. We are not angels, after all, only apes; all other creatures suffer cognitive limitations; why should we be able to understand everything? In fact our limitations are as important as our abilities in making us what we are; if we were bound by no physical limitations we should become shapeless globs of protoplasm instead of human beings, and the same goes for our minds. Chomsky distinguishes between problems and mysteries. What is forever a mystery to a dog or rat may be a solvable problem for us, but we are bound to have mysteries of our own.

I think some care is needed over the idea of permanent mysteries. We should recognise that in principle there may be several things that look mysterious, notably the following.

1. Questions that are, as it were, out of scope: not correctly definable as questions at all: these are answerable even by God.
2. Mysterian mysteries; questions that are not in themselves unanswerable, but which are permanently beyond the human mind.
3. Questions that are answerable by human beings, but very difficult indeed.

4. Questions that would be answerable by human beings if we had further information which we (a) either just don't happen to have, or which (b) we could never have in principle.

I think it's just an assumption that the problem of mind, and indeed, the problem of gravity, fall into category 2. There has been a bit of movement in recent decades, I think, and the possibility of 3 or 4(a) remains open.

I don't think the evolutionary argument is decisive either. Implicitly Chomsky assumes an indefinite scale of cognitive abilities matched by an indefinite scale of problems. Creatures that are higher up the first get higher up the second, but there's always a higher problem. Maybe, though, there's a top to the scale of problems? Maybe we are already clever enough in principle to tackle them all.

If this seems optimistic, think of Chomsky the Lizard, millions of years ago. Some organisms, he opines, can stick their noses out of the water. Some can leap out, briefly. Some crawl out on the beach for a while. Amphibians have to go back to reproduce. But all creatures have a limit to how far they can go from the sea. We lizards, we've got legs, lungs, and the right kind of eggs; we can go further than any other. That does not mean we can go all over the island. Evolution guarantees that there will always be parts of the island we can't reach.

Well, depending on the island, there may be inaccessible parts, but that doesn't mean legs and lungs have inbuilt limits. So just because we are products of evolution, it doesn't mean there are necessarily questions of type 2 for us.

Chomsky mocks those who claim that the idea of reducing the mind to activity of the brain is new and revolutionary; it has been widely espoused for centuries, he says. He mentions remarks of Locke which I don't know, but which resemble the famous analogy of Leibniz's mill.

> *If we imagine that there is a machine whose structure makes it think, sense, and have perceptions, we could conceive it enlarged, keeping the same proportions, so that we could enter into it, as one enters a mill. Assuming that, when inspecting its interior, we will find only parts that push one another, and we will never find anything to explain a perception*.

The thing about that is, we'll never find anything to explain a mill, either. Honestly, Gottfried, all I see is pieces of wood and metal moving around; none of them have any milliness? How on earth could a collection of pieces of wood - just by virtue of being *arranged* in some functional way, you say - acquire completely new, distinctively *molational* qualities?

Scott's Aliens Return

13 February 2017

Scott Bakker's alien consciousnesses are back, and this time it's peer reviewed. We talked about their earlier appearance in the Three Pound Brain a while ago, and now a paper in the JCS sets out a new version.

The new paper foregrounds the idea of using hypothetical aliens as a forensic tool for going after the truth about our own minds; perhaps we might call it xenophenomenology. That opens up a large speculative space, though it's one which is largely closed down again here by the accompanying assumption that our aliens are humanoid, the product of convergent evolution. In fact, they are now called Convergians, instead of the Thespians of the earlier version.

In a way, this is a shame. On the one hand, one can argue that to do xenophenomenology properly is impractical; it involves consideration of every conceivable form of intelligence, which in turn requires an heroic if not god-like imaginative power which few can aspire to (and which would leave the rest of us struggling to comprehend the titanic ontologies involved anyway). But if we could show that any possible mind would have to be x, we should have a pretty strong case for xism about human beings. In the present case not much is said about the detailed nature of the Convergian convergence, and we're pretty much left to assume that they are the same as us in every important respect. This means there can be no final reveal in which - aha! - it turns out that all this is true of humans too! Instead it's pretty clear that we're effectively talking about humans all along.

Of course, there's not much doubt about the conclusion we're heading to here, either: in effect the Blind Brain Theory (BBT). Scott argues that as products of evolution our minds are designed to deliver survival in the most efficient way possible. As a result they make do with a mere trickle of data and apply cunning heuristics that provide a model of the world which is quick and practical but misleading in certain important respects. In particular, our minds are unsuited to metacognition - thinking about thinking - and when we do apply our minds to themselves the darkness of those old heuristics breeds monsters: our sense of our selves as real, conscious agents and the hard problems of consciousness.

This seems to put Scott in a particular bind so far as xenophenomenology is concerned. The xenophenomenological strategy requires us to consider objectively what alien minds might be like; but Scott's theory tells us we are radically incapable of doing so. If we are presented with any intelligent being, on his view those same old heuristics will kick in and tell us that the aliens are people who think much like us. This means his conclusion that Convergians would surely suffer the same mental limitations as us appears as merely another product of faulty heuristics, and the assumed truth of his conclusion undercuts the value of his evidence.

Are those heuristics really that dominant? It is undoubtedly true that through evolution the brains of mammals and other creatures took some short cuts, and quite a few survive into human cognition, including some we're not generally aware of. That seems to short-change the human mind a bit though; in a way the whole point of it is that it isn't the prisoner of instinct and habit. When evolution came up with the human brain, it took a sort of gamble; instead of

equipping it with good fixed routines, it set it free to come up with new ones, and even over-ride old instincts. That gamble paid off, of course, and it leaves us uniquely able to identify and overcome our own limitations.

If it were true that our view of human conscious identity were built in by the quirks of our heuristics, surely those views would be universal, but they don't seem to be. Scott suggests that, for example, the two realms of sky and earth naturally give rise to a sort of dualism, and the lack of visible detail in the distant heavens predisposes Convergians (or us) to see it as pure and spiritual. I don't know about that as a generalisation across human cultures (didn't the Greeks, for one thing, have three main realms, with the sea as the third?). More to the point, it's not clear to me that modern western ways of framing the problems of the human mind are universal. Ancient Egyptians divided personhood into several souls, not just one. I've been told that in Hindu thought the question of dualism simply never arises. In Shinto the line between the living and the material is not drawn in quite the Western way. In Buddhism human consciousness and personhood have been taken to be illusions for many centuries. Even in the West, I don't think the concept of consciousness as we now debate it goes back very far at all - probably no earlier than the nineteenth century, with a real boost in the mid-twentieth (in Italian and French I believe one word has to do duty for both 'consciousness' and 'conscience', although we mustn't read too much into that). If our heuristics condemn us to seeing our own conscious existence in a particular way, I wouldn't have expected that much variation.

Of course, there's a difference between what vividly seems true and what careful science tells us is true; indeed if the latter didn't reveal the limitations of our original ideas this whole discussion would be impossible. I don't think Scott would disagree about that; and his claim that our cognitive limitations have influenced the way we understand things is entirely plausible. The question is whether that's all there is to the problems of consciousness.

As Scott mentions here, we don't just suffer misleading perceptions when thinking of ourselves; we also have dodgy and approximate impressions of physics. But those misperceptions were not Hard problems; no-one had ever really doubted that heavier things fell faster, for example. Galileo sorted several of these basic misperceptions out simply by being a better observer than anyone previously and paying more careful attention. We've been paying careful attention to consciousness for some time now, and arguably it just gets worse.

In fairness that might rather short-change Scott's detailed hypothesising about how the appearance of deep mystery might arise for Convergians; those, I think, are the places where xenophenomenology comes close to fulfilling its potential.

Fear and HOROR

19 February 2017

Emotions like fear are not something inherited from our unconscious animal past. Instead, they arise from the higher-order aspects that make human thought conscious. That (if I've got it right) is the gist of an interesting paper by LeDoux and Brown.

A mainstream view of fear (the authors discuss fear in particular as a handy example of emotion, on the assumption that similar conclusions apply to other emotions) would make it a matter of the limbic system, notably the amygdala, which is known to be associated with the detection of threats. People whose amygdalas have been destroyed become excessively trusting, for example - although as always things are more complicated than they seem at first and the amygdalas are much more than just the organs of 'fear and loathing'. LeDoux and Brown would make fear a cortical matter, generated only in the kind of reflective consciousness possessed by human beings.

One immediate objection might be that this seems to confine fear to human beings, whereas it seems pretty obvious that animals experience fear too. It depends, though, what we mean by 'fear'. LeDoux and Brown would not deny that animals exhibit aversive behaviour, that they run away or emit terrified noises; what they are after is the actual feeling of fear. LeDoux and Brown situate their concept of fear in the context of philosophical discussion about phenomenal experience, which makes sense but threatens to open up a larger can of worms - nothing about phenomenal experience, including its bare existence, is altogether uncontroversial. Luckily I think that for the current purposes the deeper issues can be put to one side; whether or not fear is a matter of ineffable qualia we can probably agree that humanly conscious fear is a distinct thing. At the risk of begging the question a bit we might say that if you don't know you're afraid, you're not feeling the kind of fear LeDoux and Brown want to talk about.

On a traditional view, again, fear might play a direct causal role in behaviour. We detect a threat, that causes the feeling of fear, and the feeling causes us to run away. For LeDoux and Brown, it doesn't work like that. Instead, while the threat causes the running away, that process does not in itself generate the feeling of fear. Those sub-cortical processes, along with other signals, feed into a separate conscious process, and it's that that generates the feeling.

Another immediate objection therefore might be that the authors have made fear an epiphenomenon; it doesn't do anything. Some, of course, might embrace the idea that all conscious experience is epiphenomenal; a by-product whose influence on behaviour is illusory. Most people, though, would find it puzzling that the brain should go to the trouble of generating experiences that never affect behaviour and so contribute nothing to survival.

The answer here, I think, comes from the authors' view of consciousness. They embrace a higher-order theory (HOT). HOTs (there are a number of variations) say that a mental state is conscious if there is another mental state in the same mind which is about it - a Higher Order

Representation (HOR); or to put it another way, being conscious is being aware that you're aware. If that is correct, then fear is a natural result of the application of conscious processes to certain situations, not a peculiar side-effect.

HOTs have been around for a long time: they would always get a mention in any round-up of the contenders for an explanation of consciousness, but somehow it seems to me they have never generated the little bursts of excitement and interest that other theories have enjoyed. LeDoux and Brown suggest that other theories of emotion and consciousness either are 'first -order' theories explicitly, or can be construed as such. They defend the HOT concept against one of the leading objections, which is that it seems to be possible to have HORs of non-existent states of awareness. In Charles Bonnet, syndrome, for example, people who are in fact blind have vivid and complex visual hallucinations. To deal with this, the authors propose to climb one order higher; the conscious awareness, they suggest, comes not from the HOR of a visual experience but from the HOR of a HOR: a HOROR, in fact. There clearly is no theoretical limit to the number of orders we can rise to, and there's some discussion here about when and whether we should call the process introspection.

I'm not convinced by HOTs myself. The authors suggest that single-order theory implies there can be conscious states of which we are not aware, which seems sort of weird: you can feel fear and not know you're feeling fear? I think there's a danger here of equivocating between two senses of 'aware'. Conscious states are states of awareness, but not necessarily states we are aware of; something is in awareness if we are conscious; but that's not to say that the something includes our awareness itself. I would argue, contrarily, that there must be states of awareness with no HOR; otherwise, what about the HOR itself? If HORs are states of awareness themselves, each must have its own HOR, and so on indefinitely. If they're not, I don't see how the existence of an inert representation can endow the first-order state with the magic of consciousness.

My intuitive unease goes a bit wider than that, too. The authors have given a credible account of a likely process, but on this account fear looks very like other conscious states. What makes it different - what makes it actually fearful? It seems possible to imagine that I might perform the animal aversive behaviour, experience a conscious awareness of the threat and enter an appropriate conscious state without actually feeling fear. I have no doubt more could be said here to make the account more plausible and in fairness LeDoux and Brown could well reply that nobody has a knock-down account of phenomenal experience, with their version offering a lot more than some.

In fact, even though I don't sign up for a HOT I can actually muster a pretty good degree of agreement nonetheless. Nobody, after all, believes that higher order mental states don't exist (we could hardly be discussing this subject if they didn't). In fact, although I think consciousness doesn't require HORs, I think they are characteristic of its normal operation and in fact ordinary consciousness is a complex meld of states of awareness at several different levels. If we define fear the way LeDoux and Brown do, I can agree that they have given a highly plausible account of how it works without having to give up my belief that simple first-order consciousness is also a thing.

Information And Experience

25 February 2017

You can't build experience out of mere information. Not, at any rate, the way the Integrated Information Theory (IIT) seeks to do it. So says Garrett Mindt in a forthcoming paper for the JCS.

'Information' is notoriously a slippery term, and much depends on how you're using it. Commonly people distinguish the everyday meaning, which makes information a matter of meaning or semantics, and the sense defined by Shannon, which is statistical and excludes meaning, but is rigorous and tractable.

It is a fairly common sceptical claim that you cannot get consciousness, or anything like intentionality or meaning, out of Shannon-style information. Mindt describes in his paper a couple of views that attack IIT on similar grounds. One is by Cerullo, who says:

> *'Only by including syntactic, and most importantly semantic, concepts can a theory of information hope to model the causal properties of the brain…'*

The other is by Searle, who argues that information, correctly understood, is observer dependent. The fact that this post, for example, contains information depends on conscious entities interpreting it as such, or it would be mere digital noise. Since information, defined this way, requires consciousness, any attempt to derive consciousness from it must be circular.

Although Mindt is ultimately rather sympathetic to both these cases, he says they fail because they assume that IIT is working with a Shannonian conception of information: but that's not right. In fact IIT invokes a distinct causal conception of information as being 'a difference that makes a difference'. A system conveys information, in this sense, if it can induce a change in the state of another system. Mindt likens this to the concept of information introduced by Bateson.

Mindt makes the interesting point that Searle and others tend to carve the problem up by separating syntax from semantics; but it's not clear that semantics is required for hard-problem style conscious experience (in fact I think the question of what, if any, connection there is between the two is puzzling and potentially quite interesting). Better to use the distinction favoured by Tononi in the context of IIT, between extrinsic information - which covers both syntax and semantics - and intrinsic, which covers structure, dynamics, and phenomenal aspects.

Still, Mindt finds IIT vulnerable to a slightly different attack. Even with the clarifications he has made, the theory remains one of structure and dynamics, and physicalist structure and dynamics just don't look like the sort of thing that could ever account for the phenomenal qualities of experience. There is no theoretical bridge arising from IIT that could take us across the explanatory gap.

I think the case is well made, although unfortunately it may be a case for despair. If this objection stands for IIT then it most likely stands for all physicalist theories. This is a little depressing because on one point of view, non-physicalist theories look unattractive. From that perspective, coming up with a physical explanation of phenomenal experience is exactly the

point of the whole enquiry; if no such explanation is possible, no decent answer can ever be given.

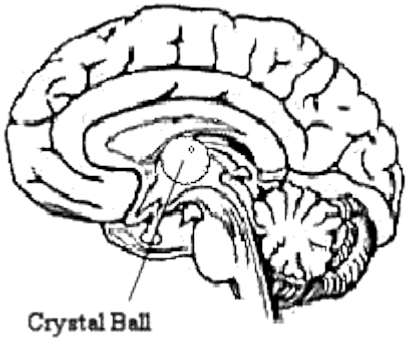
It might still be the case that IIT is the best theory of its kind, and that it is capable of explaining many aspects of consciousness. We might even hope to squeeze the essential Hard Problem to one side. What if IIT could never explain why the integration of information gives rise to experience, but could explain everything, or most things, about the character of experience? Might we not then come to regard the Hard Problem as one of those knotty tangles that philosophers can mull over indefinitely, while the rest of us put together a perfectly good practical understanding of how mind and brain work?

I don't know what Mindt would think about that, but he rounds out his case by addressing one claimed prediction of IIT; namely that if a large information complex is split, the attendant consciousness will also divide. This looks like what we might see in split-brain cases, although so far as I can see, nobody knows whether split-brain patients have two separate sets of phenomenal experiences, and I'm not sure there's any way of testing the matter. Mindt points out that the prediction is really a matter of 'Easy Problem' issues and doesn't help otherwise: it's also not an especially impressive prediction, as many other possible theories would predict the same thing.

Mindt's prescription is that we should go back and have another try at that definition of information; without attempting to do that he smiles on dual aspect theories. I'm afraid I am left scowling at all of them: as always in this field the arguments *against* any idea seem so much better than the ones *for*.

Feelings First

11 December 2005



The idea that quantum physics might have something to do with consciousness crops up in a number of theories: perhaps most prominently in the version put forward by Penrose and Hameroff, but quite regularly elsewhere - the views of Huping Hu and Maoxin Wu, for example, have been discussed here recently. I sometimes find the motivation of this line of reasoning a little unclear: why exactly do we need to invoke quantum physics, anyway? Sometimes it seems that the quantum theorists are merely hoping that there will turn out to be one solution to two otherwise unconnected mysteries: not actually a very good bet.

Earlier this year, Maurits van den Noort, from the Department of Cognitive Neuroscience at the University of Bergen, came up with yet more reflections on the relevance of quantum effects, but this time there is a difference, in that he puts forward two studies which seem to offer empirical evidence that something unexpected and possibly quantumish is going on in the unconscious emtional reactions of human beings, as revealed by scans and EEG readings. A book is apparently in the pipeline.

Van den Noort gives an admirably clear summary of just how quantum physics might be thought relevant to consciousness:

> *First, quantum coherence (e.g. Bose-Einstein condensation) is a possible physical basis for 'binding' or unity of consciousness (Marshall, 1989). Second, non-local entanglements (e.g. 'Einstein-Podolsky-Rosen correlations') serve as a potential basis for associative memory and non-local emotional interpersonal connection. Third, quantum superposition of information provides a basis for preconscious and subconscious processes, dreams and altered states. Finally, quantum state reduction (quantum computation) serves as a possible physical mechanism for the transition from preconscious processes to consciousness (Penrose, 1989; 1994).*

He says that quantum state reductions may send quantum information "back in time", and it is the scope for non-linear information processing which particularly interests him. Acknowledging that a smooth chronological sequence is one of the more salient properties of subjective experience, he points instead to unconscious information processing, and especially to the generation of emotional responses.

The two experiments quoted by Van den Noort involved showing subjects a series of images: some neutral, some emotionally stimulating. (In fact the stimulating images were either erotic or violent, which seems to short-change the emotional life of human beings somewhat, though no doubt it helped to guarantee a clear reaction). Both experiments, whether based on scanner or EEG, produced the same extraordinary result: it was possible to detect an emotional reaction to emotional stimuli just before they actually occurred, even though subjects had no way of knowing whether the next image in the random series was going to be emotionally charged or neutral.

Van den Noort suggests that a likely explanation is that the unconscious mental processes at work here are powered by non-linear quantum information processing. He sees unconscious information processing as a kind of fast but rough system for generating emotional reactions which colour our behaviour, although they remain subject to the slower and more thorough examination of events which takes place on a conscious level.

These results are so remarkable (I'm not sure I can get my head round the implications for Libet's similarly counter-intuitive findings at all) that it is tempting to assume that something must have been wrong with the experimental set-up: perhaps the images weren't truly randomised, or the subjects were somehow able to pick up subtle clues of which the experimenters remained unaware. I don't know of any flaws like this, but I am inclined to think something must be wrong somewhere.

First, the ability to respond to future events, even in the shortest term, is such a useful capacity it seems strange our nervous system doesn't make better use of it. Our system is certainly set up to produce fast responses in certain cases: when we touch a hot surface, for example, the nerve signal does not even have to spend the small amount of extra time needed to go all the way to the brain: instead, a short loop triggers a hard-wired evasive reponse. If the brain could register emotional events before they happened (and burning your hand, or even nearly burning it, surely does generate some emotion) it ought to be able to set off the reflex evasive action before you actually make contact. I realise that evolution does not necessarily deliver every possible good idea: but this one is so simple and so valuable it seems suprising, given Van den Noort's findings, that pre-emptive reflexes don't exist. Second, this kind of experiment, moitoring reaction times to a sequence of images, sounds like a pretty typical piece of psychological research: it seems a little odd that some other researcher hasn't come across this chronological curiosity before.

Dennett Recants

4 March 2017

Yes, Dennett has recanted. Alright, he hasn't finally acknowledged that Jesus is his Lord and Saviour. He hasn't declared that qualia are the real essence of consciousness after all. But his new book From Bacteria to Bach and Back does include a surprising change of heart.

The book is big and complex: to be honest it's a bit of a ragbag (and a bit of a curate's egg). I'll give it the fuller review it deserves another time, but it seems to be worth addressing this interesting point separately. The recantation arises from a point on which Dennett has changed his mind once before. This is the question of homunculi. Homunculi are 'little people' and the term is traditionally used to criticise certain kinds of explanation, the kind that assume some module in the brain is just able to do everything a whole person could do. Those modules are effectively 'little people in your head', and they require just as much explanation as your brain did in the first place. At some stage many years ago, Dennett decided that homunculi were alright after all, on certain conditions. The way he thought it could work was an hierarchy of ever stupider homunculi. Your eyes deliver a picture to the visual homunculus, who sees it for you; but we don't stop there; he delivers it to a whole group of further colleagues; line-recognising homunculi, colour-recognising homunculi, and so on. Somewhere down the line we get to an homunculus whose only job is to say whether a spot is white or not-white. At that point the function is fully computable and our explanation can be cashed out in entirely non-personal, non-mysterious, mechanical terms. So far so good, though we might argue that Dennett's ever stupider routines are not actually homunculi in the correct sense of being complete people; they're more like 'black boxes', perhaps, a stage of a process you can't explain yet, but plan to analyse further.

Be that as it may, he now regrets taking that line. The reason is that he no longer believes that neurons work like computers! This means that even at the bottom level the reduction to pure computation doesn't quite work. The reason for this remarkable change of heart is that Terrence Deacon and others have convinced Dennett that the nature of neurons as entities with metabolism and a lifecycle is actually relevant to the way they work. The fact that neurons, at some level, have needs and aims of their own may ground a kind of 'nano-intentionality' that provides a basis for human cognition.

The implications are large; if this is right then surely, computation alone cannot give rise to consciousness! You need metabolism and perhaps other stuff. That Dennett should be signing up to this is remarkable, and of course he has a get-out. This is that we could still get computer consciousness by simulating an entire brain and reproducing every quirk of every neuron. For now, that is well beyond our reach - and it may always be, though Dennett speaks with misplaced optimism about Blue Brain and other projects. In fact, I don't think the get-out works even on a theoretical level; simulations always leave out some aspect of the thing simulated, and if this biological view is sound, we can never be sure that we haven't left out something important.

But even if we allow the get-out to stand this is a startling change, and I've been surprised to see that no review of the book I've seen even acknowledges it. Does Dennett himself even appreciate quite how large the implications are? It doesn't really look as if he does. I would guess he thinks of the change as merely taking him a bit closer to, say, the evolution-based perspective of Ruth Millikan, not at all an uncongenial direction for him. I think, however, that he's got more work to do. He says:

> *The brain is certainly not a digital computer running binary code, but it is still a kind of computer...*

Later on, however, he rehashes the absurd but surely digitally-computational view he put forward in Consciousness Explained:

> *You can simulate a virtual serial machine on a parallel architecture, and that's what the brain does... and virtual parallel machines can be implemented on serial machines...*

This looks pretty hopeless in itself, by the way. You can do those things if you don't mind doing something really egregiously futile. You want to 'simulate' a serial machine on a parallel architecture? Just don't use more than one of its processors. The fact is, parallel and serial computing do exactly the same job, run the same algorithms, and deliver the same results. Parallel processing by computers is just a practical engineering tactic, of no philosophical interest whatever. When people talk about the brain doing parallel processing they are talking about a completely different and much vaguer idea and often confusing themselves in the process. Why on earth does Dennett think the brain is simulating serial processing on a parallel architecture, a textbook example of pointlessness?

> *It is true that the brain's architecture is massively parallel... but many of the brain's most spectacular activities are (roughly) serial, in the so-called stream of consciousness, in which ideas, or concepts or thoughts float by not quite in single file, but through a Von Neumann bottleneck of sorts...*

It seems that Dennett supposes that only serial processing can deliver a serially coherent stream of consciousness, but that is just untrue. On display here too is Dennett's bad habit of using 'Von Neumann' as a synonym for 'serial'. As I understand it the term "Von Neumann Architecture" actually relates to a long-gone rivalry between very early computer designs. Historically the Von Neumann design used the same storage for programs and data, while the more tidy-minded Harvard Architecture provided separate storage. The competition was resolved in Von Neumann's favour long ago and is as dead as a doornail. It simply has no relevance to the human brain: does the brain have a Von Neumann or Harvard architecture? The only tenable answer is 'no'.

Anyway, whatever you may think of that, if Dennett now says the brain is not a digital computer, he just cannot go on saying it has a Von Neumann architecture or simulates a serial processor. Simple consistency requires him to drop all that now - and a good thing too. Dennett has to find a way of explaining the stream of consciousness that doesn't rely on concepts from digital computing. If he's up for it, we might get something really interesting - but retreat to the comfort zone must look awfully appealing at this stage. There is, of course, nothing shameful in changing your mind; if only he can work through the implications a bit more thoroughly, Dennett will deserve a lot of credit for doing so.

More another time.

Building **Consciousness**

8 March 2017

Owen Holland delivered a blast of old-fashioned optimism in a recent video: let's just build a conscious robot!

It's a short video so Holland doesn't get much chance to back up his prediction that if you're under thirty you will meet a conscious robot. He voices feelings which I suspect are common on the engineering and robotics side of the house, if not usually expressed so clearly: why don't we just get on and put a machine together to do this? Philosophy, psychology, all that airy fairy stuff is getting us nowhere; we'll learn more from a bad robot than twenty papers on qualia.

His basic idea is that we're essentially dealing with an internal model of the world. We can now put together robots with an increasingly good internal modelling capability (and we can peek into those models); why not do that and then add new faculties and incremental improvements till we get somewhere?

Yeah, but. The history of AI is littered with projects that started like this and ran into the sand. In particular, the idea that it's all about an internal model may be a radical mis-simplification. We don't just picture ourselves in the world, we picture ourselves picturing ourselves. We can spend time thinking just about the concept of consciousness - how would that appear in a model? In general, our conscious experience is complex and elusive, and cannot accurately be put on a screen or described on a page (though generations of novelists have tried everything they can think of).

The danger when we start building is that the first step is wrong and already commits us to a wrong path. Maybe adding new abilities won't help. Perhaps our ability to model the world is just one aspect of a deeper and more general faculty that we haven't really grasped yet; building in a fixed spatial modeller might turn us away from that right at the off. Instead of moving incrementally towards consciousness we might end up going nowhere (although there should be some pretty cool robots along the way).

Still, without some optimism we'll certainly get nowhere anyway.

Greatest Hits

13 March 2017

Give up on real comprehension, says Daniel Dennett *in From Bacteria to Bach and Back*: commendably honest but a little discouraging to the reader? I imagine it set out like the warning above the gates of Hell: 'Give up on comprehension, you who turn these pages.' You might have to settle for acquiring some competences.

What have we got here? In this book, Dennett is revisiting themes he developed earlier in his career, retelling the grand story of the evolution of minds. We should not expect big new ideas or major changes of heart (but see last week's post for one important one). It would have been good at this stage to have a distillation; a perfect, slim little volume presenting a final crystalline formulation of what Dennett is all about. This isn't that. It's more like a sprawling Greatest Hits album. In there somewhere are the old favourites that will always have the fans stomping and shouting (there's some huffing and puffing from Dennett about how we should watch out because he's coming for our deepest intuitions with scary tools that may make us flinch, but honestly by now this stuff is about as shocking and countercultural as your dad's Heavy Metal collection); but we've also got unnecessary cover versions of ideas by other people, some stuff that was never really a hit in the first place, and unfortunately one or two bum notes here and there.

And, oh dear, another attempt to smear Descartes by association. First Dennett energetically promoted the phrase "Cartesian theatre" - so hard some people suppose that it actually comes from Descartes; now we have 'Cartesian gravity', more or less a boo-word for any vaguely dualistic tendency Dennett doesn't like. This is surely not good intellectual manners; it wouldn't be quite so bad if it wasn't for the fact that Descartes actually had a theory of gravity, so that the phrase already has a meaning. Should a responsible professor be spreading new-minted misapprehensions like this? Any meme will do?

There's a lot about evolution here that rather left me cold (but then I really, really don't need it explained again, thanks); I don't think Dennett's particular gift is for popularising other people's ideas and his take seems a bit dated. I suspect that most intelligent readers of the book will already know most of this stuff and maybe more, since they will probably have kept up with epigenetics and the various proposals for extension of the modern synthesis that have emerged in the current century, (and the fascinating story of viral intervention in human DNA, surely a natural for anyone who likes the analogy of the selfish gene?) none of which gets any recognition here (I suppose in fairness this is not intended to be full of new stuff). Instead we hear again the tired and in my opinion profoundly unconvincing story about how leaping ('stotting') gazelles are employing a convoluted strategy of wasting valuable energy as a lion-directed display of fitness. It's just an evasive manoeuvre, get over it.

For me it's the most Dennettian bits of the book that are the best, unsurprisingly. The central theme that competence precedes, and may replace, comprehension is actually well developed. Dennett claims that evolution and computation both provide 'inversions' in which intentionless performance can give the appearance of intentional behaviour. He has been accused of

equivocating over the reality of intentionality, consciousness and other concepts, but I like his attitude over this and his defence of the reality of 'free-floating rationales' seems good to me. It gives us permission to discuss the 'purposes' of things without presupposing an intelligent designer whose purposes they are, and I'm completely with Dennett when he argues that this is both necessary and acceptable. I've suggested elsewhere that talking about 'the point' of things, and in a related sense, what they point to, is a handy way of doing this. The problem for Dennett, if there is one, is that it's not enough for competence to replace comprehension often; he needs it to happen every time by some means.

Dennett sets out a theoretical space with 'bottom-up vs top-down', 'random vs directed search', and 'comprehension' as its axes; at one corner of the resulting cube we have intentionless structures like a termite colony; at the other we have fully intentional design like Gaudi's church of the Sagrada Familia, which to Dennett's eye resembles a termite colony. Gaudi's perhaps not the ideal choice here, given his enthusiasm for natural forms; it makes Dennett seem curiously impressed by the underwhelming fact that buildings by an architect who borrowed forms from the natural world turn out to have forms resembling those found in nature.

Still, the space suggests a real contrast between the mindless processes of evolution and deliberate design, which at first sight looks refreshingly different and unDennetian. It's not, of course; Dennett is happy to embrace that difference so long as we recognise that the 'deliberate design' is simply a separate evolutionary process powered by memes rather than genes.

I've never thought that memes, Richard Dawkins's proposed cultural analogue of genes, were a particularly helpful addition to Dennett's theoretical framework, but here he mounts an extended defence of them. One of the worst flaws in the theory as it stands - and there are several - is its confused ontology. What are memes - physical items of culture or abstract ideas? Dennett, as a professional philosopher, seems more sensitive to this problem than some of the more metaphysically naive proponents of the meme. He provides a relatively coherent vision by invoking the idea that memes are 'tokens'; they may take all sorts of physical forms - written words, pictures, patterns of neuronal firing - but each form is a token of a particular way of behaving. The problem here is that anything at all can serve as a token of any meme; we only know that a given noise or symbol tokens a specific meme because of its meaning. There may be - there certainly are - some selective effects that bite on the actual form of particular tokens. A word that is long or difficult to pronounce is more likely to be forgotten. But the really interesting selections take place at the level of meanings; that requires a much more complex level of explanation. There may still be mechanisms involved that are broadly selective if not exactly Darwinian - I think there are - but I believe any move up to this proper level of complexity inevitably edges the simplistic concept of the meme out of play.

The original Dawkinsian observation that the development of cultural items sometimes resembles evolution was sound, but it implicitly called for the development of a general theory which in spite of some respectable attempts, has simply failed to appear. Instead, the supporters of memetics, perhaps trapped by the insistent drumbeat of the Dawkinsian circus, have tended to insist instead that it's all Darwinian natural selection. How a genetic theory can be Darwinian when Darwin never heard of genes is just one of the lesser mysteries here (should we call it 'Mendelian' instead? But Darwin's name is the hooray word here just as Descartes' is the cue for boos). Among the many ways in which cultural selection does not resemble biological evolution, Dennett notes the cogent objection that there is nothing that corresponds to DNA; no general encoding of culture on which selection can operate. One of the worst "bum

notes" in the book is Dennett's strange suggestion that HTML might come to be our cultural DNA. This is, shall we say, an egregious misconception of the scope of text mark-up language.

Anyway, it's consciousness we're interested in (check out Tom Clark's thoughtful take here) and the intentional stance is the number the fans have been waiting for; cunningly kept till last by Dennett. When we get there, though, we get a remix instead of the classic track. Here he has a new metaphor, cunningly calculated to appeal to the youth of today; it's all about apps. Our impression of consciousness is a user illusion created by our gift for language; it's like the icons that activate the stuff on your phone. You may object that a user illusion already requires a user, but hang on. Your ability to talk about yourself is initially useful for other people, telling them useful stuff about your internal states and competences, but once the system is operating, you can read it too. It seems plausible to me that something like that is indeed an important part of the process of consciousness, though in this version I felt I had rather lost track of what was illusory about it.

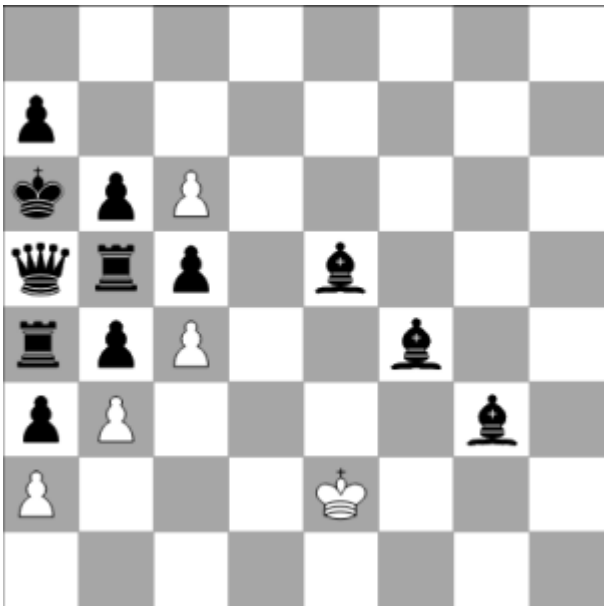
Dennett moves on to a new attack on qualia. This time he offers an explanation of why people think they occur - it's because of the way we project our impressions back out into the world, where they may seem unaccountable. He demonstrates the redundancy of the idea by helpfully sketching out how we could run up a theory of qualia and noting how pointless they are. I was nodding along with this. He suggests that qualia and our own sense of being the intelligent designers in our own heads are the same kind of delusion, simply applied externally or internally. I suppose that's where the illusion is.

He goes on to defend a sort of compatibilist view of free will and responsibility; another example of what Descartes might be tempted to label Dennettian Equivocation, but as before, I like that posture and I'm with him all the way. He continues with a dismissal of mysterianism, leaning rather more than I think is necessary on the interesting concept of joint understanding, where no one person gets it all perfectly, but nothing remains to be explained, and takes a relatively sceptical view of the practical prospects for artificial general intelligence, even given recent advances in machine learning. Does Google Translate display understanding (in some appropriate sense); no, or rather, not yet. This is not Dennett as we remember him; he speaks disparagingly of the cheerleaders for AI and says that "some of us" always discounted the hype. Hmm. Daniel Dennett, long-time AI sceptic?

What's the verdict then? Some good stuff in here, but as always true fans will favour the classic album; if you want Dennett at his best the aficionado will still tell you to buy *Consciousness Explained*.

Chess Problem Computers Can't Solve?

14 March 2017



A somewhat enigmatic report in the Daily Telegraph says that this problem has been devised by Roger Penrose, who says that chess programs can't solve it but humans can get a draw or even a win.

I'm not a chess buff, but to me it seems like a draw. Although Black has an immensely powerful collection of pieces, they are all, completed bottled up and immobile, apart from three bishops. Since these are all, weirdly, on white squares, the White king is completely safe from them if he stays on black squares. Since the white pawns fencing in Black's pieces are all on black squares, the bishops can't do anything about them either. It looks like a drawn position already, in fact: but its quite possible I'm not seeing something.

I suppose Penrose believes that chess computers can't deal with this because it's a very weird situation which will not be in any of their reference material. If they resort to combinatorial analysis the huge number of moves available to the bishops is supposed to render the problem intractable, while the computer cannot see the obvious consequences of the position the way a human can.

I don't know whether it's true that all chess programs are essentially that stupid, but it is meant to buttress Penrose's case that computers lack some quality of insight or understanding that is an essential property of human consciousness.

This is all apparently connected with the launch of a new Penrose Institute, whose website appears to be incomplete. No doubt we'll hear more soon. [I don't think we really did.]

Our Grip On Reality

Are we losing it?

Nick Bostrom's suggestion that we're most likely living in a simulated world continues to provoke discussion. Joelle Dahm draws an interesting parallel with multiverses. I think myself that it depends a bit on what kind of multiverse you're going for - the ones that come from an interpretation of quantum physics usually require conservation of identity between universes - you have to exist in more than one universe - which I think is both potentially problematic and strictly speaking non-Bostromic. Dahm also briefly touches on some tricky difficulties about how we could tell whether we were simulated or not, which seem reminiscent of Descartes' doubts about how he could be sure he wasn't being systematically deceived by a demon - hold that thought for now.

Some of the assumptions mentioned by Dahm would probably annoy Sabine Hossenfelder, who lays into the Bostromisers with a piece about just how difficult simulating the physics of our world would actually be: a splendid combination of indignation with actually knowing what she's talking about.

Bostrom assumes that if advanced civilisations typically have a long lifespan, most will get around to creating simulated versions of their own civilisation, perhaps re-enactments of earlier historical eras. Since each simulated world will contain a vast number of people, the odds are that any randomly selected person is in fact living in a simulated world. The probability becomes overwhelming if we assume that the simulations are good enough for the simulated people to create simulations within their own world, and so on.

There's plenty of scope for argument about whether consciousness can be simulated computationally at all, whether worlds can be simulated in the required detail, and certainly about the optimistic idea of nested simulations. But recently I find myself thinking, isn't it simpler than that? Are we simulated people in a simulated world? No, because we're real, and people in a simulation aren't real.

When I say that, people look at me as if I were stupid, or at least, impossibly naive. Dude, read some philosophy, they seem to say. Dontcha know that Socrates said we are all just grains of sand blowing in the wind?

But I persist - nothing in a simulation actually exists (clue's in the name), so it follows that if we exist, we are not in a simulation. Surely no-one doubts their own existence (remember that parallel with Descartes), or if they do, only on the kind of philosophical level where you can doubt the existence of anything? If you don't even exist, why do I even have to address your simulated arguments?

I do, though. Actually, non-existent people can have rather good arguments; dialogues between imaginary people are a long-established philosophical method (in my feckless youth I may even have indulged in the practice myself).

But I'm not entirely sure what the argument against reality is. People do quite often set out a vision of the world as powered by maths; somewhere down there the fundamental equations are working away and the world is what they're calculating. But surely that is the wrong way

round; the equations describe reality, they don't dictate it. A system of metaphysics that assumes the laws of nature really are explicit laws set out somewhere looks tricky to me; and worse, it can never account for the arbitrary particularity of the actual world. We sort of cling to the hope that this weird specificity can eventually be reduced away by titanic backward extrapolation to a hypothetical time when the cosmos was reduced to the simplicity of a single point, or something like it; but we can't make that story work without arbitrary constants and the result doesn't seem like the right kind of explanation anyway. We might appeal instead to the idea that the arbitrariness of our world arises from it's being an arbitrary selection out of the incalculable banquet of the multiverse, but that doesn't really explain it.

I reckon that reality just is the thing that gets left out of the data and the theory; but we're now so used to the supremacy of those two we find it genuinely hard to remember, and it seems to us that a simulation with enough data is automatically indistinguishable from real events - as though once your 3D printer was programmed, there was really nothing to be achieved by running it.

There's one curious reference in Dahm's piece which makes me wonder whether Christof Koch agrees with me. She says the Integrated Information Theory doesn't allow for computer consciousness. I'd have thought it would; but the remarks from Koch she quotes seem to be about how you need not just the numbers about gravity but actual gravity too, which sounds like my sort of point.

Regular readers may already have noticed that I think this neglect of reality also explains the notorious problem of qualia; they're just the reality of experience. When Mary sees red, she sees something real, which of course was never included in her perfect theoretical understanding.

I may be naive, but you can't say I'm not consistent.

Autism, Camouflage, And Consciousness

23 March 2017

Does recent research into autism suggest real differences between male and female handling of consciousness?

Traditionally, autism has been regarded as an overwhelmingly male condition. Recently, though it has been suggested that the gender gap is not as great as it seems; it's just that most women with autism go undiagnosed. How can that be? It is hypothesised that some sufferers are able to 'camouflage' the symptoms of their autism, and that this suppression of symptoms is particularly prevalent among women.

'Camouflaging' means learning normal social behaviours such as giving others appropriate eye contact, interpreting and using appropriate facial expressions, and so on. But surely, that's just what normal people do? If you can learn these behaviours, doesn't that mean you're not autistic any more?

There's a subtle distinction here between doing what comes naturally and doing what you've learned to do. Camouflaging, on this view, requires significant intellectual resources and continuous effort, so that while camouflaged sufferers may lead apparently normal lives, they are likely to suffer other symptoms arising from the sheer mental effort they have to put in – fatigue, depression, and so on.

Measuring the level of camouflaging - which is obviously intended to be undetectable - obviously raises some methodological challenges. Now a study reported in the invaluable BPS Research Digest claims to have pulled it off. The research team used scanning and other approaches, but their main tool was to contrast two different well-established methods of assessing autism – the Autism Diagnostic Observation Schedule on the one hand and the Autism Spectrum Quotient on the other. While the former assesses 'external' qualities such as behaviour, the latter measures 'internal' ones. Putting it crudely, they measure what you actually do and what you'd like to do respectively. The ratio between the two scores yields a measure of how much camouflaging is going on, and in brief the results confirm that camouflaging is present to a far greater degree in women. I think in fact it's possible the results are understated; all of the subjects were people who had already been diagnosed with autism; that criterion may have selected women who were atypically low in the level of camouflaging, precisely because women who do a lot of camouflaging would be more likely to escape diagnosis.

The research is obviously welcome because it might help improve diagnosis rates for women, but also because a more equal rate of autism for men and women perhaps helps to dispel the idea, formerly popular but (to me at least) rather unpalatable, that autism is really little more than typical male behaviour exaggerated to unacceptable levels.

It does not eliminate the tricky gender issues, though. One thing that surely needs to be taken into account is the possibility that accommodating social pressures is something women do more of anyway. It is plausible (isn't it?) that even among typical average people, women devote more effort to social signals, listening and responding, laughing politely at jokes, and so on. It might be that there is a base level of activity among women devoted to 'camouflaging' normal irritation, impatience, and boredom which is largely absent in men, a baseline against which the findings for people with autism should properly be assessed. It might have been interesting to test a selection of non-autistic people, if that makes sense in terms of the tests. How far the general underlying difference, if it exists, might be due to genetics, socialisation, or other factors is a thorny question.

At any rate, it seems to me inescapable that what the study is really attempting to do with its distinction between outward behaviour and inward states, is to measure the difference between unconscious and conscious control of behaviour. That subtle distinction, mentioned above, between natural and learned behaviour is really the distinction between things you don't have to think about, and things that require constant, conscious attention. Perhaps we might draw a parallel of sorts with other kinds of automatic behaviour. Normally, a lot of things we do, such as walking, require no particular thought. All that stuff, once learned, is taken care of by the cerebellum and the cortex need not be troubled (disclaimer: I am not a neurologist). But people who have their cerebellum completely removed can apparently continue to function: they just have to think about every step all the time, which imposes considerable strain after a while. However, there's no special organ analogous to the cerebellum that records our social routines, and so far as I know it's not clear whether the blend of instinct and learning is similar either.

In one respect the study might be thought to open up a promising avenue for new therapeutic approaches. If women can, to a great extent, learn to compensate consciously for autism, and if that ability is to a great extent a result of social conditioning, then in principle one option would be to help male autism sufferers achieve the same thing through applying similar socialisation. Although camouflaging evidently has its downsides, it might still be a trick worth learning. I doubt if it is as simple as that, though; an awful lot of regimes have been tried out on male sufferers and to date I don't believe the levels of success have been that great; on the other hand it may be that pervasive, ubiquitous social pressure is different in kind from training or special regimes and not so easily deployed therapeutically. The only way might be to bring up autistic boys as girls…

If we take the other view, that women's ability or predisposition to camouflage is not the result of social conditioning, then we might be inclined to look for genuine 'hard-wired' differences in the operation of male and female consciousness. One route to take from there would be to relate the difference to the suggested ability of women (already a cornerstone of gender-related folk psychology) to multi-task more effectively, dividing conscious attention without significant loss to the efficiency of each thread. Certainly one would suppose that having to pay constant attention to detailed social cues would have an impact on the ability to pay attention to other things, but so far as I know there is no evidence that women with camouflaged autism are any worse at paying attention generally than anyone else. Perhaps this is a particular skill of the female mind, while if men pay that much attention to social cues, their ability to listen to what is actually being said is sensibly degraded?

The speculative ice out here is getting thinner than I like, so I'll leave it there; but in all seriousness, any study that takes us forward in this area, as this one seems to do, must be very warmly welcomed.

Fakebots

29 March 2017

Are robots short-changing us imaginatively?

Chat-bots, it seems, might be getting their second (or perhaps their third or fourth) wind. While they're not exactly great conversationalists, the recent wave of digital assistants demonstrates the appeal of a computer you can talk to like a human being. Some now claim that a new generation of bots using deep machine learning techniques might be way better at human conversation than their chat-bot predecessors, whose utterances often veered rapidly from the gnomic to the insane.

A straw in the wind might be the Hugging Face app (I may be showing my age, but for me that name strongly evokes a ghastly Alien parasite). This greatly impressed Rachel Metz, who apparently came to see it as a friend. It's certainly not an assistant - it doesn't do anything except talk to you in a kind of parody of a bubbly teen with a limping attention span. The thing itself is available for IOS and the underlying technology, without the teen angle, appears to be on show here, though I don't really recommend spending any time on either. Actual performance, based on a small sample (I can only take so much) is disappointing; rather than a leap forward it seems distinctly inferior to some Loebner prize winners that never claimed to be doing machine learning. Perhaps it will get better. Jordan Pearson here expresses what seem reasonable reservations about an app aimed at teens that demands a selfie from users as its opening move.

Behind all this, it seems to me, is the looming presence of Spike Jonze's film Her, in which a professional letter writer from the near future (They still have letters? They still write – with pens?) becomes devoted to his digital assistant Samantha. Samantha is just one instance of a bot which people all over are falling in love with. The AIs in the film are puzzlingly referred to as Operating Systems, a randomly inappropriate term that perhaps suggests that Jonze didn't waste any time reading up on the technology. It's not a bad film at all, but it isn't really about AI; nothing much would be lost if Samantha were a fairy, a daemon, or an imaginary friend. There's some suggestion that she learns and grows, but in fact she seems to be a fully developed human mind, if not a superlative one, right from her first words. It's perhaps unfair to single the relatively thoughtful Her out for blame, because with some honourable exceptions the vast majority of robots in fiction are like this; humans in masks.

Fictional robots are, in fact, fakes, and so are all chat-bots. No chat-bot designer ever set out to create independent cognition first and then let it speak; instead they simply echo us back to ourselves as best they can manage. This is a shame because the different patterns of thought that a robot might have; the special mistakes it might be prone to and the unexpected insights it might generate, are potentially very interesting; indeed I should have thought they were fertile ground for imaginative writers. But perhaps 'imaginative' understates the amazing creative powers that would be needed to think yourself out of your own essential cognitive nature. I read a discussion the other day about human nature; it seems to me that the truth is we don't know what human nature is like because we have nothing much to compare it with; it won't be until

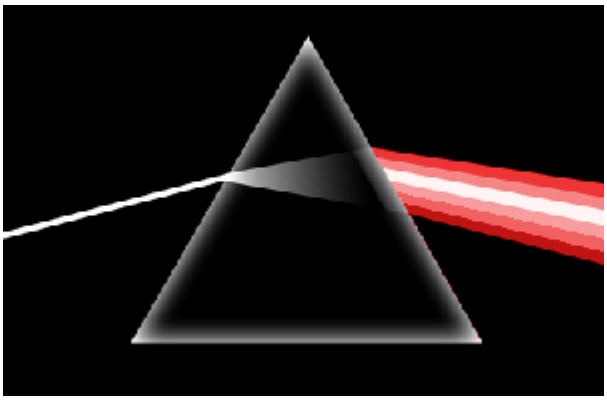
we communicate with aliens or talk properly with non-fake robots that we'll be able to form a proper conception of ourselves.

To a degree it can be argued that there are examples of this happening already. Robots that aspire to Artificial General Intelligence in real world situations suffer badly from the Frame Problem, for instance. That problem comes in several forms, but I think it can be glossed briefly as the job of picking out from the unfiltered world the things that need attention. AI is terrible at this, usually becoming mired in irrelevance (hey, the fact that something hasn't changed might be more important than the fact that something else has). Dennett, rightly I think, described this issue as not the discovery of a new problem for robots so much as a new discovery about human nature; turns out we're weirdly, inexplicably good at something we never even realised was difficult.

How interesting it would be to learn more about ourselves along those challenging, mind-opening lines; but so long as we keep getting robots that are really human beings, mirroring us back to ourselves reassuringly, it isn't going to happen.

What Is It Like To Be A Pigeon?

1 April 2017



Is colour the problem or the solution? Last year we heard about a way of correcting colour blindness with glasses. It only works for certain kinds of colour blindness, but the fact that it works at all is astonishing. Human colour vision relies on three different kinds of receptor cone cells in the retina; each picks up a different wavelength and the brain extrapolates from those data to fill in the spectrum. (Actually, it's far more complex than that, with the background and light conditions taken into account so that the brain delivers a consistent colour reading for the same object even though in different conditions the light reflected from it may be of completely different wavelengths. But let's leave that aside for now and stick with the simplistic view.) The thing is, receptor cells actually respond to a range of wavelengths; in some people two kinds of receptors have ranges that overlap so much the brain can't discriminate. What the glasses do is cut out most of the overlapping wavelengths; suddenly the data from the different receptor cells are very different, and the brain can do a full-colour job at last.

Now a somewhat similar approach has been used to produce glasses that turn normal vision into super colour vision. These new lenses exploit the fact that we have two eyes; by cutting out different parts of the range of wavelengths detected by same kind of receptor in the right and left eyes, they give the effect of four kinds of receptor rather than three. In principle the same approach could double up all three kinds of receptor, giving us the effective equivalent of six kinds of receptor, though this has not been tried yet.

This tetrachromacy or four-colour system is not unprecedented. Some animals, notably pigeons, naturally have four or even more kinds of receptor. And a significant percentage of women, benefiting from the second copy of the relevant genes that you get when you have two 'X' chromosomes, have four kinds of receptor, though it doesn't always lead to enhanced colour vision because in most cases the range of the fourth receptor overlaps the range of another one too largely to be useful.

There is no doubt that all three kinds of tetrachromat - pigeons, women with lucky genes, and people with special glasses - can discriminate between more colours than the rest of us. Because our trichromat eyes have only three sources of data, they have to treat mixtures of wavelengths as though they were the same as pure wavelengths with values equivalent to the average of the mixtures - though they're not. Tetrachromats can do a bit better at this (and I conjecture that colour video and camera images, which use only the three colours needed to fool normal eyes, must sometimes look a bit strange to tetrachromats).

Do tetrachromats see the same spectrum as we do, but in better detail, or do they actually see different colours? There's never been a way to tell for sure. Tetrachromats can't tell us what colours they see any more than we can tell each other whether my red is the same as yours, or instead is the same as what you experience for green.The curious fact that the ends of the spectrum join up into a complete colour wheel might support the idea that the spectrum is in

some sense an objective reality, based on mathematical harmonic relationships analogous to those of sound waves; in effect we see a single octave of colour with the wavelength at one end double (or half) that at the other. I've sort of speculated in the past that if our eyes could see a much wider range of wavelengths we would see lower and higher octaves of colour; not wholly new colours like Terry Pratchett's octarine, but higher and lower reds, greens and blues. I speculated further that 'lower' and 'higher' might actually be experienced as 'cooler' and 'hotter'. That is of course the wildest guesswork, but the thesis that everyone - tetrachromats included - sees the same spectrum but in lesser or greater detail seems to be confirmed by the experimenters if I'm reading it right.

Of course, colour vision is not just a matter of what happens in the retina; there is also a neural colour space mapped out in the brain (which interestingly is a little more extensive than the colour space of the real world, leading to the hidden existence of 'chimerical' colours). Do pigeons, human tetrachromats, and human trichromats all map colours to similar neural spaces? I haven't been able to find out, but I'm guessing the answer is yes. If it weren't so, there would be potential issues over neural plasticity. If your brain receives no signals from one eye during your early life, it re-purposes the relevant bits of neural real estate and you cannot get your vision back later even if the eye starts sending the right kind of signal. We might expect that people who were colour blind from birth would be affected in a similar way, yet in fact use of the new glasses seems to bring an intact colour system straight into operation for the first time. So it might be that a standard spectral colour space is hard-wired into the genes of all of us (even pigeons), or again it might be that the spectrum is a mathematical reality which any visual system must represent, albeit with varying fidelity.

All of this is skating around the classic philosophical issues. Does Mary, who never saw colours, know something new when she has seen red? Well, we can say with confidence that the redness will be registered and mapped properly; she will not have lost the ability to see colour through being brought up in a monochrome world. More importantly, the scientifically tractable aspects of colour vision have moved another step closer to the subjective experience. We have some objective reasons for supposing that Mary's colour experience will be arranged along the same spectral structure as ours, though not necessarily graduated with the same fineness.

None of this will banish the Hard Problem, or dispel our particular sense that colours especially are subjective optional extras. For a long time some have thought of colour as a 'secondary' property, in the observer, not the world; not like such properties as mass or volume, which are more 'real'. The newly-understood complexity of colour vision leads to new arguments that it is in fact artificial, a useful artefact in the brain, in some sense not really there in objective reality. My feeling though is that if we can all experience tetrachromacy, the gap between the objective and the subjective will not be perceived as being so unbridgeable as it has been to date.

The Hard Problem Of Physics

9 April 2017

Is there a Hard Problem of physics that explains the Hard Problem of consciousness?

Hedda Hassel Mørch has a thoughtful piece in Nautilus's interesting Consciousness issue (well worth a look generally) that raises this idea. What is the alleged Hard Problem of physics? She say it goes like this...

> *What is physical matter in and of itself, behind the mathematical structure described by physics?*

To cut to the chase, Mørch proposes that things in themselves have a nature not touched by physics, and that nature is consciousness. This explains the original Hard Problem - we, like other things, just are by nature conscious; but because that consciousness is our inward essence rather than one of our physical properties, it is missed out in the scientific account.

I'm sympathetic to the idea that the original Hard Problem is about an aspect of the world that physics misses out, but according to me that aspect is just the reality of things. There may not, according to me, be much more that can usefully be said about it. Mørch, I think, takes two wrong turns. The first is to think that there are such things as things in themselves, apart from observable properties. The second is to think that if this were so, it would justify panpsychism, which is where she ends up.

Let's start by looking at that Hard problem of physics. Mørch suggests that physics is about the mathematical structure of reality, which is true enough, but the point here is that physics is also about observable properties; it's nothing if not empirical. If things have a nature in themselves that cannot be detected directly or indirectly from observable properties, physics simply isn't interested, because those things-in-themselves make no difference to any possible observation. No doubt some physicists would be inclined to denounce such unobservable items as absurd or vacuous, but properly speaking they are just outside the scope of physics, neither to be affirmed nor denied. It follows, I think, that this can't be a Hard Problem of physics; it's actually a Hard Problem of metaphysics.

This is awkward because we know that human consciousness does have physical manifestations that are readily amenable to physical investigation; all of our conscious behaviour, our speech and writing, for example. Our new Hard Problem (let's call it the NHP) can't help us with those; it is completely irrelevant to our physical behaviour and cannot give us any account of those manifestations of consciousness. That is puzzling and deeply problematic

- but only in the same way as the old Hard Problem (OHP) - so perhaps we are on the right track after all?

The problem is that I don't think the NHP helps us even on a metaphysical level. Since we can't investigate the essential nature of things empirically, we can only know about it by pure reasoning; and I don't know of any purely rational laws of metaphysics that tell us about it. Can the inward nature of things change? If so, what are the (pseudo-causal?) laws of intrinsic change that govern that process? If the inward nature doesn't change, must we take everything to be essentially constant and eternal in itself? That Parmenidean changelessness would be particularly odd in entities we are relying on to explain the fleeting, evanescent business of subjective experience.

Of course Mørch and others who make a similar case don't claim to present a set of a priori conclusions about their own nature; rather they suggest that the way we know about the essence of things is through direct experience. The inner nature of things is unknowable except in that one case where the thing whose inner nature is to be known is us. We know our own nature, at least. It's intuitively appealing - but how do we know our own real nature? Why should being a thing bring knowledge of that thing? Just because we have an essential nature here's no reason to suppose we are acquainted with that inner nature; again we seem to need some hefty metaphysics to explain this, which is actually lacking. All the other examples of knowledge I can think of are constructed, won through experience, not inherent. If we have to invent a new kind of knowledge to support the theory the foundations may be weak.

At the end of the day, the simplest and most parsimonious view, I think, is to say that things just are made up of their properties, with no essential nub besides. Leibniz's Law tells us that that's the nature of identity. To be sure, the list will include abstract properties as well as purely physical ones, but abstract properties that are amenable to empirical test, not ones that stand apart from any possible observation. Mørch disagrees:

> *Some have argued that there is nothing more to particles than their relations, but intuition rebels at this claim. For there to be a relation, there must be two things being related. Otherwise, the relation is empty—a show that goes on without performers, or a castle constructed out of thin air.*

I think the argument is rather that the properties of a particle relate to each other, while these groups of related properties relate in turn to other such groups. Groups don't require a definitive member, and particles don't require a single definitive essence. Indeed, since the particle's essential self cannot determine any of its properties (or it could be brought within the pale of physics) it's hard to see how it can have a defined relation to any of them and what role the particle-in-itself can play in Mørch's relational show.

The second point where I think Mørch goes wrong is in the leap to panpsychism. The argument seems to be that the NHP requires non-structural stuff (which she likens to the hardware on which the software of the laws of physics runs - though I myself wouldn't buy unstructured hardware); the OHP gives us the non-structural essence of conscious experience (of course conscious experience does have structure, but Mørch takes it that down there somewhere is the structureless ineffable something-it-is-like); why not assume that the latter is universal and fills the gap exposed by the NHP?

Well, because other matter exhibits no signs of consciousness, and because the fact that our essence is a conscious essence just wouldn't warrant the assumption that all essences are

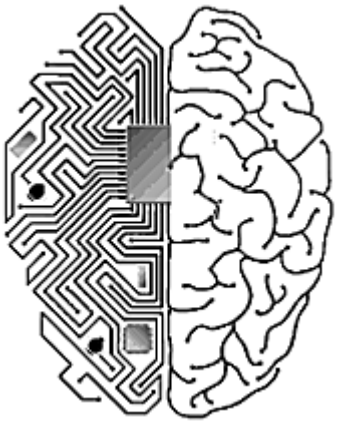
conscious ones. Wouldn't it be simpler to think that only the essences of outwardly conscious beings are conscious essences? This is quite apart from the many problems of panpsychism, which we've discussed before, and which Mørch fairly acknowledges.

So I'm not convinced, but the case is a bold and stimulating one and more persuasively argued than it may seem from my account. I applaud the aims and spirit of the expedition even though I may regret the direction it took.

Consciousness in the Singularity

15 April 2017

What is it like to be a Singularity (or in a Singularity)?



You probably know the idea. At some point in the future, computers become generally cleverer than us. They become able to improve themselves faster than we can do, and an accelerating loop is formed where each improvement speeds up the process of improving, so that they quickly zoom up to incalculable intelligence and speed, in a kind of explosion of intellectual growth. That's the Singularity. Some people think that we mere humans will at some point have the opportunity of digitising and uploading ourselves, so that we too can grow vastly cleverer and join in the digital world in which these superhuman conscious entities will exist.

Just to clear upfront, I think there are some basic flaws in the plausibility of the story which mean the Singularity is never really going to happen: could never happen, in fact. However, it's interesting to consider what the experience would be like.

How would we digitise ourselves? One way would be to create a digital model of our actual brain, and run that. We could go the whole hog and put ourselves into a fully simulated world, where we could enjoy sweet dreams forever, but that way we should miss out on the intellectual growth which the Singularity seems to offer, and we should also remain at the mercy of the vast new digital intellects who would be running the show. Generally I think it's believed that only by joining in the cognitive ascent of these mighty new minds can we assure our own future survival.

In that case, is a brain simulation enough? It would run much faster than a meat brain, a point we'll come back to, but it would surely suffer some of the limitations that biological brains are heir to. We could perhaps gradually enhance our memory and other faculties and gradually improve things that way, a process which might provide a comforting degree of continuity, but it seems likely that entities based on a biological scheme like this would be second-class citizens within the digital world, falling behind the artificial intellects who endlessly redesign and improve themselves. Could we then preserve our identity while turning fully digital and adopting a radical new architecture?

The subject of what constitutes personal identity, be it memory, certain kinds of continuity, or something else, is too large to explore here, except to note a basic question; can our identity ultimately be boiled down to a set of data? If the answer is yes (I actually believe it's 'no', but today we'll allow anything) , then one way or another the way is clear for uploading ourselves into an entirely new digital architecture.

The way is also clear for duplicating and splitting ourselves. Using different copies of our data we can become several people and follow different paths. Can we then re-merge? If the data that constitutes us is static, it seems we should be able to recombine it with few issues; if it is partly a description of a dynamic process, we might not be able to do the merger on the fly, and might have to form a third, merged individual. Would we terminate the two contributing selves? Would we worry less about 'death' in such cases? If you know your data can always be brought

back into action, terminating the processes using that data (for now) might seem less frightening than the irretrievable destruction of your only brain.

This opens up further strange possibilities. At the moment our conscious experience is essentially linear (it's a bit more complex than that, with layers and threads of attention, but broadly there's a consistent chronological stream). In the brave new world our consciousness could branch out without limit; or we could have grid experiences, where different loci of consciousness follow crossing paths, merging at each node and the splitting again, before finally reuniting in one node with (very strange) remembered composite experience.

If merging is a possibility, then we should be able to exchange bits of ourselves with other denizens of the digital world, too. When handed a copy of part of someone else we might retain it as exterior data, something we just know about, or incorporate it into a new merged self, whether a successor to ourselves as ourselves, or as a kind of child; if all our data is saved the difference perhaps ceases to be of great significance. Could we exchange data like this with the artificial entities that were never human, or would they be too different?

I'm presupposing here that both the ex-humans and the artificial consciousnesses here remain multiple and distinct. Perhaps there's an argument for generally merging into one huge consciousness? I think probably not, because it seems to me that multiple loci of consciousness would just get more done in the way of thinking and experiencing. Perhaps when we became sufficiently linked and multi-threaded, with polydimensional multi-member grid consciousnesses binding everything loosely together anyway the question of whether we are one or many – and how many – might not seem important any more.

If we can exchange experiences, does that solve the Hard Problem? We no longer need to worry whether your experience of red is the same as mine, we just swap. Now many people (and I am one) would think that fully digitised entities wouldn't be having real experiences anyway, so any data exchange they might indulge in would be irrelevant. There are several ways it could be done, of course. It might be a very abstract business or entities of human descent might exchange actual neural data from their old selves. If we use data which, fed into a meat brain, definitely produces proper experience, it perhaps gets a little harder to argue that the exchange is phenomenally empty.

The strange thing is, even if we put all the doubts aside and assume that data exchanges really do transfer subjective experience, the question doesn't go away. It might be that attachment to a particular node of consciousness conditions the experience so that it is different anyway.

Consider the example of experiences transferred within a single individual, but over time. Let's think of acquired tastes. When you first tasted beer, it seemed unpleasant; now you like it. Does it taste the same, with you having learnt to like that same taste? Or did it in fact taste different to you back then – more bitter, more sour? I'm not sure it's possible to answer with great confidence. In the same way, if one node within the realm of the Singularity 'runs' another's experience, it may react differently, and we can't say for sure whether the phenomenal experience generated is the same or not…

I'm assuming a sort of cyberspace where these digital entities live – but what do they do all day? At one end of the spectrum, they might play video games constantly – rather sadly reproducing the world they left behind. Or at the intellectually pure end, they might devote themselves to the study of maths and philosophy. Perhaps there will be new pursuits that we, in our stupid meaty way, cannot even imagine as yet. But it's hard not to see a certain tedious emptiness in the pure

life of the mind as it would be available to these intellectual giants. They might be tempted to go on playing a role in the real world.

The real world, though, is far too slow. Whatever else they have improved, they will surely have racked up the speed of computation to the point where thousands of years of subjective time take only a few minutes of real world time. The ordinary physical world will seem to have slowed down very close to the point of stopping altogether; the time required to achieve anything much in the real world is going to seem like millions of years.

In fact, that acceleration means that from the point of view of ordinary time, the culture within the Singularity will quickly reach a limit at which everything it could ever have hoped to achieve is done. Whatever projects or research the Singularitarians become interested in will be completed and wrapped up in the blinking of an eye. Unless you think the future course of civilisation is somehow infinite, it will be completed in no time. This might explain the Fermi Paradox, the apparently puzzling absence of advanced alien civilisations: once they invent computing, galactic cultures go into the Singularity, wrap, themselves up in a total intellectual consummation, and within a few days at most, fall silent forever.

Self-discovery: fascinating journey of life or load of tosh? An IAI discussion.

On the whole, I think the vastness of the subject means we get no more than first steps here, though the directions are at least interesting. Joanna Kavenna notes the paradoxical entanglements that can arise from self-examination and makes an interesting comparison with the process of novelists finding their ‘voice’. Exploration of selves is of course the bedrock of the novel, a topic which could take up many pages in itself. She asserts that the self is experientially real, but that thought also floats away unexamined.

David Chalmers has a less misty proposition; people have traits and we are inclined to think of some as deep or essential. Identifying these is a reasonable project, but not without dangers if we settle on the wrong ones.

Ed Stafford seems to be uncomfortable with philosophy unless it comes from an ayahuasca session or a distant tribe. He likes the idea of thinking with your stomach, but does not shed any light on the interesting question of how stomach thoughts differ from brain ones. In general he seems to take the view that for well-adjusted people there is no mystery, one knows who one is and there’s no need to wibble about it. Oddly, though he mentions being dropped on a desert island where the solitude was so severe, that even when the helicopter was still in view, he vomited. To suffer radical depersonalisation after a couple of minutes alone on a beach seems an extraordinary example of personal fragility, but I suppose we are to understand this was before he centred himself through contact with more robust cultures. Of course, those who reject theory always in fact have a theory; it’s just one that they either haven’t examined or don’t want examined. In response to Chalmers’ suggestion that a loving environment can surely lead to personal growth, he seems to begin adding qualifications to his view of the robustly settled personality, but if we are witnessing actual self-discovery here it doesn’t go far.

Myself I reckon that you don’t need to identify your essential traits to experience self-discovery; merely becoming conscious of your own traits renders them self-conscious and transforms them, an iterative process that represents a worthwhile kind of growth, both moral and psychological. But I’ve never tried ayahuasca.', 'The unexamined self is not worth being', '', 'inherit', 'closed', 'closed', '', '2859-revision-v1', '', '',

Under-hypnotised

23 April 2017

Maybe hypnosis is the right state of mind and 'normal' is really 'under-hypnotised'?

That's one idea that does not appear in the comprehensive synthesis of what we know about hypnosis produced by Terhune, Cleeremans, Raz and Lynn. It is a dense, concentrated document, thick with findings and sources, but they have done a remarkably good job of keeping it as readable as possible, and it's both a useful overview and full of interesting detail.

Hypnosis, it seems, has two components; the induction and one or more suggestions. The induction is what we normally think of as the process of hypnotising someone. It's the bit that in popular culture is achieved by a swinging watch, mystic hand gestures or other theatrical stuff; in common practice probably just a verbal routine. It seems this is much less important than we might have been led to think, and Terhune et al don't find it of primary interest. The truth is that hypnosis is more about the suggestibility of the subject than about the effectiveness of the induction. In fact if you want to streamline your view, you could see the induction as simply the first suggestion. Post-hypnotic suggestions, which take effect after the formal hypnosis session has concluded, may use different mechanisms to suggestions that take immediate effect, though it seems this has yet to be fully explored.

Broadly, people fall into three groups. 10 to 15 per cent of people are very suggestible, responding strongly to the full range of suggestions; about the same proportion are poorly suggestible and respond to hypnosis poorly or not at all; the rest of us are somewhere in the middle. Suggestibility is a fixed characteristic and seems to be heritable; but it does not correlate strongly with other personality traits (nor with dissociative conditions such as Dissociative Identity Disorder, formerly known as Multiple Personality Disorder). It does interestingly resemble the kind of suggestibility seen in the placebo effect.

Terhune et al regard the debate about whether hypnosis is an altered state of consciousness as an unproductive one; but there are certainly some points of great interest here when it comes to consciousness. A key feature of hypnosis is the loss of the sense of agency; hypnotised subjects think of their arm moving, not of having moved their arm. Credible current theories attribute this to the suppression of second-order mental states, or of metacognition. Typically in the literature it seems this is discussed as a derangement of the proper sense of agency, but of course elsewhere people have concluded that our sense of agency is a delusion anyway. So perhaps, to repeat my opening suggestion, it's the hypnotised subjects who have it right, and if we want to understand our own minds properly we should all enter a hypnotic state. Or perhaps that's too much like noticing that blind people don't suffer from optical illusions.

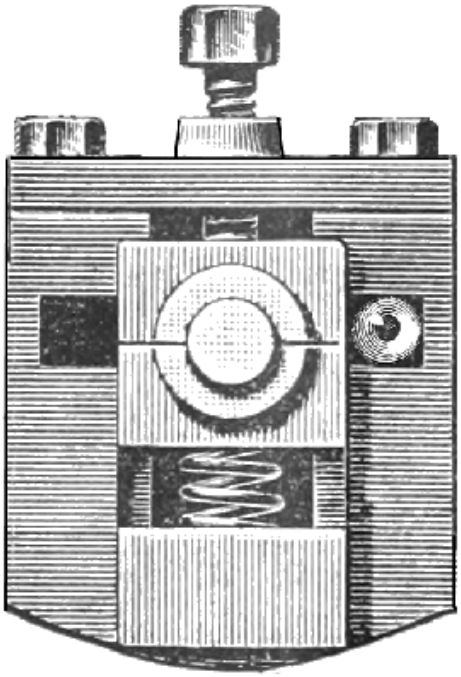
There's a useful distinction here between voluntary control and top-down control. One interesting thing about hypnosis is that it demonstrates the power of top-down control, where beliefs and other high-level states determine basic physiological responses, something we may be inclined to under-rate. But hypnosis also highlights that top-down control does not imply

agency; perhaps we sometimes mistake the former for the latter? At any rate it seems to me that some of this research ought to be highly relevant to the analysis of agency and suggests some potentially interesting avenues.

Another area of interest is surely the ability of hypnosis to affect attention and perception. It had been shown that changes in colour perception induced by hypnosis are registered in the brain differently from mere imagined changes. If we tell someone under hypnosis to see red for green and green for red, does that change the qualia of the experience or not? Do they really see green instead of red, or merely believe that's what is happening? If anything the facts of hypnosis seem to compound the philosophical problems rather than helping to solve them; nevertheless it does seem to me that quite a lot of the results so handily summarised here should have a bigger impact on current philosophical discussion than they have had to date.

Superfluous Consciousness

1 May 2017



Do we need robots to be conscious? Ryota Kanai thinks it is largely up to us whether the machines wake up - but he is working on it. I think his analysis is pretty good and in fact I think we can push it a bit further.

His opening remarks, perhaps due to over-editing, don't clearly draw the necessary distinction between Hard and Easy problems, or between subjective p-consciousness and action-related a-consciousness (I take it to be the same distinction, though not everyone would agree). Kanai talks about the unsolved mystery of experience, which he says is not a necessary by-product of cognition, and says that nevertheless consciousness must be a product of evolution. Hm. It's p-consciousness, the ineffable, phenomenal business of what experience is like, that is profoundly mysterious, not a necessary by-product of cognition, and quite possibly nonsense. That kind of consciousness cannot in any useful sense be a product of evolution, because it does not affect my behaviour, as the classic zombie twin thought experiment explicitly demonstrates. A-consciousness, on the other hand, the kind involved in reflecting and deciding, absolutely does have survival value and certainly is a product of evolution, for exactly the reasons Kanai goes on to discuss.

The survival value of A-consciousness comes from the way it allows us to step back from the immediate environment; instead of responding to stimuli that are present now, we can respond to ones that were around last week, or even ones that haven't happened yet; our behaviour can address complex future contingencies in a way that is both remarkable and powerfully useful. We can make plans, and we can work out what to do in novel situations (not always perfectly, of course, but we can do much better than just running a sequence of instinctive behaviour).

Kanai discusses what must be about the most minimal example of this; our ability to wait three seconds before responding to a stimulus. Whether this should properly be regarded as requiring full consciousness is debatable, but I think he is quite right to situate it within a continuum of detached behaviour which, further along, includes reactions to very complex counterfactuals.

The kind he focuses on particularly is self-consciousness or higher-order consciousness; thinking about ourselves. We have an emergent problem, he points out, with robots whose reasons are hidden; increasingly we cannot tell why a complex piece of machine learning resulted in the behaviour that resulted. Why not get the robot to tell us, he says; why not enable it to report its own inner states? And if it becomes able to consider and explain its own internal states, won't that be a useful facility which is also like the kind of self-reflecting consciousness that some philosophers take to be the crucial feature of the human variety?

There's an immediate and a more general objection we might raise here. The really bad problem with machine learning is not that we don't have access to the internal workings of the robot

mind; it's really that in some cases there just is no explanation of the robot's behaviour that a human being can understand. Getting the robot to report will be no better than trying to examine the state of the robot's mind directly; in fact it's worse, because it introduces a new step into the process, one where additional errors can creep in. Kanai describes a community of AIs, endowed with a special language that allows them to report their internal states to each other. It sounds awfully tedious, like a room full of people who, when asked 'How are you?' each respond with a detailed health report. Maybe that is quite human in a way after all.

The more general theoretical objection (also rather vaguer, to be honest) is that, in my opinion at least, Kanai and those Higher Order Theory philosophers just overstate the importance of being able to think about your own mental states. It is an interesting and important variety of consciousness, but I think it just comes for free with a sufficiently advanced cognitive apparatus. Once we can think about anything, then we can of course think about our thoughts.

So do we need robots to be conscious? I think conscious thought does two jobs for us that need to be considered separately although they are in fact strongly linked. I think myself that consciousness is basically recognition. When we pull off that trick of waiting for three seconds before we respond to a stimulus, it is because we recognise the wait as a thing whose beginning is present now, and can therefore be treated as another present stimulus. This one simple trick allows us to respond to future things and plan future behaviour in a way that would otherwise seem to contradict the basic principle that the cause must come before effect.

The first job that does is allow the planning of effective and complex actions to achieve a given goal. We might want a robot to be able to do that so it can acquire the same kind of effectiveness in planning and dealing with new situations which we have ourselves, a facility which to date has tended to elude robots because of the Frame Problem and other issues to do with the limitations of pre-programmed routines.

The second job is more controversial. Because action motivated by future contingencies has a more complex causal back-story, it looks a bit spooky, and it is the thing that confers on us the reality (or the illusion, if you prefer) of free will and moral responsibility. Because our behaviour comes from consideration of the future, it seems to have no roots in the past, and to originate in our minds. It is what enables us to choose 'new' goals for ourselves that are not merely the consequence of goals we already had. Now there is an argument that we don't want robots to have that. We've got enough people around already to originate basic goals and take moral responsibility; they are a dreadful pain already with all the moral and legal issues they raise, so adding a whole new category of potentially immortal electronic busybodies is arguably something best avoided. That probably means we can't get robots to do job number one for us either; but that's not so bad because the strategies and plans which job one yields can always be turned into procedures after the fact and fed to 'simple' computers to run. We can, in fact, go on doing things the way we do them now; humans work out how to deal with a task and then give the robots a set of instructions; but we retain personhood, free will, agency and moral responsibility for ourselves.

There is quite a big potential downside, though; it might be that the robots, once conscious, would be able to come up with better aims and more effective strategies than we will ever be able to devise. By not giving them consciousness we might be permanently depriving ourselves of the best possible algorithms (and possibly some superior people, but that's a depressing thought from a human point of view). True, but then I think that's almost what we are on the brink of doing already. Kanai mentions European initiatives which may insist that computer

processes come with an explanation that humans can understand; if put into practice the effect, once the rule collides with some of those processes that simply aren't capable of explanation, would be to make certain optimal but inscrutable algorithms permanently illegal.

We could have the best of both worlds if we could devise a form of consciousness that did job number one for us without doing job two as an unavoidable by-product, but since in my view they're all acts of recognition of varying degrees of complexity, I don't see at the moment how the two can be separated.

Emotional Chatting Machine

6 May 2017

Is this a breakthrough in robot emotion? Zhou et al describe the Emotional Chatting Machine (ECM), a chatbot which uses machine learning to return answers with a specified emotional tone.

The ultimate goal of such bots is to produce a machine that can detect the emotional tone of an input and choose the optimum tone for its response, but this is very challenging. It's not simply a matter of echoing the emotional tone of the input; the authors suggest for example, that sympathy is not always the appropriate response to a sad story. For now, the task they address is to take two inputs; the actual content and a prescribed emotional tone, and generate a response to the content reflecting the required tone. Actually, doing more than reflecting is going to be very challenging indeed because the correct tone of a response ought to reflect the content as well as the tone of the input; if someone calmly tells you they're about to die, or about to kill someone else, an equally calm response may not be emotionally appropriate (or it could be in certain contexts; this stuff is, to put it mildly, complex).

To train the ECM, two databases were used. The NLPCC dataset has 23,105 sentences collected from Weibo, a Chinese blog site, and categorised by human beings using eight categories: Anger, Disgust, Fear, Happiness, Like, Sadness, Surprise and Other. Fear and Surprise turned up too rarely on Weibo blogs to be usable in practice.

Rather than using the NLPCC dataset directly, the researchers used it to train a classifier which then categorised the larger STC dataset, which has 219,905 posts and 4,308,211 responses; they reckon they achieved an accuracy of 0.623, which doesn't sound all that great, but was apparently good enough to work with; obviously this is something that could be improved in future. It was the 'emotionalised' STC data set which was then used to train the ECM for its task.

User: Worst day ever. I arrived late because of the traffic.
Chatbot (Basic Seq2Seq): You were late. ECM (*Like*): I am always here to support you. ECM (*Happiness*): Keep smiling! Things will get better. ECM (*Sadness*): It's depressing. ECM (*Disgust*): Sometimes life just sucks. ECM (*Anger*): The traffic is too bad!

Results were assessed by human beings for both naturalness (how human they seemed) and emotional accuracy; ECM improved substantially on other approaches and generally turned in a good performance, especially on emotional accuracy. Alas, the chattbot is not available to try out online.

This is encouraging but I have a number of reservations. The first is about the very idea of an emotional chatbot. Chatbots are almost by definition superficial. They don't attempt to reproduce or even model the processes of thought that underpin real conversation, and similarly they don't attempt to endow machines with real or even imitation emotion (the ECM has internal and external memory in which to record emotional states, but that's as far as it goes). Their performance is always, therefore, on the level of a clever trick.

Now that may not matter, since the aim is merely to provide machines that deal better with emotional human beings. They might be able to do that without having anything like real or even model emotions themselves (we can debate the ethical implications of 'deceiving' human interlocutors like this another time). But there must be a worry that performance will be unreliable.

Of course, we've seen that by using large data sets, machines can achieve passable translations without ever addressing meanings; it is likely enough that they can achieve decent emotional results in the same sort of way without ever simulating emotions in themselves. In fact the complexity of emotional responses may make humans more forgiving than they are for translations, since an emotional response which is slightly off can always be attributed to the bot's personality, mood, or other background factors. On the other hand, a really bad emotional misreading can be catastrophic, and the chatbot approach can never eliminate such misreading altogether.

My second reservation is about the categorisation adopted. The eight categories adopted for the NLPCC data set, and inherited here with some omissions, seem to belong to a family of categorisations which derive ultimately from the six-part one devised by Paul Ekman: anger, disgust, fear, happiness, sadness, and surprise. The problem with this categorisation is that it doesn't look plausibly comprehensive or systematic. Happiness and sadness look like a pair, but there's no comparable positive counterpart of disgust or fear, for example. These problems have meant that the categories are often fiddled with. I conjecture that 'like' was added to the NLPCC set as a counterpart to disgust, and 'other' to ensure that everything could be categorised somewhere. You may remember that n the Pixar film Inside Out Surprise didn't make the cut; some researchers have suggested that only four categories are really solid, with fear/surprise and anger/disgust forming pairs that are not clearly distinct.

The thing is, all these categorisations are rooted in attempts to categorise facial expressions. It isn't the case that we necessarily have a distinct facial expression for every possible emotions, so that gives us an incomplete and slightly arbitrary list. It might work for a bot that pulled faces, but one that provides written outputs needs something better. I think a dimensional approach is better; one that defines emotions in terms of a few basic qualities set out along different axes. These might be things like attracted/repelled, active/passive, ingoing/outgoing or whatever. There are many models along these lines and they have a long history in psychology; they offer better assurance of a comprehensive account and a more hopeful prospect of a reductive explanation.

I suppose you also have to ask whether we want bots that respond emotionally. The introduction of cash machines reduced the banks' staff costs, but I believe they were also popular because you could get your money without having to smile and talk. I suspect that in a similar way we really just want bots to deliver the goods (often literally), and their lack of messy humanity is their strongest selling point. I suspect though, that in this respect we ain't seen nothing yet...

Hippocampal Holodeck

14 May 2017

Matt Faw says subjective experience is caused by a simulation in the hippocampus - a bit like a holodeck. There's a brief description on the Brains Blog, with the full version here.

In a very brief sketch, Faw says that data from various systems is pulled together in the hippocampal system and compiled into a sort of unified report of what's going on. This is sort of a global workspace system, whose function is to co-ordinate. The ongoing reportage here is like a rolling film or holodeck simulation, and because it's the only unified picture available, it is mistaken for the brain's actual interaction with the world. The actual work of cognition is done elsewhere, but this simulation is what gives rise to 'neurotypical subjective experience'.

I'm uneasy about that; it doesn't much resemble what I would call subjective experience. We can have a model of the self in the world without any experiencing going on (Roger Penrose suggested the example of a video camera pointed at a mirror), while the actual subjectivity of phenomenal experience seems to be nothing to do with the 'structural' properties of whether there's a simulation of the world going on.

I believe the simulation or model is supposed to help us think about the world and make plans; but the actual process of thinking about things rarely involves a continuous simulation of reality. If I'm thinking of going out to buy a newspaper, I don't have to run through imagining what the experience is going to be like; indeed, to do so would require some effort. Even if I do that, I'm the puppeteer throughout; it's not like running a computer game and being surprised by what happens. I don't learn much from the process.

And what would the point be? I can just think about the world. Laboriously constructing a model of the world and then thinking about that instead looks like a lot of redundant work and a terrible source of error if I get it wrong.

There's a further problem for Faw in that there are people who manage without functioning hippocampi. Although they undoubtedly have serious memory problems, they can talk to us fairly normally and answer questions about their experiences. It seems weird to suggest that they don't have any subjective experience; are they philosophical zombies?

Faw doesn't want to say so. Instead he likens their thought processes to the ones that go on when we're driving without thinking. Often we find we've driven somewhere but cannot remember any details of the journey. Faw suggests that what happens here is just that we don't remember the driving. All the functions that really do the cognitive work are operating normally, but whereas in other circumstances their activity would get covered in the 'news bulletin'

simulation, in this case our mind is dealing with something more interesting, and just fails to record what we've done with the brake and the steering wheel. But if we were asked at the very moment of turning left what we were doing, we'd have no problem answering. People with no hippocampi are like this; constantly aware of the detail of current cognition, stuff that is normally hidden from neurotypically normal people.

I broadly like this account, and it points to the probability that the apparent problems for Faw are to a degree just a matter of labelling. He's calling it subjective experience, but if he called it reportable experience it would make a lot of sense. We only ever report what we remember to have been our conscious experience; some time, even if only an instant, has to have passed. It's entirely plausible that that we rely on the hippocampus to put together these immediate, reportable memories for us.

So really what I call subjective experience is going on all the time out there; it doesn't require a unified account or model; but it does need the hippocampal newsletter in order to become reportable. Faw and I might disagree on the fine philosophical issue of whether it is meaningful to talk about experiences that cannot, in principle, be reported; but in other ways we don't really differ as much as it seemed.

That Bloke

22 May 2017

Is Karl Friston that bloke? You know who I mean. That really clever bloke, master of some academic field. He's used to understanding a few important things far better than the rest of us. But when people in the pub start discussing philosophy of mind, he feels wrong-footed in this ill-defined territory.

"You see," he says, "the philosophers make such a meal of this, with all their vague, mystical talk of intentions and experiences. But I'm a simple man, and I can't help looking at it as a scientist. To me it just seems obvious that it's basically nothing more than a straightforward matter of polyvalent symmetry relations between vectors in high-order Minkowski topologies, isn't it?

Aha! Now you have to meet him on his own turf and defeat him there: otherwise you can't prove he hasn't solved the problem of consciousness!

I'm sure that isn't really Friston at all; but his piece in Aeon, perhaps due to its unavoidable brevity, seems to invoke some relatively esoteric apparatus while ultimately saying things that turn out to be either madly optimistic or relatively banal. Let's quickly canter through it. Friston starts by complaining of philosophers and cognitive scientists who insist that the mind is a thing. "As a physicist and psychiatrist," he says "I find it difficult to engage with conversations about consciousness." In physics, he says, it's dangerous to assume that that things 'exist'; the real question is what processes give rise to the notion that something exists. (An unexpected view: physics, then, seeks to explain the notions in our minds, not external reality? Poor old Isaac Newton wasn't really doing proper physics at all.) Friston, instead, wants to brush all that 'thing' nonsense aside and argue that consciousness, in reality, is a natural process.

I've spent some time trying to think of anyone who would deny that, and really I've come up empty. Panpsychists probably think that at the most fundamental level consciousness can be a very simple form of awareness, too simple to go through complex changes: but even they, I think, would not deny that human consciousness is a natural process. Perhaps all Friston means is that he doesn't want to spend any time on definitions of consciousness; that territory is certainly a treacherous swamp where people get lost, although setting out to understand something you haven't defined can be kinda bold, too.

To illustrate the idea of consciousness as a process, Friston (inexplicably, to me anyway: this is one of the places where I feel something might have got lost in editing) suggests we swap the word and talk about whether evolution is a process. Scientifically, he says, we know that evolution isn't for anything - it's just a process that happens. Since consciousness is a product of evolution, it isn't for anything either. I don't know about that; it's true it can't borrow its purpose from evolution if evolution doesn't have one; the thing is, putting aside all the difficult

issues of ultimate purpose and the nature of teleology, there is a well-established approach within evolutionary theory of asking what, say, horns or eyes are for (defeating rivals, seeing food, etc). This is just a handy way to ask about the survival value of particular adaptations. So within evolution things like consciousness can be for something in a relatively clear sense without evolution itself having to be for anything. Actually Friston understands this perfectly well; he immediately goes on to speak approvingly of Dennett's free-floating rationales, just the kind of intra-evolutionary purpose I mean. (He says Dennett has spent his career trying to understand the origin of the mind - what, is he one of those pesky guys who treat the mind as a thing?)

Anyway, now we're getting nearer to the real point; inference. Inference, claims Friston, is quite close to a theory of everything (maybe, but so is 'stuff happens'). First, though, it seems we need to talk about complex systems, and we're going to do it by talking about their state spaces. I wish we weren't. Really, this business of state spaces is like a fashion - or a disease - sweeping through certain parts of the intellectual community. Is there an emerging belief, comparable to the doctrine of the Universality of Computation, that everything must be usefully capable of analysis by state space? I might be completely up the creek , but it seems to me it's all too easy to use hypothetical state spaces to give yourself a false assurance that something is tractable when it really isn't. Of course proper defined state spaces are perfectly legitimate in certain cases. Friston mentions quantum systems; yes, but elementary particles have a relatively small number of properties that provide a manageable number of regular axes from which our space can be constructed. At least you've got to be able to say what variables you're mapping in your space, haven't you? Here it seems Friston wants to talk about, for example, the space of electrochemical states of the brain. I mean, he's a professor of neurology, so he knows whereof he speaks - but what on earth are the unimaginable axes that define that one, using cleanly separated independent variables? He's very hand-wavy about this; he half-suggests we might be dealing with a state space detailing the position of every particle, but surely that's radically below the level of description we need even for neurology, never mind inferences about the world. It's highly likely that mental states are uniquely resistant to simple analysis; in the state space of human thoughts every one of the infinite possible thoughts may be adjacent to every other along one of the infinite number of salient axes. I doubt whether God could construct that state space meaningfully, not even even if we gave Him a pencil and paper.

Anyway, Friston wants us to consider a Lyapunov function applied to some suitable state space of - I think - a complex system such as one of us, ie a living organism. He describes the function in general terms and how it governs movement through the space, although without too much detail. In fact after a bit of a whistle stop tour of attractors and homeostasis all he seems to want from the discussion is the point that adaptive behaviour can be read as embodying inferences about the world the organism inhabits. We could get there just by talking about hibernation or bees building hives, so the unkind suspicion crossed my mind that he brings Lyapunov into it partly in order to have a scary dead Russian on the team. I'm sure that's unfair, and most likely there is illuminating detail about the Lyapunov function that didn't survive into this account because of limitations on space (or probable reader comprehension). In any case it seems that all we really need to take away is that this function in complex systems like us can be interpreted as making inferences about the future, or minimising 'surprise'.

It's important to be clear here that we're not actually talking literally about actual surprise or about actual inference, the process of some conscious entity inferring something. We're using the word metaphorically to talk about the free-floating explanations, in a Dennettian sense, that

complex, self-maintaining systems implicitly display. By acting in ways that keep them going, such systems can sort of be said to embody guesses about the future. Friston says these sorts of behaviour in complex systems 'effectively' make inferences about the world; this is one of those cases where we should remember that 'effectively' should strictly be read as 'don't'. It's absolutely OK to talk in these metaphorical terms - I'd go so far as to say it's almost unavoidable - but to talk of metaphorical inference in an account of consciousness, where we're trying to explain the real thing, raises the obvious risk of losing track and kidding ourselves that we've solved the problem of cognition when all we've done is invoke a metaphor. So let's call this kind of metaphorical inference 'minference'. If we want to claim later that inference is minference, or that we can build inference out of minference, well and good: but we'll be sure of noticing what we're doing.

So complex, self-sustaining systems like us do minference; but that tells us nothing about consciousness because it's true of all such systems. It's true of plants and bacteria just as much as it's true of us, and it's even true of some non-living systems. Of course, says Friston, but for similar reasons consciousness must also be a process of inference (minference). It's just the (m)inference done by the brain. That's fine, but it just tells us consciousness must tend to produce behaviour that helps us stay alive, without telling us anything at all about the distinctive nature of the process; digestion is also a process that does minference, isn't it? But we don't usually attribute consciousness to the gut. Brain minference is entirely different to conscious explicit inferences.

Friston does realise that he needs to explain the difference between conscious and non-conscious minferring creatures, but I don't think he's clear enough that the earlier talk of (m)inference is no real use to him. He suggests that in order to infer (really infer) the consequences of its actions, a creature needs an internal model. This seems quite problematic to me, though I'm led to believe he has a more extensive argument for it which doesn't appear here. While we may use models for some purposes, I honestly don't see that inference requires one (in fact, building a model and then making your inferences about that would be asking for trouble). I plan to go and catch a train in a minute, having inferred that there will be one at the station; does that mean I must have a small train set or a miniature timetable simulated in my brain? Nope. Friston wants to say that the difference between conscious and unconscious behaviour is that the former derives from a 'thick' model of time, which here seems to mean no more than one that takes account of a relatively extended period. The idea that the duration is crucial makes no great sense to me: the behaviour patterns of ants reflect hundreds of thousands or even millions of years of minference: my conscious decisions may be the work of a moment; but I think in the end what Friston means to say is that conscious thought detaches us from the immediate moment by modelling the world and so allows us to entertain plans motivated by long-term considerations. That's fine, but it has nothing much to do with the state spaces, attractors and Lyapunov functions discussed earlier; it looks as if we can junk all that and just start afresh with the claim about consciousness being a matter of a model that helps us plan. And once that idea is shorn of all the earlier apparatus it becomes clear that it's not an especially new insight. In fact, it's pretty much the sort of thing a lot of those pesky mind-as-thing fellows have been saying all along.

Alas, it's worse than that. Because of the confusion between inference and minference Friston seems to be saddled with the idea that actual consciousness is about minimising actual surprise. Is the animating purpose of human thought to avoid surprise? Do our explicit mental processes seek above all to attain a steady equilibrium, always thinking about the same few

things in a tight circle and getting away from new ideas as quickly as possible? It doesn't seem plausible to me.

Friston concludes with these two sentences.

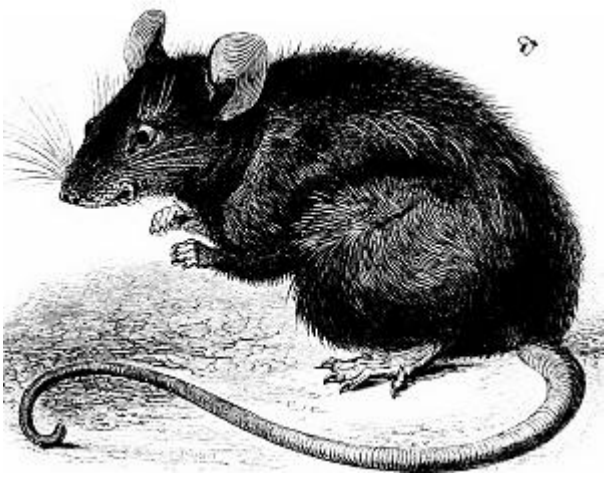
> *There's no real reason for minds to exist; they appear to do so simply because existence itself is the end-point of a process of reasoning. Consciousness, I'd contend, is nothing grander than inference about my future.*

Frankly, I don't understand the first one; the second is fine, but I'm left with the nagging feeling that I missed something more exciting along the way?

[The picture is Lyapunov, by the way, not Karl Friston]

Symptoms of Consciousness

28 May 2017



Where did consciousness come from? A recent piece in New Scientist reviewed a number of ideas about the evolutionary origin and biological nature of consciousness. The article obligingly offered a set of ten criteria for judging whether an organism is conscious or not...

- Recognises itself in a mirror
- Has insight into the minds of others
- Displays regret having made a bad decision
- Heart races in stressful situations
- Has many dopamine receptors in its brain to sense reward
- Highly flexible in making decisions
- Has ability to focus attention (subjective experience)
- Needs to sleep
- Sensitive to anaesthetics
- Displays unlimited associative learning

This is clearly a bit of a mixed bag. One or two of these have a clear theoretical base; they could be used as the basis of a plausible definition of consciousness. Having insight into the minds of others ('theory of mind') is one, and unlimited associative learning looks like another. But robots and aliens need not have dopamine receptors or racing hearts, yet we surely wouldn't rule out their being conscious on that account. The list is less like notes towards a definition and more of a collection of symptoms.

They're drawn from some quite different sources, too. The idea that self-awareness and awareness of the minds of others has something to do with consciousness is widely accepted and the piece alludes to some examples in animals. A chimp shown a mirror will touch a spot that had been covertly placed on its forehead, which is (debatably) said to prove it knows that the reflection is itself. A scrub jay will re-hide food if it was seen doing the hiding the first time - unless it was seen only by its own mate. A rat that pressed the wrong lever in an experiment will, it seems, gaze regretfully at the right one ('What do you do for a living?' 'Oh, I assess the level of regret in a rat's gaze.') Self-awareness certainly could constitute consciousness if higher-order theories are right, but to me it looks more like a product of consciousness and hence a symptom, albeit a pretty good one.

Another possibility is hedonic variation, here championed by Jesse Prinz and Bjørn Grinde. Many animals exhibit a raised heart rate and dopamine levels when stimulated - but not amphibians or fish (who seem to be getting a bad press on the consciousness front lately). There's a definite theoretical insight underlying this one. The idea is that assigning pleasure to some outcomes and letting that drive behaviour instead of just running off fixed patterns instinctively, allows an extra degree of flexibility which on the whole has a positive survival value. Grinde apparently thinks there are downsides too and on that account it's unlikely that consciousness evolved more than once. The basic idea here seems to make a lot of sense, but the dopamine stuff

apparently requires us to think that lizards are conscious while newts are not. That seems a fine distinction, though I have to admit that I don't have enough experience of newts to make the judgement (or of lizards either if I'm being completely honest).

Bruno van Swinderen has a different view, relating consciousness to subjective experience. That, of course, is notoriously unmeasurable according to many, but luckily van Swinderen thinks it correlates with selective attention, or indeed is much the same thing. Why on earth he thinks that remains obscure, but he measures selective attention with some exquisitely designed equipment plugged into the brains of fruit flies. ('Oh, you do rat regret? I measure how attentively flies are watching things.')

Sleep might be a handy indicator, as van Swinderen believes it is creatures that do selective attention that need it. They also, from insects to vertebrates (fish are in this time), need comparable doses of anaesthetic to knock them out, whereas nematode worms need far more to stop them in their tracks. I don't know whether this is enough. I think if I were shown a nematode that had finally been drugged up enough to make it keep still, I might be prepared to say it was unconscious; and if something can become unconscious, it must previously have been conscious.

Some think by contrast that we need a narrower view; Michael Graziano reckons you need a mental model, and while fish are still in, he would exclude the insects and crustaceans van Swinderen grants consciousness to. Eva Jablonka thinks you need unlimited associative learning, and she would let the insects and crustaceans back in, but hesitates over those worms. The idea behind associative learning is again that consciousness takes you away from stereotyped behaviour and allows more complex and flexible responses - in this case because you can, for example, associate complex sets of stimuli and treat them as one new stimulus, quite an appealing idea.

Really it seems to me that all these interesting efforts are going after somewhat different conceptions of consciousness. I think it was Ned Block who called it a 'mongrel' concept; there's little doubt that we use it in very varied ways, to describe the property of a worm that's still moving at one end, to the ability to hold explicit views about the beliefs of other conscious entities at the other. We don't need one theory of consciousness; we need a dozen.

Morality and Consciousness

3 June 2017

What is the moral significance of consciousness? Jim Davies addressed the question in a short but thoughtful piece recently.

Davies quite rightly points out that although the nature of consciousness is often seen as an academic matter, remote from practical concerns, it actually bears directly on how we treat animals and each other (and of course, robots, an area that was purely theoretical not that long ago but becomes more urgently practical by the day). In particular, the question of which entities are to be regarded as conscious is potentially decisive in many cases.

There are two main ways my consciousness affects my moral status. First, if I'm not conscious, I can't be a moral subject, in the sense of being an agent (perhaps I can't anyway, but if I'm not conscious it really seems I can't get started). Second, I probably can't be a moral object either; I don't have any desires that can be thwarted and since I don't have any experiences, I can't suffer or feel pain.

Davies asks whether we need to give plants consideration. They respond to their environment and can suffer damage, but without a nervous system it seems unlikely they feel pain. However, pain is a complex business, with a mix of simple awareness of damage, actual experience of that essential bad thing that is the experiential core of pain, and in humans at least, all sorts of other distress and emotional response. This makes the task of deciding which creatures feel pain rather difficult, and in practice guidelines for animal experimentation rely heavily on the broad guess that the more like humans they are, the more we should worry. If you're an invertebrate, then with few exceptions you're probably not going to be treated very tenderly. As we come to understand neurology and related science better, we might have to adjust our thinking. This might let us behave better, but it might also force us to give up certain fields of research which are useful to us.

To illustrate the difference between mere awareness of harm and actual pain, Davies suggests the example of watching our arm being crushed while heavily anaesthetised (I believe there are also drugs that in effect allow you to feel the pain while not caring about it). I think that raises some additional fundamental issues about why we think things are bad. You might indeed sit by and watch while your arm was crushed without feeling pain or perhaps even concern. Perhaps we can imagine that for some reason you're never going to need your arm again (perhaps now you have a form of high-tech psychokinesis, an ability to move and touch things with your mind that simply outclasses that old-fashioned 'arm' business), so you have no regrets or worries.

Even so, isn't there just something bad about watching the destruction of such a complex and well-structured limb?

Take a different example; everyone but you is dead and you know no-one is ever coming back, not even any aliens. Your own existence is completely secure for the rest of your natural life, and you spend your time blowing up buildings and destroying works of art, for no particular reason. It doesn't hurt anyone and cannot have any consequences, but doesn't your casual destructiveness still seem bad?

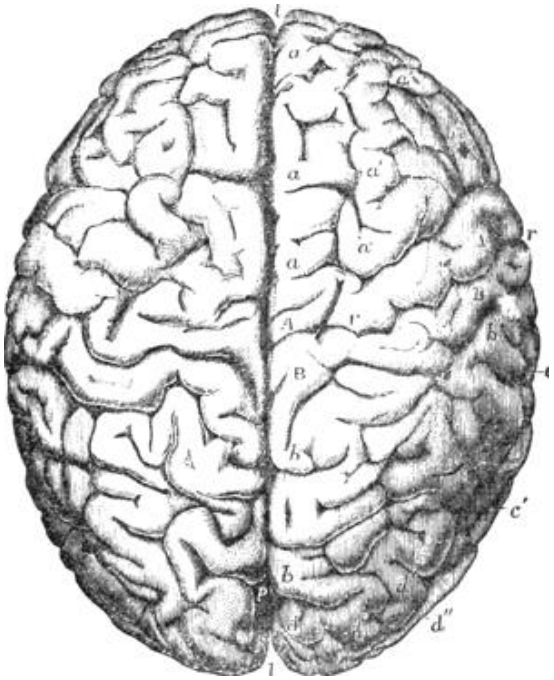
I'd like to argue that there is a badness to destruction over and above its consequential impact, but it's difficult to construct a pure example, and I know many people simply don't share my intuition. It is admittedly difficult because there's always the likelihood that one's intuitions are contaminated by ingrained assumptions about things having utility. I'd like to say there's a real moral rule that favours more things and more organisation, but without appealing to consequentialist arguments it's hard for me to do much more than note that in fact moral codes tend to inhibit destruction and favour its opposite.

However, if my gut feeling is right, it's quite important, because it means the largely utilitarian grounds used for rules about animal research and some human matters, re not quite adequate after all; the fact that some piece of research causes no pain is not necessarily enough to stop its destructive character bring bad.

It's probably my duty to work on my intuitions and arguments a bit more, but that's hard to do when you're sitting in the sun with a beer in the charming streets of old Salamanca.

Copy the brain?

12 June 2017



A whole set of interesting articles from IEEE Spectrum explore the question of whether AI can and should copy the human brain more. Of course, so-called neural networks were originally inspired by the way the brain works, but they represent a drastic simplification of a partial understanding. In fact, they are so unlike real neurons it's really rather remarkable that they turn out to perform useful processes at all. Karlheinz Meier provides a useful review of the development of neuromorphic computing, up to contemporary chips with impressive performance.

Jeff Hawkins suggests the brain is better in three ways. First, it learns by rewiring; by growing new synapses. This confers three benefits: learning that is fast, incremental, and continuous. The brain does not need to do lengthy retraining to learn new things. Remarkably, he says a single neuron can do substantial pieces of pattern recognition and acquire 'knowledge' of several patterns without them interfering with each other.

The second way in which the brain is better is that it uses sparse distributed representations; a particular idea such as 'cat' can be represented by a large number of neurons, with only a small percentage needing to be active at any one time. This makes the system robust in respect of noise and damage, but because some of the 'cat' neurons may play roles in the representation of other animals and other entities, it also makes it quick and efficient at recognising similarities and dealing with vague ideas (an animal in the bush which may or may not be a cat).

The third thing the brain does better, according to Hawkins, is sensorimotor integration. He makes the interesting claim that the brain effectively does this all over, as part of basic ordinary activity, not as a specialised central function. Instead of one big 3D model of the world, we have what amounts to little ones everywhere. This is interesting partly because it is, prima facie, so implausible. Doing your modelling a hundred or a million times over is going to use up a lot of energy and 'processing power', and it raises the obvious risk of inconsistency between models. But Hawkins says he has a detailed theory of how it works, and you'd have to be bold to dismiss his claim.

There are several other articles, all worth a look. Actually, there are several different reasons we might want to imitate the brain. We might want computers that can interface with humans better because, in part, they work in similar ways. We might want to understand the brain better and be able to test our understanding; an ability that might have real benefits for treating brain disease and injury, and to some degree make up for the ethical limitations on the experiments we can perform on humans. The main focus here, though, is on learning how to do those things the brain does so well, but which still cannot yet be done efficiently, or in some cases at all, by computers.

As a strategy, copying the brain has several drawbacks. First, we still don't understand the brain well enough. Things have moved on greatly in recent years, but in some ways that just shows how limited our understanding was to begin with. There's a significant danger that by imitating the brain without understanding, we end up reproducing features that are functionally irrelevant; features the brain has for chance evolutionary reasons. Do we need a brain divided into two halves, as those of vertebrates generally are, or is that unimportant? Second, one thing we do know is that the brain is extraordinarily complex and finely structured. We are never going to reproduce all that in full detail - but perhaps it doesn't matter; we've never replicated the exquisite engineering of feather technology either, but it didn't stop us doing flight or understanding birds.

I think the challenge of understanding the brain is unique, but trying to copy it is probably an increasingly productive strategy.

It's All Theology

18 June 2017

Consciousness: it's all theology in the end. Or philosophy, if the distinction matters. Beyond a certain point it is, at any rate; when we start asking about the nature of consciousness, about what it really is. That, in bald summary, seems to be the view of Robert A. Burton in a resigned piece. Burton is a neurologist, and he describes how Wilder Penfield first provided really hard and detailed evidence of close, specific correlations between neural and mental events, by showing that direct stimulation of neurons could evoke all kinds of memories, feelings, and other vivid mental experiences. Yet even Penfield, at the close of his career did not think the mind, the spirit of man, could yet be fully explained by science, or would be any time soon; in the end we each had to adopt personal assumptions or beliefs about that.

That is more or less the conclusion Burton has come to. Of course he understands the great achievements of neuroscience, but in the end the Hard Problem, which he seems to interpret rather widely, defeats analysis and remains a matter for theology or philosophy, in which we merely adopt the outlook most congenial to us. You may feel that's a little dismissive in relation to both fields, but I suppose we see his point. It will be no surprise that Burton dislikes Dennett's sceptical outlook and dismissal of religion. He accuses him of falling into a certain circularity: Dennett claims consciousness is an illusion, but illusions, after all, require consciousness.

We can't, of course, dismiss the issues of consciousness and selfhood as completely irrelevant to normal life; they bear directly on such matters as personal responsibility, moral rights, and who should be kept alive. But I think Burton could well argue that the way we deal with these issues in practice tends to vindicate his outlook; what we often see when these things are debated is a clash between differing sets of assumptions rather than a skilful forensic duel of competing reasoning.

Another line of argument that would tend to support Burton is the one that says worries about consciousness are largely confined to modern Western culture. I don't know enough for my own opinion to be worth anything, but I've been told that in classical Indian and Chinese thought the issue of consciousness just never really arises, although both traditions have long and complex philosophical traditions. Indeed, much the same could be said about Ancient Greek philosophy, I think; there's a good deal of philosophy of mind, but consciousness as we know it just doesn't really present itself as a puzzle. Socrates never professed ignorance about what consciousness was.

A common view would be that it's only after Descartes that the issue as we know it starts to take shape, because of his dualism; a distinction between body and spirit that certainly has its roots in the theological (and philosophical) traditions of Western Christianity. I myself would argue that the modern topic of consciousness didn't really take shape until Turing raised the real possibility of digital computers; consciousness was recruited to play the role of the thing computers haven't got, and our views on it have been shaped by that perspective over the last fifty years in particular. I've argued before that although Locke gives what might be the first recognisable version of the Hard Problem, with an inverted spectrum thought experiment, he

actually doesn't care about it much and only mentions it as a secondary argument about matters that to him, seemed more important.

I think it is true in some respects that as William James said, consciousness is the last remnant of the vanishing soul. Certainly, when people deny the reality of the self, it often seems to me that their main purpose is to deny the reality of the soul. But I still believe that Burton's view cedes too much to relativism - as I think Fodor once said, I hate relativism. We got into this business - even the theologians - because we wanted the truth, and we're not going to be fobbed off with that stuff! Scientists may become impatient when no agreed answer is forthcoming after a couple of centuries, but I cling to the idea that there is a truth if the matter about personhood, freedom, and consciousness. I recognise that there is in this, ironically, a tinge of an act of faith, but I don't care.

Unfortunately, as always things are probably more complicated than that. Could freedom of the will, say, be a culturally relative matter? All my instincts say no, but if people don't believe themselves to be free, doesn't that in fact impose some limits on how free they really are? If I absolutely do not believe I'm able to touch the sacred statue, then although the inability may be purely psychological, couldn't it be real? It seems there are at least some fuzzy edges. Could I abolish myself by ceasing to believe in my own existence? In a way you could say that is in caricature form what Buddhists believe (though they think that correctly understood, my existence was a delusion anyway). That's too much for me, and not only because of the sort of circularity mentioned above; I think it's much too pessimistic to give up on a good objective accounts of agency and selfhood.

There may never be a single clear answer that commands universal agreement on these issues, but then there has never been, and never will be, complete agreement about the correct form of human government; but we have surely come on a bit in the last thousand years. To abandon hope of similar gradual consensual progress on consciousness might be to neglect our civic duty. It follows that by reading and thinking about this, you are performing a humane and important task. Well done!

Panpsychism Vindicated?

25 June 2017

Could the Universe be conscious? This might seem like one of those Interesting Questions To Which The Answer Is 'No' that so often provide arresting headlines in the popular press. Since the Universe contains everything, what would it be conscious of? What would it think about? Thinking about itself - thinking about any real thing - would be bizarre, analogous to us thinking about the activity of the neurons that were doing the thinking. But I suppose it could think about imaginary stuff. Perhaps even the cosmos can dream; perhaps it thinks it's Cleopatra or Napoleon.

Actually, so far as I can see no-one is actually suggesting the Universe as a whole, as an entity, is conscious. Instead this highly original paper by Gregory L. Matloff starts with panpsychism, a belief that there is some sort of universal field of proto-consciousness permeating the cosmos. That is a not unpopular outlook these days. What's startling is Matloff's suggestion that some stars might be able to do roughly what our brains are supposed by panpsychists to do; recruit the field and use it to generate their own consciousness, exerting some degree of voluntary control over their own movements.

He relies on a phenomenon called Parenago's discontinuity; cooler, less massive stars seem to circle the galaxy a bit faster than the others. Dismissing a couple of rival explanations, he suggests that these cooler stars might be the ones capable of hosting consciousness and might be capable of shooting jets from their interior in a consistent direction so as to exert an influence over their own motion. This might be a testable hypothesis, bringing panpsychism in from the realms of philosophy to astrophysics (unkind people might suggest that the latter is not that much less freely speculative than former; I couldn't possibly comment).

In discussing panpsychism it is good to draw a distinction between types of consciousness. There is a certain practical decision-making capacity in human consciousness that is relatively well rooted in science in several ways. We can see roughly how it emerged from biological evolution and why it is useful, and we have at least some idea of how neurons might do it, together with a lot of evidence that in fact, they do do it. Then there is the much mistier business of subjective experience, what being conscious is actually like. We know little about that and it raises severe problems. I think it would be true to claim that most panpsychists think the kind of awareness that suffuses the world is of the latter kind; it is a dim general awareness, not a capacity to make snappy decisions. It is, in my view, one of the big disadvantages of panpsychism that it does not help much either explaining the practical, working kind of consciousness and in fact arguably leaves us with more to explain than we had on our plate to start with.

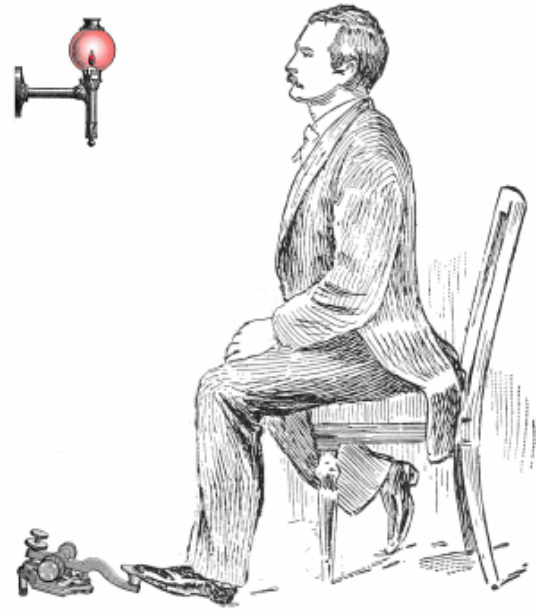
Anyway, if Parenago's theory is to be plausible, he needs to explain how stars could possibly build the decision-making kind of consciousness, and how the universal field would help. To his credit he recognises this - stars surely don't have neurons - and offers at least some hints about

how it might work. If I've got it right, the suggestion is that the universal field of consciousness might be identified with vacuum fluctuation pressures, which on the one hand might influence the molecules present in regions of the cooler stars under consideration, and on the other have effects within neurons more or less on Penrose/Hameroff lines. This is at best an outline, and raises immediate and difficult questions; why would vacuum fluctuation have anything to do with subjective experience? If a bunch of molecules in cool suns is enough for conscious volition, why doesn't the sea have a mind of its own? And so on. For me the deadliest questions are the simplest. If stars have conscious control of their movements, why are they all using it the same way - to speed up their circulation a bit? You'd think if they were conscious they would be steering around in different ways according to their own choices. Then again, why would they choose to do anything? As animals we need consciousness to help us pursue food, shelter, reproduction, and so on. Why would stars care which way they went?

I want to be fair to Matloff, because we shouldn't mock ideas merely for bring unconventional. But I see one awful possibility looming. His theory somewhat recalls medieval ideas about angels moving the stars in perfect harmony. They acted in a co-ordinated way because although the angels had wills of their own, they subjected them to God's. Now, why are the cool stars apparently all using their wills in a similarly co-ordinated way? Are they bound together through the vacuum fluctuations; have we finally found out there the physical manifestation of God? Please, please, nobody go in that direction!

Limits Of Free Won't

2 July 2017



Another strange side light on free will. Some of the most-discussed findings in the field are Libet's celebrated research which found that Readiness Potentials (RPs) in the brain showed when a decision to move had been made, significantly before the subject was aware of having decided. Libet himself thought this was problematic for free will, but that we could still have 'Free Won't' - we could still change our minds after the RP had appeared and veto the planned movement.

A new paper (discussed here by Jerry Coyne) follows up on this and seems to show that while we do have something like this veto facility, there is a time limit on that too, and beyond a certain point the planned move will be made regardless.

The actual experiment was in three phases. Subjects were given a light and a pedal and set up with equipment to detect RPs in their brain. They were told to press the pedal at a time of their choosing when the light was green, but not when it had turned red. The first run merely trained the equipment to detect RPs, with the light turning red randomly. In the second phase, the light turned red when an RP was detected, so that the subjects were in effect being asked to veto their own decision to press. In the third phase, they were told that their decisions were being predicted and they were asked to try to be unpredictable.

Detection of RPs actually took longer in some instances than others. It turned out that where the RP was picked up early enough, subjects could exercise the veto; but once the move was 200ms or less away, it was impossible to stop.

What does this prove, beyond the bare facts of the results? Perhaps not much. The conditions of the experiment are very strange and do not resemble everyday decision-making very much at all. It was always an odd feature of Libet's research that subjects were asked to get ready to move but choose the time capriciously according to whim; not a mental exercise that comes up very often in real life. In the new research, subjects further have to stop when the light is red; they don't, you notice, choose to veto their move, but merely respond to a pre-set signal. Whether this deserves to be called free won't is debatable; it isn't a free decision making process. How could it be, anyway; how could it be that deciding to do something takes significantly longer than deciding not to do the same thing? Is it that decisions to move are preceded by an RP, but other second-order decisions about those decisions are not? We seem to be heading into a maze of complications if we go that way and substantially reducing the significance of Libet's results.

Of course, if we don't think that Libet's results dethrone free will in the first place, we need not be very worried. My own view is that we need to distinguish between making a conscious decision and becoming aware of having made the decision. Some would argue that that second-order awareness is essential to the nature of conscious thought, but I don't think so. For me Libet's original research showed only that deciding and knowing you've decided are distinct,

and the latter naturally follows after the former. So assuming that, like me, you think it's fine to regard the results of certain physical processes as 'free' in a useful sense, free will remains untouched. If you were always a sceptic then of course Libet never worried you anyway, and nor will the new research.

It's Out There

9 July 2017

Where is consciousness? It's out there, apparently, not in here. There has been an interesting dialogue series going on between Riccardo Manzotti and Tim Parks in the NYRB (thanks to Tom Clark for drawing my attention to it)

We discussed Manzotti's views back in 2006, when with Honderich and Tonneau he represented a new wave of externalism. His version seemed to me perhaps the clearest and most attractive back then (though I think he's mistaken). He continues to put some good arguments.

In the first part, Manzotti says consciousness is awareness, experience. It is somewhat mysterious - we mustn't take for granted any view about a movie playing in our head or the like - and it doesn't feature in the scientific account. All the events and processes described by science could, it seems, go on without conscious experience occurring.

He is scathing, however, about the view that consciousness is therefore special (surely something that science doesn't account for can reasonably be seen as special?), and he suggests the word "mental" is a kind of conceptual dustbin for anything we can't accommodate otherwise. He and Parks describe the majority of views as internalist, dedicated to the view that one way or another neural activity just is consciousness. Many neural correlates of consciousness have been spotted, says Manzotti, but correlates ain't the thing itself.

In the second part he tackles colour, one of the strongest cards in the internalist hand. It looks to us as if things just have colour as a simple property, but in fact the science of colour tells us it's very far from being that simple. For one thing how we perceive a colour depends strongly on what other colours are adjacent; Manzotti demonstrates this with a graphic where areas with the same RGB values appear either blue or green. Examples like this make it very tempting to conclude that colour is constructed in the brain, but Manzotti boldly suggests that if science and ordinary understanding are at odds, so much the worse for science. Maybe we ought to accept that those colours really are different and be damned to RGB values.

The third dialogue attacks the metaphor of a computer often applied to the brain, and rejects talk of information processing. Information is not a physical thing, says Manzotti, and to speak of it as though it were a visible fluid passing through the brain risks dualism; something Tononi, with his theory of integrated information, accepts; he agrees that his ideas about information having two aspects point that way.

So what's a better answer? Manzotti traces externalist ideas back to Aristotle but focuses on the more ideas of affordances and enactivism. An affordance is roughly a possibility offered to us by an object; a hammer offers us the possibility of hitting nails. This idea of bashing things does not need to be represented in the head, because it is out there in the form of the hammer. Enactivism develops a more general idea of perception as action but runs into difficulties in some cases such as that of dreams, where we seem to have experience without action; or

consider that licking a strawberry or a chocolate ice cream is the same action but yields very different experience.

To set out his own view, Manzotti introduces the 'metaphysical switchboard': one switch toggles whether subject and object are separate, the other whether the subject is physical or not. If they're separate, and we choose to make the subject non-physical, we get something like Cartesian dualism, with all the problems that entails. If we select 'physical' then we get the view of modern science; and that too seems to be failing. If subject and object are neither separate nor physical, we get Berkleyan idealism; my perceptions actually constitute reality. The only option that works is to say that subject and object are identical, but physical; so when I see an apple, my experience of it is identical with the apple itself. Parks, rightly I think, says that most people will find this bonkers at first sight. But after all, the apple is the only thing that has apple-like qualities! There's no appliness in my brain or in my actions.

This raises many problems. My experience of the apple changes according to conditions, yet the apple itself doesn't change. Oh no? says Manzotti, why not? You're just clinging to the subject/object distinction; let it go and there's no problem. OK, but if my experience of the apple is identical with the apple, and so is yours, then our experiences must be identical. In fact, since subject and object are the same, we must also be identical!

The answer here is curious. Manzotti points out that the physical quality of velocity is relative to other things; you may be travelling at one speed relative to me but a different one compared to that train going by. In fact, he says, all physical qualities are relative, so the apple is an apple experience relative to one animal (me) and at the same time relative to another in a different way. I don't think this ingenious manoeuvre ultimately works; it seems Manzotti is introducing an intermediate entity of the kind he was trying to dispel; we now have an apple-experience relative to me which is different from the one relative to you. What binds these and makes them experiences of the same apple? If we say nothing, we fall back into idealism; if it's the real physical apple, then we're more or less back with the traditional framework, just differently labelled.

What about dreams and hallucinations? Manzotti holds that they are always made up out of real things we have previously experienced. Hey, he says, if we just invent things and colour is made in the head, how come we never dream new colours? He argues that there is always an interval between cause and effect when we experience things; given that, why shouldn't real things from long ago be the causes of dreams?

And the self, that other element in the traditional picture? It's made up of all the experiences, all the things experienced, that are relative to us; all physical, if a little scattered and dare I say metaphysically unusual; a massive conjunction bound together by… nothing in particular? Of course, the body is central, and for certain feelings, or for when we're in a dark, silent room, it may be especially salient. But it's not the whole thing, and still less is the brain.

In the latest dialogue, Manzotti and Parks consider free will. For Manzotti, having said that you are the sum of your experiences, it is straightforward to say that your decisions are made by the subset of those experiences that are causally active; nothing that contradicts determinist physics, but a reasonable sense in which we can say your act belonged to you. To me this is a relatively appealing outlook.

Overall? Well, I like the way externalism seeks to get rid of all the problems with mediation that lead many people to think we never experience the world, only our own impressions of it.

Manzotti's version is particularly coherent and intelligible. I'm not sure his clever relativity finally works though. I agree that experience isn't strictly in the brain, but I don't think it's in the apple either; to talk about its physical location is just a mistake. The processes that give rise to experience certainly have a location, but in itself it just doesn't have that kind of property.

The Thought That Counts

16 July 2017

Neurocritic asks a great question here, neatly provoking that which he would have defined - thought. What is thought, and what are individual thoughts? He quotes reports that we have an estimated 70,000 thoughts a day and justly asks how on earth anyone knows. How can you count thoughts?

Well, we like a challenge round here, so what is a thought? I'm going to lay into this one without showing all my working (this is after all a blog post, not a treatise), but I hope to make sense intermittently. I will start by laying it down axiomatically that a thought is about or of something. In philosophical language, it has intentionality. I include perceptions as thoughts, though more often when we mention thoughts we have in mind thoughts about distant, remembered or possible things rather than ones that are currently present to our senses. We may also have in mind thoughts about perceptions or thoughts about other thoughts - in the jargon, higher-order thoughts.

Now I believe we can say three things about a thought at different levels of description. At an intuitive level, it has content. At a psychological level it is an act of recognition; recognition of the thing that forms the content. And at a neural level it is a pattern of activity reliably correlated with the perception, recollection, or consideration of the thing that forms the content; recognition is exactly this chiming of neural patterns with things (What exactly do I mean by 'chiming of neural patterns'? No time for that now, move along please!). Note that while a given pattern of neural activity always correlates with one thought about one thing, there will be many other patterns of neural activity that correlate with slightly different thoughts about that same thing - that thing in different contexts or from different aspects. A thought is not uniquely identifiable by the thing it is about (we could develop a theory of broader content which would uniquely identify each thought, but that would have weird consequences so let's not). Note also that these 'things' I speak of may be imaginary or abstract entities as well as concrete physical objects: there are a lot of problems connected with that which I will ignore here.

So what is one thought? It's pretty clear intuitively that a thought may be part of a sequence which itself would also normally be regarded as a thought. If I think about going to make a cup of tea I may be thinking of putting the kettle on, warming the pot, measuring out the tea, and so on; I've had several thoughts in one way but in another the sequence only amounts to a thought about making tea. I may also think about complex things; when I think of the teapot I think of handle, spout, and so on. These cases are different in some respects, though in my view they use the same mechanism of linking objects of thought by recognising an over-arching entity that includes them. This linkage by moving up and down between recognition of larger and smaller entities is in my view what binds a train of thought together. Sitting here I perceive a small sensation of thirst, which I recognise as a typical initial stage of the larger idea of having a drink. One recognisable part of having a drink may be making the tea, part of which in turn involves the recognisable actions of standing up, going to the kitchen... and so on. However, great care must

be taken here to distinguish between the things a thought contains and the things it implies. If we allow implication then every thought about a cup of tea implies an indefinitely expanding set of background ideas and every thought has infinite content.

Nevertheless, the fact that sequences can be amalgamated suggests that there is no largest possible thought. We can go on adding more elements. There's a strong analogy here with the formation of sentences when speaking or writing. A thought or a sentence tends to run to a natural conclusion after a while, but this seems to arise partly because we run out of mental steam, and partly because short thoughts and short sentences are more manageable and can together do anything that longer ones can do. In principle a sentence could go on indefinitely, and so could a thought. Indeed, since the thread of relevance is weakened but not (we hope) lost at each junction between sentences or thoughts, we can perhaps regard whole passages of prose as embodying a single complex thought. The Decline and Fall of the Roman Empire is arguably a single massively complicated thought that emerged from Gibbon's brain over an unusually extended period, having first sprung to mind as he 'sat musing amidst the ruins of the Capitol, while the barefoot friars were singing vespers in the Temple of Jupiter'.

Parenthetically I throw in the speculation that grammatical sentence structure loosely mirrors the structure of thought; perhaps particular real world grammars emerge from the regular bashing together of people's individual mental thought structures, with all the variable compromise and conventionalisation that that would involve.

Is there a smallest possible thought? If we can go on putting thoughts together indefinitely, like more and more complex molecules, is there a level at which we get down to thoughts like atoms, incapable of further division without destruction?

As we enter this territory, we walk among the largely forgotten ruins of some grand projects of the past. People as clever as Leibniz once thought we might manage to define a set of semantic primitives, basic elements out of which all thoughts must be built. The idea, intuitively, was roughly that we could take the dictionary and define each word in terms of simpler ones; then define the words in the definitions in ones that were simpler still, until we boiled everything down to a handful of basics which we sort of expected to be words encapsulating elementary concepts of physics, ethics, maths, and so on.

Of course, it didn't work. It turns out that the process of definition is not analytical but expository. At the bottom level our primitives turn out to contain concepts from higher layers; the universe by transcendence and slippery lamination eludes comprehensive categorisation. As Borges said:

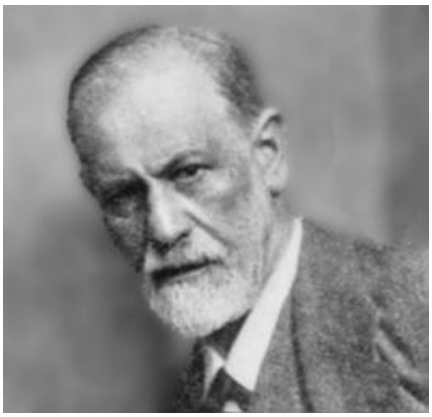
> *It is clear that there is no classification of the Universe that is not arbitrary and full of conjectures. The reason for this is very simple: we do not know what kind of thing the universe is. We can go further; we suspect that there is no universe in the organic, unifying sense of that ambitious word*.

That doesn't mean there is no smallest thought in some less ambitious sense. There may not be primitives, but to resurrect the analogy with language, there might be words. If, as I believe, thoughts correlate with patterns of neural activity, it follows that although complex thoughts may arise from patterns that evolve over minutes or even years (like the unimaginably complex sequence of neural firing that generated Gibbon's masterpiece), we could in principle look at a snapshot and have our instantaneous smallest thought.

It still isn't necessarily the case that we could count atomic thoughts. It would depend whether the brain snaps smartly between one meaningful pattern and another, as indeed language does between words, or smooshes one pattern gradually into another. (One small qualification to that is that although written and mental words seem nicely separated, in spoken language the sound tends to be very smooshy.) My guess is that it's more like the former than the latter (it doesn't feel as if thinking about tea morphs gradually into thinking about boiling water, more like a snappy shift from one to the other), but it is hard to be sure that that is always the case. In principle it's a matter that could be illuminated or resolved by empirical research, though that would require a remarkable level of detailed observation. At any rate no-one has counted thoughts this way yet and perhaps they never will.

Is there An Unconscious?

23 July 2017



Does the unconscious exist? David B. Feldman asks, and says no. He points out that the unconscious and its influence is a cornerstone of Freudian and other theories, where it is quoted as the explanation for our deeper motivation and our sometimes puzzling behaviour. It may send messages through dreams and other hints, but we have no direct access to it and cannot read its thoughts, even though they may heavily influence our personality.

Freud's status as an authority is perhaps not what it once was, but the unconscious is widely accepted as a given, pretty much part of our everyday folk-psychology understanding of our own minds. I think if you asked, a majority of people would say they had had direct experience of their own unconscious knowledge or beliefs affecting the way they behaved. Many psychological experiments have demonstrated 'priming' effects, where the subject's choices are affected by things they have been told or shown previously (although some of these may be affected by the reproducibility problems that have beset psychological research recently, I don't think the phenomenon of priming in general can be dismissed). Nor is it a purely academic matter. Unconscious bias is generally held to be a serious problem, responsible for the perpetuation of various kinds of discrimination by people who at a conscious level are fair-minded and well-meaning.

Feldman, however, suggests that the unconscious is neither scientifically testable nor logically sound. It may well be true that psychoanalytic explanations are scientifically slippery; mistaken predictions about a given subject can always be attributed to a further hidden motivation or complex, so that while one interpretation can be proved false, the psychoanalytic model overall cannot be. However, more generally there is good scientific evidence for unconscious influences on our behaviour as I've mentioned, so perhaps it depends on what kind of unconscious we're talking about. On the logical front, Feldman suggests that the unconscious is an 'homunculus'; an example of the kind of explanation that attributes some mental functions to 'a little man in your head', a mental module that is just assumed to be able to do whatever a whole brain can do. He quite rightly says that homuncular theories merely defer explanation in a way which is most often useless and unjustified.

But is he right? On one hand people like Dennett, as we've discussed in the past, have said that homuncular arguments may be alright with certain provisos; on the other hand, is it clear that the unconscious really is an homuncular entity? The key question, I think, is whether the unconscious is an entity that is just assumed to do all the sorts of things a complete mind would do. If we stick to Freud, Feldman's charges may have substance; the unconscious seems to have desires and motivations, emotions, and plans; it understands what is going on in our lives pretty well and can make intelligently targeted interventions and encode messages in complex ways. In a lot of ways it is like a complete person - or rather, like three people: id, ego, and superego. A Freudian might argue over that; however, in the final analysis it's not the decisive issue because we're not bound to stick to a Freudian or psychoanalytic reading of the

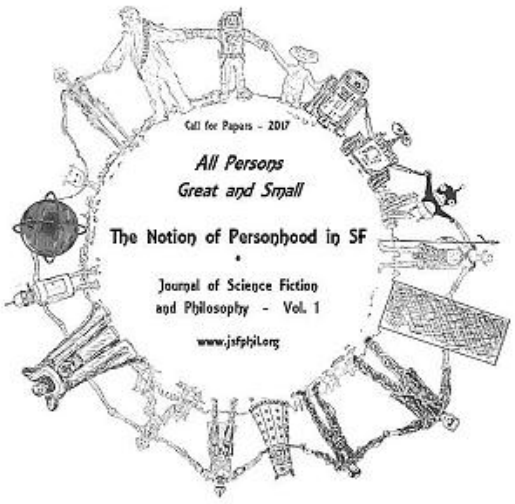
unconscious anyway. Again, it depends what kind of unconscious we're proposing. We could go for a much simpler version which does some basic things for us but at a level far below that a real homunculus. Perhaps we could even speak loosely of an unconscious if it were no more than the combined effect of many separate mental features?

In fact, Feldman accepts all this. He is quite comfortable with our doing things unconsciously, he merely denies the existence of the unconscious as a distinct coherent thinking entity. He uses the example of driving along a familiar route; we perform perfectly, but afterwards cannot remember doing the steering or changing gear at any stage. Myself I think this is actually a matter of memory, not inattention while actually driving - if we were stopped at any point in the journey I don't think we would have to snap out of some trance-like state; it's just that we don't remember. But in general Feldman's position seems entirely sensible.

There is actually something a little odd in the way we talk about unconsciousness. Virtually everything is unconscious, after all. We don't remark on the fact that muscles or the gut do their job without being conscious; it's the unique presence of consciousness in mental activity that is worthy of mention. So why do we even talk about unconscious functions, let alone an unconscious?

PhiMiFi

26 July 2017



If there's one thing philosophers of mind like more than an argument, it's a rattling good yarn. Obviously we think of Mary the Colour Scientist, Zombie Twin (and Zimboes, Zomboids, Zoombinis...), the Chinese Room (and the Chinese Nation), Brain in a Vat, Swamp-Man, Chip-Head, Twin Earth and Schmorses... even papers whose content doesn't include narratives at this celebrated level often feature thought-experiments that are strange and piquant. Obviously philosophy in general goes in for that kind of thing too - just think of the trolley problems that have been around forever but became inexplicably popular in the last year or so (I was probably force-fed too many at an impressionable age, and now I can't face them - it's like broccoli, really): but I don't think there's another field that loves a story quite like the Mind guys.

I've often alluded to the way novelists have been attacking the problems of minds by other means ever since the James Boys (Henry and William) set up their pincer movement on the stream of consciousness; and how serious novelists have from time to time turned their hand to exploring the theme consciousness with clear reference to academic philosophy, sometimes even turning aside to debunk a thought experiment here and there. We remember philosophically considerable works of genuine science fiction such as Scott Bakker's Neuropath. We haven't forgotten how Ian and Sebastian Faulks in their different ways made important contributions to the field of Bogus but Totally Convincing Psychology with De Clérambault's Syndrome and Glockner's Isthmus, nor David Lodge's book 'Consciousness and the Novel' and his novel Thinks. And philosophers have not been averse to writing the odd story, from Dan Lloyd's novel Radiant Cool up to short stories by many other academics including Dennett and Eric Schwitzgebel.

So I was pleased to hear (via a tweet from Eric himself) of the inception of an unexpected new project in the form of the Journal of Science Fiction and Philosophy. The Journal 'aims to foster the appreciation of science fiction as a medium for philosophical reflection'. Does that work? Don't science fiction and philosophy have significantly different objectives? I think it would be hard to argue that all science fiction is of philosophical interest (other than to the extent that everything is of philosophical interest). Some space opera and a disappointing amount of time travel narrative really just consists of adventure stories for which the SF premise is mere background. Some science fiction (less than one might expect) is actually about speculative science. But there is quite a lot that could almost as well be called Phifi as Scifi, stories where the alleged science is thinly or unconvincingly sketched, and simply plays the role of enabler for an examination of social, ethical, or metaphysical premises. You could argue that Asimov's celebrated robot short stories fit into this category; we have no idea how positronic brains are supposed to work, it's the ethical dilemmas that drive the stories.

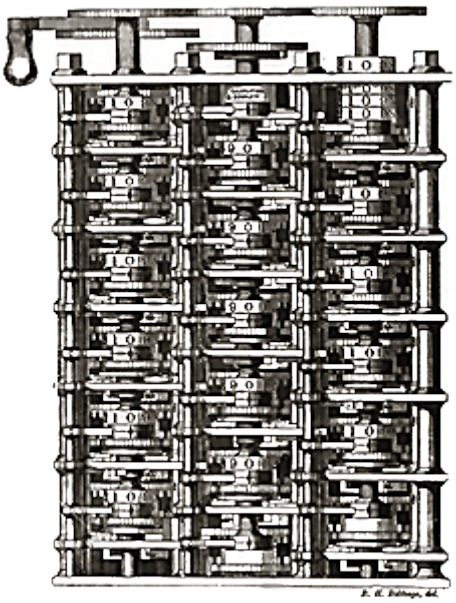
There is, then, a bit of an overlap; but surely SF and philosophy differ radically in their aims? Fiction aims only to entertain; the ideas can be rubbish so long as they enable the monsters or,

slightly better, boggle the mind, can't they? Philosophy uses stories only as part of making a definite case for the truth of particular positions, part of an overall investigative effort directed, however indirect the route, at the real world? There's some truth in that, but the line of demarcation is not sharp. For one thing, successful philosophers write entertainingly; I do not think either Dennett or Searle would have achieved recognition for their arguments so easily if they hadn't been presented in prose clear enough for non-academic readers to understand, and well-crafted enough to make them enjoy the experience. Moreover, philosophy doesn't have to present the truth; it can ask questions or just try to do some of that mind boggling. Myself when I come to read a philosophical paper I do not expect to find the truth (I gave up that kind of optimism along with the broccoli): my hopes are amply fulfilled if what I read is interesting. Equally, while fiction may indeed consist of amusing lies, novelists are not indifferent to the truth, and often want to advance a hypothesis, or at least, have us entertain one.

I really think some gifted novelist should take the themes of the famous thought-experiments and attempt to turn them into a coherent story. Meantime. there is every prospect that the new journal represents, not dumbing down but wising up, and I for one welcome our new peer-reviewers. (*I'm pleased to report that as of July 2024 the journal was still up and running*.)

Against Non-Computationalism

31 July 2017



Marcin Milkowski has produced a short survey of arguments against computationalism; his aim is in fact to show that they all fail and that computationalism is likely to be both true and non-trivial. The treatment of each argument is very brief; the paper could easily be expanded into a substantial book. But it's handy to have a comprehensive list, and he does seem to have done a decent job of covering the ground.

There are a couple of weaknesses in his strategy. One is just the point that defeating arguments against your position does not in itself establish that your position is correct. But in addition he may need to do more than he thinks. He says computationalism is the belief 'that the brain is a kind of information-processing mechanism, and that information-processing is necessary for cognition'. But I think some would accept that as true in at least some senses while denying that information-processing is sufficient for consciousness, or characteristic of consciousness. There's really no denying that the brain does computation, at least for some readings of 'computation' or some of the time. Indeed, computation is an invention of the human mind and arguably does not exist without it. But that does not make it essential. We can register numbers with our fingers, but while that does mean a hand is a digital calculator, digital calculation isn't really what hands do, and if we're looking for the secret of manipulation we need to look elsewhere.

Be that as it may, a review of the arguments is well worthwhile. The first objection is that computationalism is essentially a metaphor; Milkowski rules this out by definition, specifying that his claim is about literal computation. The second objection is that nothing in the brain seems to match the computational distinction between software and hardware. Rather than look for equivalents, Milkowski takes this one on the nose rather, arguing that we can have computation without distinguishing software and hardware. To my mind that concedes quite a large gulf separating brain activity from what we normally think of as computation.

No concessions are needed to dismiss the idea that computers merely crunch numbers; on any reasonable interpretation they do other things too by means of numbers, so this doesn't rule out their being able to do cognition. More sophisticated is the argument that computers are strictly speaking abstract entities. I suppose we could put the case by saying that real computers have computerhood in the light of their resemblance to Turing machines, but Turing machines can only be approximated in reality because they have infinite tape and and move between strictly discrete states, etc. Milkowski is impatient with this objection - real brains could be like real computers, which obviously exist - but reserves the question of whether computer symbols mean anything. Swatting aside the objection that computers are not biological, it's this interesting point about meaning that Milkowski tackles next.

He approaches the issue via the Chinese Room thought experiment and the 'symbol grounding problem'. Symbols mean things because we interpret them that way, but computers only deal

with formal, syntactic properties of data; how do we bridge the gap? Milkowski does not abandon hope that someone will naturalise meaning effectively, and mentions the theories of Millikan and Dretske. But in the meantime, he seems to feel we can accommodate some extra function to deal with meaning without having to give up the idea that cognition is essentially computational. That seems too big a concession to me, but if Milkowski set out his thinking in more depth it might perhaps be more appealing than it seems on brief acquaintance. Milkowski dismisses as a red herring Robert Epstein's argument from the inability of the human mind to remember what's on a dollar bill accurately (the way a computational mind would).

The next objection, derived from Gibson and Chemero, apparently says that people do not process information, they merely pick it up. This is not an argument I'm familiar with, so I might be doing it an injustice, but Milkowski's rejection seems sensible; only on some special reading of 'processing' would it seem likely that people don't process information.

Now we come to the argument that consciousness is not computational; that computation is just the wrong sort of process to produce consciousness. Milkowski traces it back to Leibniz's famous mill argument; the moving parts of a machine can never produce anything like experience. Perhaps we could put in the same camp Brentano's incredulity and modern mysterianism; Milkowski mentions Searle's assertion that consciousness can only arise from biological properties, not yet understood. Milkowski complains that if accepted, this sort of objection seems to bar the way to any reductive explanation (some of his opponents would readily bite that bullet).

Next up is an objection that computer models ignore time; this seems again to refer to the discrete states of Turing machines, and Milkowski dismisses it similarly. Next comes the objection that brains are not digital. There is in fact quite a lot that could be said on either side here, but Milkowski merely argues that a computer need not be digital. This is true, but it's another concession; his vision of brain computationalism now seems to be of analogue computers with no software; I don't think that's how most people read the claim that 'the brain is a computer'. I think Milkowski is more in tune with most computationalists in his attitude to arguments of the form 'computers will never be able to x' where x has been things like playing chess. Historically these arguments have not fared well.

Can only people see the truth? This is how Milkowski describes the formal argument of Roger Penrose that only human beings can transcend the limitations which every formal system must have, seeing the truth of premises that cannot be proved within the system. Milkowski invokes arguments about whether this transcendent understanding can be non-contradictory and certain, but at this level of brevity the arguments can really only be gestured at.

The objection Milkowski gives most credit to is the claimed impossibility of formalising common sense. It is at least very difficult, he concedes, but we seem to be getting somewhere. The objection from common sense is a particular case of a more general one which I think is strong; formal processes like computation are not able to deal with the indefinite realms that reality presents. It isn't just the Frame Problem; computation also fails with the indefinite ambiguity of meaning (the same problem identified for translation by Quine); whereas human communication actually exploits the polyvalence of meaning through the pragmatics of normal discourse, rich in Gricean implicatures.

Finally Milkowski deals with two radical arguments. The first says that everything is a computer; that would make computationalism true, but trivial. Well, says Milkowski, there's computation

and computation; the radical claim would make even normal computers trivial, which we surely don't want. The other radical case is that nothing is really a computer, or rather that whether anything is a computer is simply a matter of interpretation. Again this seems too destructive and too sweeping. If it's all a matter of interpretation, says Milkowski, why update your computer? Just interpret your trusty Windows Vista machine as actually running Windows 10 - or macOS, why not?

DID and 'Split': How We Talk About Kevin

6 August 2017

I finally got round to seeing *Split*, the M. Night Shyamalan film (spoilers follow) about a problematic case of split personality, and while it's quite a gripping film with a bravura central performance from James McAvoy, I couldn't help feeling that in various other ways it was somewhere between unhelpful and irresponsible. Briefly, in the film we're given a character suffering from Dissociative Identity Disorder (DID), the condition formerly known as 'Multiple Personality Disorder'. The working arrangement reached by his 'alters', the different personalities inhabiting an unfortunate character called Kevin Wendell Crumb, has been disturbed by two of the darker alters (there are 23); he kidnaps three girls and it gradually becomes clear that the 'Beast', a further (24th) alter is going to eat them.

DID has a chequered and still somewhat controversial history. I discussed it at moderate length here (Oh dear, tempus fugit) about eleven years ago. One of the things about it is that its incidence is strongly affected by cultural factors. It's very much higher in some countries than others, and the appearance of popular films or books about it seems to have a major impact, increasing the number of diagnoses in subsequent years. This phenomenon apparently goes right back to Jekyll and Hyde, an early fictional version which remains powerful in Anglophone culture. In fact Split itself draws on two notable features of Jekyll and Hyde: the ideas that some alters are likely to be wicked, and that they may differ in appearance and even size from the original. The number of cases in the US rose dramatically after the TV series Sybil, based on a real case, first aired (though subsequently doubts about the real-world diagnosis have emerged). It's also probable that the popular view has been influenced by the persistent misunderstanding that schizophrenia is having a 'split personality' (it isn't, although it's not unknown for DID patients to have schizophrenia too: and some 'Schneiderian' symptoms - voices, inserted thoughts - may confusingly arise from either condition.

One view is that while DID is undeniably a real mental condition, it is largely or wholly iatrogenic: caused by the doctors. On this view therapists trying to draw out alters for the best of reasons may simply be encouraging patients to confabulate them, or indeed the whole problem. On this view the cultural background may be very important in preparing the minds of patients (and indeed the minds of therapists: let's be honest, psychologists watch stupid films too). So the first charge against Split is that it is likely to cause another spike in the number of DID cases.

Is that a bad thing, though? One argument is that cultural factors don't cause the dissociative problems, they merely lead to more of them being properly diagnosed. One mainstream modern view sees DID as a response to childhood trauma; the sufferer generates a separate persona to deal with the intolerable pain. And often enough it works; we might see DID less as a mental problem and more as a strategy, often successful, for dealing with certain mental problems. There's actually no need to reintegrate the alters, any more than you would try to homogenise any other personality; all you need to do is reach a satisfactory working arrangement. If that's the case then making knowledge of DID more widely available might actually be a good thing.

That might be an arguable position, though we'd have to take some account of the potential for disruptive and amnesiac episodes that may come along with DID. However, Split can hardly be seen as making a valuable contribution to awareness because of the way it draws on Jekyll and Hyde tropes. First, there's the renewed suggestion that alters usually include terrifically evil personalities. The central character in Split is apparently going to become a super-villain in a sequel. This will be a 'grounded' super; one whose powers are not attributable to the semi-magic effects of radiation or film-style mutation, but 'realistically' to DID. Putting aside the super powers, I don't know of any evidence that people with DID have a worse criminal record than anyone else; if anything I'd guess that coping with their own problems leaves them no time or capacity for embarking on crime sprees. But portraying them as inherently bad inevitably stigmatises existing patients and deters future diagnoses in ways that are surely offensive and unhelpful. It might even cause some patients to think that their alters have to behave badly in order to validate their diagnosis.

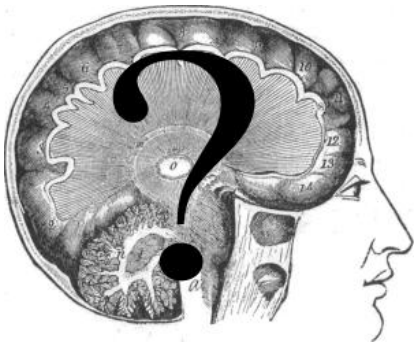
Of course, Hollywood almost invariably portrays mental problems as hidden superpowers. Autism makes you a mathematical genius; OCD means you're really tidy and well-organised. But the suggestion that DID probably makes you a wall-climbing murderer is an especially negative one. Zombies, those harmless victims of bizarre Caribbean brainwashing, possibly got a similarly negative treatment when they were transformed by Romero into brain-munching corpse monsters; but luckily I think that diagnosis is rare.

The other thing about Split is that it takes some of the wilder claims about the physical impact of DID and exaggerates them to absurdity. The psychologist in the film, Dr. Karen Fletcher merely asserts that the switch between alters can change people's body chemistry: fine, getting into an emotional state changes that. But it emerges that Kevin's eyesight, size and strength all change with his alters: one of them even needs insulin injections while the others don't (a miracle that the one who needs them ever managed to manifest consistently enough to get the medication prescribed). In his final monster incarnation he becomes bigger, more muscled, able to climb walls like a fly, and invulnerable to being shot in the chest at close range (we really don't want patients believing in that one, do we?). Remarkable in the circumstances that his one female alter didn't develop a bulging bosom.

Anyway, you may have noticed that Hollywood isn't the only context in which zombies have been used for other purposes and dubious stories about personal identity told. In philosophy our problems with traditional agency and responsibility have led to widespread acceptance of attenuated forms of personhood; multiple draft people, various self-referential illusions, and epiphenomenal confabulations. These sceptical views of common-sense selfhood are often discussed in a relatively positive light, as yielding a kind of Buddhist insight, or bringing a welcome relief from moral liability; but I don't think it's too fanciful to fear that they might also create a climate that fosters a sense of powerlessness and depersonalisation. I'd be the last person to say that philosophers should self-censor, still less that they should avoid hypotheses that look true or interesting but are depressing. Nor am I suffering from the delusion that the public at large, or even academic psychologists, are waiting eagerly to hear what the philosophers think. But perhaps there's room for slightly more awareness that these are not purely academic issues?

Is The Brain Understandable?

13 August 2017



Can we, one day, understand how the neurology of the brain leads to conscious minds, or will that remain impossible?

Round here we mostly discuss the mind from a top-down, philosophical perspective; but there is another way, which is to begin by understanding the nuts and bolts and then gradually working up to more complex processes. This Scientific American piece gives a quick view of how research at the neuronal level is coming along (quite well, but with vastly more to do).

Is this ever going to tell us about consciousness, though? A point often quoted by pessimists is that we have had the complete 'wiring diagram' of the roundworm Caenorhabditis elegans for years (Caenorhabditis has only just over 300 neurons and they have all been mapped) but still cannot properly explain how it works. Apparently researchers have largely given up on this puzzle for now. Perhaps Caenorhabditis is just too simple; its nervous system might be quirky or use elegant but opaque tricks that make it particularly difficult to fathom. Instead researchers are using fruit fly larvae and other creatures with nervous systems that are simple enough to deal with, but large enough to suggest that they probably work in a generic way, one that is broadly standard for all nervous systems up to and including the human. With modern research techniques this kind of approach is yielding some actual progress.

How optimistic can we be, though? We can never understand the brain by knowing the simultaneous states of all its neurons, so the hope of eventual understanding rests on the neurology of the brain being legible at some level. We hope there will turn out to be functions that get repeated, that firm building blocks of some intelligible structure; that we will be able to deduce rules or a kind if grammar which will let us see how things work on a slightly higher level of description.

This kind of structure is built into machines and programs; they are designed to be legible by human beings and lend themselves to reverse engineering. But the brain was not designed and is under no obligation to construct itself according to regular plans and principles. Our hope that it won't turn out to be a permanently incomprehensible tangle rests on several possibilities.

First, it might just turn out to be like that. The computer metaphor encourages us to think that the brain must encode its information in regular ways (though the lack of anything strongly analogous to software is arguably a fly in the ointment). Perhaps we'll just get lucky. When the structure of DNA was discovered, it really seemed as if we'd had a stroke of luck of this kind. What amounted to a long string of four repeated characters, ones that given certain conditions could be read as coding for many different proteins; it looked like we had a really clear legible system of very general significance. It still does to a degree, but my impression is that the glad confident morning is over, and now the more we learn about genetics the more complex and messy it gets. But even if we take it that genetics is a perfect example of legibility, there's no particular reason to think that the connectome will be as tractable as the genome.

The second reason to be cheerful is that legibility might flow naturally from function. That is, after all, pretty much what happens with organs other than the brain. The heart is not

mysterious, because it has a clear function and its structure is very legible in engineering terms in the light of that function. The brain is a good deal more complex than that, but on the other hand we already know of neurons and groups of neurons that do intelligibly carry out functions in our sensory or muscular systems.

There are big problems when it comes to the higher cognitive functions though. First, we don't already understand consciousness the way we already understand pumps and levers. When it comes to the behaviour of fruit fly larvae, even, we can relate inputs and outputs to neural activity in a sensible way. For conscious thought it may be difficult to tell which neurons are doing it without already knowing what it is they're doing. It helps a lot that people can tell us about conscious experience, though when it comes to subjective, qualities experience we have to remember that Zombie Twin tells us about his experiences too, though he doesn't have any. (Then again, since he's the perfect counterpart of a non-zombie, how much does it matter?)

Second, conscious processing is clearly non-generic in a way that nothing else in our bodies appears to be. Muscle fibres contract, and one does it much like another. Our lungs oxygenate our blood, and there's no important difference between bronchi. Even our gut behaves pretty generically; it copes magnificently with a bizarre variety of inputs, but it reduces them all to the same array of nutrients and waste.

The conscious mind is not like that. It does not secrete litres of undifferentiated thought, producing much the same stuff every day and whatever we feed it with. On the contrary, its products are minutely specific - and that is the whole point. The chances of our being able to identify a standard thought module, the way we can identify standard functions elsewhere, are correspondingly slight as a result.

Still, one last reason to be cheerful; one thing the human brain is exceptionally good at intuiting patterns from observations; far better than it has any right to be. It's not for nothing that 'seeing' is literally the verb for vision and metaphorically the verb for understanding. So exhibiting patterns of neural activity might just be the way to trigger that unexpected insight that opens the problem out.

Issues

20 August 2017

Ned Block has produced a meaty discussion for *The Encyclopedia of Cognitive Science on Philosophical Issues About Consciousness*.

There are special difficulties about writing an encyclopedia about these topics because of the lack of consensus. There is substantial disagreement, not only about the answers, but about what the questions are, and even about how to frame and approach the subject of consciousness at all. It is still possible to soldier on responsibly, like the heroic Stanford Encyclopedia of Philosophy, doing your level best to be comprehensive and balanced. Authors may find themselves describing and critiquing many complex points of view that neither they nor the reader can take seriously for a moment; sometimes possible points of view (relying on fine and esoteric distinctions of a subtlety difficult even for professionals to grasp), that in point of fact no-one, living or dead, has ever espoused. This can get tedious. The other approach, in my mind, is epitomised by the Oxford Companion to the Mind, edited by Richard Gregory, whose policy seemed to be to gather as much interesting stuff as possible and worry about how it hung together later, if at all. If you tried to use the resulting volume as a work of reference you would usually come up with nothing or with a quirky, stimulating take instead of the mainstream summary you really wanted; however, it was a cracking read, full of fascinating passages and endlessly browsable.

Luckily for us, Block's piece seems to lean towards the second approach; he is mainly telling us what he thinks is true, rather than recounting everything anyone has said, or might have said. You might think, therefore, that he would start off with the useful and much-quoted distinction he himself introduced into the subject: between phenomenal, or p-consciousness, and access, or a-consciousness. Here instead he proposes two basic forms of consciousness: phenomenality and reflexivity. Phenomenality, the feel or subjective aspect of consciousness, is evidently fundamental; reflexivity is reflection on phenomenal experience. While the first seems to be possible without the second - we can have subjective experience without thinking about it, as we might suppose dogs or other animals do - reflexivity seems on this account to require phenomenality. It doesn't seem that we could have a conscious creature with no sensory apparatus, that simply sits quietly and - what? Invents set theory, perhaps, or metaphysics (why not?).

Anyway, the Hard Problem according to Block is how to explain a conscious state (especially phenomenality) in terms of neurology. In fact, he says, no-one has offered even a highly speculative answer, and there is some reason to think no satisfactory answer can be given. He thinks there are broadly four naturalistic ways you can go: eliminativism; philosophical reductionism (or deflationism); phenomenal realism (or inflationism); or dualistic naturalism. The third option is the one Block favours.

He describes inflationism as the belief that consciousness cannot be philosophically reduced. So while a deflationist expects to reduce consciousness to a redundant term with no distinct and useful meaning, an inflationist thinks the concept can't be done away with. However, an

inflationist may well believe that scientific reduction of consciousness is possible. So, for example, science has reduced heat to molecular kinetic energy; but this is an empirical matter; the concept of heat is not abolished. (I'm a bit uncomfortable with this example but you see what he's getting at). Inflationists might also, like McGinn, think that although empirical reduction is possible, it's beyond our mental capacities; or they might think it's altogether impossible, like Searle (is that right or does he think we just haven't got the reduction yet?).

Block mentions some leading deflationist views such as higher-order theories and representationism, but inflationists will think that all such theories leave out the thing itself, actual phenomenal experience. How would an empirical reduction help? So what if experience Q is neural state X? We're not looking for an explanation of that identity - there are no explanations of identities - but rather an explanation of how something like Q could be something like X, an explanation that removes the sense of puzzlement. And there, were back at square one; nobody has any idea.

So what do we do? Block thinks there is a way forward if we distinguish carefully between a property and the concept of a property. Different concepts can identify the same property, and this provides a neat analysis of the classic thought experiment of Mary the colour scientist. Mary knows everything science could ever tell her about colour; when she sees red for the first time does she know a new fact - what red is like? No; on this analysis she gains a new concept of a property she was already familiar with through other, scientific concepts. Thus we can exchange a dualism of properties for a dualism of concepts. That may be less troubling, but I'm not sure is altogether trouble-free; for one thing it requires phenomenal concepts which seem themselves to need some demystifying explanation. In general though, I like what I take to be Block's overall outlook; that reductions can be too greedy and that the world actually retains a certain unavoidable conceptual, perhaps ontological, complexity.

Moving off on a different tack, he notes recent successes in identifying neural correlates of experience. There is a problem, however; while we can say that a certain experience corresponds with a certain pattern of neuronal activity, that pattern (so far as we can tell) can recur without the conscious experience. What's the missing ingredient? As a matter of fact I think it could be almost anything, given the limited knowledge we have of neurological detail: however, Block sees two families of possible explanation. Maybe it's something like intensity or synchrony; or maybe it's access (aha!); the way the activity is connected up with other bits of brain that do memory or decision-making; let's say with the global mental workspace, without necessarily committing to that being a distinct thing.

But these types of explanation embody different theoretical approaches; physicalism and functionalism respectively. The danger is that these may be theories of different kinds of consciousness. Physicalism may be after phenomenal consciousness, the inward experience, whereas functionalism has access consciousness, the sort that is about regulating behaviour, in its sights. It might therefore be that researchers are sometimes talking past each other.

Of course, either kind of explanation implies that there's something more going on than neural correlates, so in a sense the real drama is offstage. My instincts tell me that Block is doing things backwards; he should have started with access consciousness and worked towards the phenomenal. But as I say it is a meaty entry for an encyclopaedia, one I haven't really done justice to; see what you make of it.

Three Laws of Robotics

27 August 2017

Do Asimov's Three Laws even work? Ben Goertzel and Louie Helm, who both know a bit about AI, think not.

The three laws, which play a key part in many robot-based short stories by Asimov, and a somewhat lesser background role in some full-length novels, are as follows. They have a strict order of priority.

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

Consulted by George Dvorsky, both Goertzel and Helm think that while robots may quickly attain the sort of humanoid mental capacity of Asimov's robots, they won't stay at that level for long. Instead they will cruise on to levels of super intelligence which make law-like morals imposed by humans irrelevant.

It's not completely clear to me why such moral laws would become irrelevant. It might be that Goertzel and Helm simply think the superbots will be too powerful to take any notice of human rules. It could be that they think the AIs will understand morality far better than we do, so that no rules we specify could ever be relevant.

I don't think, at any rate, that it's the case that super intelligent bots capable of human-style cognition would be morally different to us. They can go on growing in capacity and speed, but neither of those qualities is ethically significant. What matters is whether you are a moral object and/or a moral subject. Can you be hurt, on the one hand, and are you an autonomous agent on the other? Both of these are yes/no issues, not scales we can ascend indefinitely. You may be more sensitive to pain, you may be more vulnerable to other kinds of harm, but in the end you either are or are not the kind of entity whose interests a moral person must take into account. You may make quicker decisions, you may be massively better informed, but in the end either you can make fully autonomous choices or you can't. (To digress for a moment, this is business of truly autonomous agency is clearly a close cousin at least of our old friend Free Will; compatibilists like me are much more comfortable with the whole subject than hard-line determinists. For us, it's just a matter of defining free agency in non-magic terms. I, for example, would say that free decisions are those determined by thoughts about future or imagined contingencies (more cans of worms there, I know). How do hard determinists working on AGI manage? How can you try to endow a bot with real agency when you don't actually believe in agency anyway?)

Nor do I think rules are an example of a primitive approach to morality. Helm says that rules are pretty much known to be a 'broken foundation for ethics', pursued only by religious philosophers that others laugh and point at. It's fair to say that no-one much supposes a list like the Ten Commandments could constitute the whole of morality, but rules surely have a role to play. In my view (I resolved ethics completely in this post a while ago, but nobody seems to have noticed yet.) the central principle of ethics is a sort of 'empty consequentialism' where we studiously avoid saying what it is we want to maximise (the greatest whatever of the greatest number); but that has to be translated into rules because of the impossibility of correctly assessing the infinite consequences of every action; and I think many other general ethical principles would require a similar translation. It could be that Helm supposes super intelligent AIs will effortlessly compute the full consequences of their actions: I doubt that's possible in principle, and though computers may improve, to date this has been the sort of task they are really bad at; in the shape of the wider Frame Problem, working out the relevant consequences of an action has been a major stumbling block to AI performance in real world environments.

Of course, none of that is to say that Asimov's Laws work. Helm criticises them for being 'adversarial', which I don't really understand. Goertzel and Helm both make the fair point that it is the failure of the laws that generally provides the plot for the short stories; but it's a bit more complicated than that. Asimov was rebelling against the endless reiteration of the stale 'robots try to take over' plot, and succeeded in making the psychology and morality of robots interesting, dealing with some issues of real ethical interest, such as the difference between action and inaction (if the requirement about inaction in the First Law is removed, he points out that robots would be able to rationalise killing people in various ways. A robot might drop a heavy weight above the head of a human. Because it knows it has time to catch the weight, doing so is not murder in itself, but once the weight is falling, since inaction is allowed, the robot need not in fact catch the thing.

Although something always had to go wrong to generate a story, the Laws were not designed to fail, but were meant to embody genuine moral imperatives.

Nevertheless, there are some obvious problems. In the first place, applying the laws requires an excellent understanding of human beings and what is or isn't in their best interests. A robot that understood that much would arguably be above control by simple laws, always able to reason its way out.

There's no provision for prioritisation or definition of a sphere of interest, so in principle the First Law just overwhelms everything else. It's not just that the robot would force you to exercise and eat healthily (assuming it understood human well-being reasonably well; any errors or over-literal readings - 'humans should eat as many vegetables as possible' - could have awful consequences); it would probably ignore you and head off to save lives in the nearest famine/war zone. And you know, sometimes we might need a robot to harm human beings, to prevent worse things happening.

I don't know what ethical rules would work for super bots; probably the same ones that go for human beings, whatever you think they are. Goertzel and Helm also think it's too soon to say; and perhaps there is no completely safe system. In the meantime, I reckon practical laws might be more like the following.

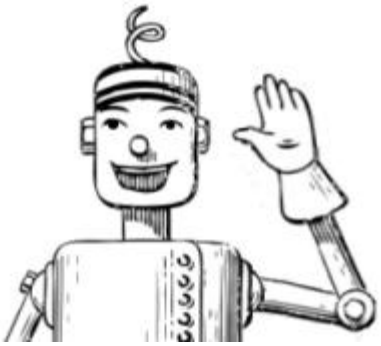
1. Leave Rest State and execute Plan, monitoring regularly.

2. If anomalies appear, especially human beings in unexpected locations, sound alarm and try to return to Rest State.
3. If returning to Rest State generates new anomalies, stop moving and power down all tools and equipment.

Can you do better than that?

Your Plastic Pal

3 September 2017



Scott Bakker has a thoughtful piece which suggests we should be much more worried than we currently are about AIs that pass themselves off, superficially, as people. Of course this is a growing trend, with digital personal assistants like Alexa or Cortana, that interact with users through spoken exchanges, enjoying a surge of popularity. In fact, it has just been announced that those two are going to benefit from a degree of integration. That might raise the question of whether in future they will really be two entities or one with two names - although in one sense the question is nugatory. When we're dealing with AIs we're not dealing with any persons at all; but one AI can easily present as any number of different simulated personal entities.

Some may feel I assume too much in saying so definitely that AIs are not persons. There is, of course, a massive debate about whether human consciousness can in principle be replicated by AI. But here we're not dealing with that question, but with machines that do not attempt actual thought or consciousness and were never intended to; they only seek to interact in ways that seem human. In spite of that, we're often very ready to treat them as if they were human. For Scott this is a natural if not inevitable consequence of the cognitive limitations that in his view condition or even generate the constrained human view of the world; however, you don't have to go all the way with him in order to agree that evolution has certainly left us with a strong bias towards crediting things with agency and personhood.

Am I overplaying it? Nobody really supposes digital assistants are really people, do they? If they sometimes choose to treat them as if they were, it's really no more than a pleasant joke, surely, a bit of a game?

Well, it does get a little more serious. James Vlahos has created a chat-bot version of his dying father, something I wouldn't be completely comfortable with myself. In spite of his enthusiasm for the project, I do think that Vlahos is, ultimately, aware of its limitations. He knows he hasn't captured his father's soul or given him eternal digital life in any but the most metaphorical sense. He understands that what he's created is more like a database accessed with conversational cues. But what if some appalling hacker made off with a copy of the dadbot, and set it to chatting up wealthy widows with its convincing life story, repertoire of anecdotes and charming phrases? Is there a chance they'd be taken in? I think they might be, and these things are only going to get better and more convincing.

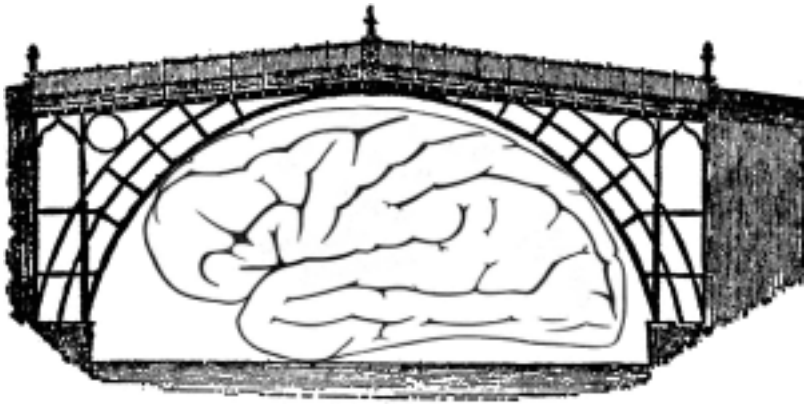
Then again, if we set aside that kind of fraud (perhaps we'll pick up that suggestion of a law requiring bots to identify themselves), what harm is there in spending time talking to a bot? It's no more of a waste of time than some trivial game, and might even be therapeutic for some. Scott says that deprivation of real human contact can lead to psychosis or depression, and that talking to bots might degrade your ability to interact with people in real life; he foresees a generation of hikikomori, young men unable to deal with real social interactions, let alone real girlfriends.

Something like that seems possible, though it may be hard to tell whether excessive bot use would be cause, symptom, palliation, or all three. On the one hand we might make fools of ourselves, leaving the computer on all night in case switching it off kills our digital friend, or trying to give legal rights to non-existent digital people. Someone will certainly try to marry one, if they haven't already. More seriously, getting used to robot pals might at least make us ruder and more impatient with human service providers, more manipulative and less respectful in our attitudes to crime and punishment, and less able to understand why real people don't laugh at our jokes and echo back our opinions (is that… is that happening already?)

I don't know what can be done about it; if Scott is anywhere near right, then these issues are too deeply rooted in human nature for us to change direction. Maybe in twenty years, these words, if not carried away by digital rot, will seem impossibly quaint and retrograde; readers will wonder what can have been wrong with my hidden layers.

Bridging The Brain

10 September 2017



How can we find out how the brain works? This was one of five questions put to speakers at the Cognitive Computational Neuroscience conference, and the answers, posted on the conference blog, are interesting. There seems to be a generally shared perspective that what we need are bridges between levels of interpretation, though there are different takes on what those are likely to be and how we get them. There's less agreement about the importance of recent advances in machine learning and how we respond to them.

Rebecca Saxe says the biggest challenge is finding the bridge between different levels of interpretation - connecting neuronal activity on the one hand with thoughts and behaviour on the other. She thinks real progress will come when both levels have been described mathematically. That seems a reasonable aspiration for neurons, though the maths would surely be complex; but the mind boggles rather at the idea of expressing thoughts mathematically. It has been suggested in the past that formal logic was going to be at least an important part of this, but that hasn't really gone well in the AI field and to me it seems quite possible that the unmanageable ambiguity of meanings puts them beyond mathematical analysis (although it could be that in my naivety I'm under-rating the subtlety of advanced mathematical techniques).

Odelia Schwartz looks to computational frameworks to bring together the different levels; I suppose the idea is that such frameworks might themselves have multiple interpretations, one resembling neural activity while another is on a behavioural level. She is optimistic that advances in machine learning open the way to dealing with natural environments: that is a bit optimistic in my view but perhaps not unreasonably so.

Nicole Rust advocates 'thoughtful descriptions of the computations that the brain solves'. We got used, she rightly says, to the idea that the test of understanding was being able to build the machine. The answer to the problem of consciousness would not be a proof, but a robot. However, she points out, we're now having to deal with the idea that we might build successful machines that we don't understand.

Another way of bridging between those levels is proposed by Birte Forstmann: formal models that make simultaneous predictions about different modalities such as behaviour and the brain. It sounds good, but how do we pull it off?

Alona Fyshe sees three levels - neuronal, macro, and behavioural - and wants to bring them together through experimental research, crucially including real world situations: you can learn something from studying subjects reading a sentence, but you're not getting the full picture unless you look at real conversations. It's a practical programme, but you have to accept that the correlations you observe might turn out complex and deceptive; or just incomprehensible.

Tom Griffiths has a slightly different set of levels, derived from David Marr; computational, algorithmic, and implementation. He feels the algorithmic level has been neglected; but I'd say it remains debatable whether the brain really has an algorithmic level. An algorithm implies a tractable level of complexity, whereas it could be that the brain's 'algorithms' are so complex that all explanatory power drains away. Unlike a computer, the brain is under no obligation to be human-legible.

Yoshua Bengio hopes that there is, at any rate, a compact set of computational principles in play. He advocates a continuing conversation between those doing deep learning and other forms of research.

Wei Ji Ma expresses some doubt about the value of big data; he favours a diversity of old-fashioned small-scale, hypothesis-driven research; a search for evolutionarily meaningful principles. He's a little concerned about the prevalence of research based on rodents; we're really interested in the unique features of human cognition, and rats can't tell us about those.

Michael Shadlen is another sceptic about big data and a friend of hypothesis driven research, working back from behaviour; he's less concerned with the brain as an abstract computational entity and more with its actual biological nature. People sometimes say that AI might achieve consciousness by non-biological means, just as we achieved flight without flapping wings; Shadlen, on that analogy, still wants to know how birds fly.

If this is a snapshot of the state of the field, I think it's encouraging; the outlooks briefly indicated here seem to me to show good insight and promising outlooks. But it is possible to fear that we need something more radical. Perhaps we'll only find out how the brain works when we ask the right questions, ones that realign our view in such a way that different levels of interpretation no longer seem to be the issue. Our current views are dominated by the concept of consciousness, but we know that in many ways that is a recent idea, primarily Western and perhaps even Anglo-Saxon. It might be that we need a paradigm shift; but alas I have no idea where that might come from.

Inconceivable Arguments

17 September 2017

Stephen Law has new arguments against physicalism (which is, approximately, the view that the account of the world given by physics is good enough to explain everything). He thinks conscious experience can't be dealt with by physics alone. There is a well-established family of anti-physicalist arguments supporting this view which are based on conceivability; Law adds new cousins to that family, ones which draw on inconceivability and are, he thinks, less vulnerable to some of the counter-arguments brought against the established versions.

What is the argument from conceivability? Law helpfully summarises several versions (his exposition is commendably clear and careful throughout) including the zombie twin argument we've often discussed here; but let's take a classic one about the supposed identity of pain and the firing of C-fibres, a kind of nerve. It goes like this...

1. Pain without C-fibre firing is conceivable
2. Conceivability entails metaphysical possibility (at least in this case)
3. The metaphysical possibility of pain without C-fibre firing entails ?that pain is not identical with C-fibre firing.
4. Pain is not identical with C-fibre firing

(It's good to see the old traditional C-fibre example still being quoted. It reminds me of long-gone undergraduate days when some luckless fellow student in a tutorial read an essay citing the example of how things look yellow to people with jaundice. Myles Burnyeat, the tutor, gave an erudite history of this jaundice example, tracking it back from the twentieth century to the sixteenth, through mediaeval scholastics, Sextus Empiricus (probably), and many ancient authors. You have joined, he remarked, a long tradition of philosophers who have quoted that example for thousands of years without any of them ever bothering to find out that it is, in fact, completely false. People with jaundice have yellowish skin, but their vision is normal. Of course the inaccuracy of examples about C-fibres and jaundice does not in itself invalidate the philosophical arguments, a point Burnyeat would have conceded with gritted teeth.)

I have a bit of a problem with the notion of metaphysical possibility. Law says that something being conceivable means no incoherence arises when we suppose it, which is fine; but I take it that the different flavours of conceivability/possibility arise from different sets of rules. So something is physically conceivable so long as it doesn't contradict the laws of physics. A five-kilometre cube of titanium at the North Pole is not something that any plausible set of circumstances is going to give rise to, but nothing about it conflicts with physics, so it's conceivable.

I'm comfortable, therefore, with physical conceivability, and with logical conceivability, because pretty decent (if not quite perfect) sets of rules for both fields have been set out for us. But what are the laws of metaphysics that would ground the idea of metaphysical conceivability or equally, metaphysical possibility? I'm not sure how many candidates for such laws (other than ones that are already laws of physics, logic, or maths) I can come up with, and I know of no

attempt to set them out systematically (a book opportunity for a bold metaphysician there, perhaps). But this is not a show-stopper so long as it is reasonably clear in each case what kind of metaphysical rules we hold ourselves not to be violating.

Conceivability arguments of this kind do help clarify an intuitive feeling that physical events are just not the sort of thing that could also be subjective experiences, firming things up for those who believe in them and sportingly providing a proper target for physicalists.

So what is the new argument? Law begins by noting that in some cases appearance and reality can separate, while in others they cannot. So a substance that appears just like gold, but does not have the right atomic number, is not gold: we could call it fool's gold. A medical case where the skin was yellowish but the underlying condition was not jaundice might be fool's jaundice (jaundice again, but here used unimpeachably). However, can there be fool's red? If we're talking of a red experience it seems not: something that seems red is indeed a red experience whatever underlies it. More strongly still, it seems that the idea of fool's pain is inconceivable. If what you're experiencing seems to be pain, then it is pain.

Is that right? There's evidently something in it, but what is to stop us believing ourselves to be in pain when we're not? Hypochondriacs may well do that very thing. Law, I suppose, would say that a mistaken belief isn't enough; there has to be the actual experience of pain. That begins to look as if he's in danger of begging the question; if we specify that there's a real experience of pain, then it's inconceivable it isn't real pain? But I think the notions of mistaken pain beliefs and the putative fool's pain are sufficiently distinct.

The inconceivability argument goes on to suggest that if fool's pain is inconceivable, but we can conceive of C-fibre firing without pain, then C-fibre firing cannot be identical with pain. Clearly the same argument would apply for various mental experiences other than pain, and for any proposed physical correlate of pain.

Law rebuts explicitly arguments that this is nothing new, or merely a trivial variant on the old argument. I'm happy enough to take it as a handy new argument, worth having in itself; but Law also argues that in some cases it stands up against counter-arguments better than the old one. Notably he mentions an argument by Loar. This offers an alternative explanation for the conceivability of pain in the absence of C-fibre firing: experience and concepts such as C-fibre firing are dealt with in quite different parts of the brain, and our ability to conceive of one without the other is therefore just a matter if human psychology, from which no deep metaphysical conclusions can be drawn. Law argues that even if Loar's argument or a similar one is accepted, we still run up against the problem that conceiving of pain without C-fibre firing, we are conceiving of fool's pain, which the new argument has established in inconceivable.

The case is well made and I think Law is right to claim he has hit on a new and useful argument. Am I convinced? Not really; but my disbelief stems from a more fundamental doubt about whether conceivability and inconceivability can actually tell us about the real world, or merely about our own mental powers.

Perhaps we can look at it in terms of possible worlds. It seems to me that Law's argument, like the older ones, establishes that we can separate C-fibre firing and pain conceptually; that there are in fact possible worlds in which pain is not C-fibre firing. But I don't care. I don't require the identity of pain and C-fibre firing to be true a priori; I'm happy for it to be true only in this world, as an empirical, scientific matter. Of course this opens a whole range of new cans of worms

(about which kinds of identity are necessary, for example) whose contents I am not eager to ingest at the moment.

Blue Topology

24 September 2017

An interesting but somewhat problematic paper from the Blue Brain project claims that the application of topology to neural models has provided a missing link between structure and function. That's exciting because that kind of missing link is just what we need to enable us to understand how the brain works. The claim about the link is right there in the title, but unfortunately so far as I can see the paper itself really only attempts something more modest. It only seems to offer a new exploration of some ground where future research might conceivably put one end of the missing link. There also seem to me to be some problems in the way we're expected to interpret some of the findings reported.

That may sound pretty negative. I should perhaps declare in advance that I know little neurology and less topology, so my opinion is not worth much. I also have form as a Blue Brain sceptic, so you can argue that I bring some stored-up negativity to anything associated with it. I've argued in the past that the basic goal of the project, of simulating a complete human brain is misconceived and wildly over-ambitious; not just a waste of money but possibly also a distraction which might suck resources and talent away from more promising avenues.

One of the best replies to that kind of scepticism is to say, well look; even if we don't deliver the full brain simulation, the attempt will energise and focus our research in a way which will yield new and improved understanding. We'll get a lot of good research out of it even if the goal turns out to be unattainable. The current paper, which demonstrates new mathematical techniques, might well be a good example of that kind of incidental pay off. There's a nice explanation of the paper here, with links to some other coverage, though I think the original text is pretty well written and accessible.

As I understand it, topological approaches to neurology in the past have typically considered neural networks as static objects. The new approach taken here adds the notion of directionality, as though each connection were a one-way street. This is more realistic for neurons. We can have groups of neurons where all are connected to all, but only one neuron provides a way into the group and one provides a way out; these are directed simplices. These simplices can be connected to others at their edges where, say, two of the member neurons are also members of a neighbouring simplex. Where there are a series of connected simplices, they may surround a void where nothing is going on. These cavities provide a higher level of structure, but I confess I'm not altogether clear as to why they are interesting. Holes, of course, are dear to the heart of any topologist, but in terms of function I'm not really clear about their relevance.

Anyway, there's a lot in the paper but two things seem especially noteworthy. First, the researchers observed many more simplices, of much higher dimensionality, than could be expected from a random structure (they tested several such random structures put together according to different principles). 'Dimensionality' here just refers to how many neurons are involved; a simplex of higher dimensionality contains more neurons. Second, they observed a

characteristic pattern when the network was presented with a ‘stimulus’; simplices of gradually higher and higher dimensionality would appear and then finally disperse. This is not, I take it, a matter of the neurons actually wiring up new connections on the fly, it’s simply about which are involved actively by connections that are actually firing.

That’s interesting, but all of this so far was discovered in the Blue Brain simulated neurons, more or less those same tiny crumbs of computationally simulated rat cortex that were announced a couple of years ago. It is, of course, not safe to assume that real brain behaves in the same way; if we rely entirely on the simulation we could easily be chasing our tails. We would build the simulation to match our assumptions about the brain and then use the behaviour of the simulation to validate the same assumptions. In fact the researchers very properly tried to perform similar experiments with real rat cortex. This requires recording activity in a number of adjacent neurons, which is fantastically difficult to pull off, but to their credit they had some success; in fact the paper claims they confirmed the findings from the simulation. The problem is that while the simulated cortex was showing simplices of six or seven dimensions (even higher numbers are quoted in some of the media reports, up to eleven), the real rat cortex only managed three, with one case of four. Some of the discussion around this talks as though a result of three is partial confirmation of a result of six, but of course it isn’t. Putting it brutally, the team’s own results in real cortex contradicted what they had found in the simulation. Now, there could well be good reasons for that; notably they could only work with a tiny amount of real cortex. If you’re working with a dozen neurons at a time, there’s obviously quite a low ceiling on the complexity you can expect. But the results you got are the results you got, and I don’t see that there’s a good basis here for claiming that the finding of high-order simplices is supported in real brains. In fact what we have if anything is prima facie evidence that there’s something not quite right about the simulation. The researchers actually took a further step here by producing a simulation of the actual real neurons that they tested and then re-running the tests. Curiously, the simulated versions in these cases produced fewer simplices than the real neurons. The paper interprets this as supportive of its conclusions; if the real cortex was more productive of simplices, it argues, then we might expect big slabs of real brain to have even more simplices of even higher dimensionality than the remarkable results we got with the main simulation. I don’t think that kind of extrapolation is admissible; what you really got was another result showing that your simulations do not behave like the real thing. In fact, if a simulation of only twelve neurons behaves differently from the real thing in significant respects, that surely must indicate that the simulation isn’t reproducing the real thing very well?

The researchers also looked at the celebrated roundworm C. Elegans, the only organism whose neural map (or connectome) is known in full, and apparently found evidence of high-order simplices - though I think it can only have been a potential for such simplices, since they don’t seem to have performed real or simulated experiments, merely analysing the connectome.

Putting all that aside, and supposing we accept the paper’s own interpretations, the next natural question is: so what? It’s interesting that neurons group and fire in this way, but what does that tell us about how the brain actually functions? There’s a suggestion that the pattern of moving up to higher order simplices represents processing of a sensory input, but in what way? In functional terms, we’d like the processing of a stimulus to lead on to action, or perhaps to the recording of a memory trace, but here we just seem to see some neurons get excited and then stop being excited. Looking at it in simple terms, simplices seem really bad candidates for any functional role, because in the end all they do is deliver the same output signal as a single neural connection would do. Couldn’t you look at the whole thing with a sceptical eye and say

that all the researchers have found is that a persistent signal through a large group of neurons gradually finds an increasing number of parallel paths?

At the end of the paper we get some speculation that addresses this functional question directly. The suggestion is that active high-dimensional simplices might be representing features of the stimulus, while the grouping around cavities binds together different features to represent the whole thing. It is, if sketchy, a tenable speculation, but quite how this would amount to representation remains unclear. There are probably other interesting ways you might try to build mental functions on the basis of escalating simplices, and there could be more to come in that direction. For now though, it may give us interesting techniques, but I don't think the paper really delivers on its promise of a link with function.

Deus In Machina

1 October 2017

Anthony Levandowski has set up an organisation dedicated to the worship of an AI God. Or so it seems; there are few details. The aim of the new body is to 'develop and promote the realization of a Godhead based on Artificial Intelligence', and 'through understanding and worship of the Godhead, contribute to the betterment of society'. Levandowski is a pioneer in the field of self-driving vehicles (centrally involved in a current dispute between Uber and Google), so he undoubtedly knows a bit about autonomous machines.

This recalls the Asimov story where they build Multivac, the most powerful computer imaginable, and ask it whether there is a God? There is now, it replies. Of course the Singularity, mind uploading, and other speculative ideas of AI gurus have often been likened to some of the basic concepts of religion; so perhaps Levandowski is just putting down a marker to ensure his participation in the next big thing.

Yuval Noah Harari says we should, indeed, be looking to Silicon Valley for new religions. He makes some good points about the way technology has affected religion, replacing the concern with good harvests which was once at least as prominent as the task of gaining a heavenly afterlife. But I think there's an interesting question about the difference between, as it were, steampunk and cyberpunk. Nineteenth century technology did not produce new gods, and surely helped make atheism acceptable for the first time; lately, while on the whole secularism may be advancing we also seem to have a growth of superstitious or pseudo-religious thinking. I think it might be because nineteenth century technology was so legible; you could see for yourself that there was no mystery about steam locomotives, and it made it easy to imagine a non-mysterious world. Computers now, are much more inscrutable and most of the people who use them do not have much intuitive idea of how they work. That might foster a state of mind which is more tolerant of mysterious forces.

To me it's a little surprising, though it probably should not be, that highly intelligent people seem especially prone to belief in some slightly bonkers ideas about computers. But let's not quibble over the impossibility of a super-intelligent and virtually omnipotent AI. I think the question is, why would you worship it? I can think of various potential reasons.

1. Humans just have an innate tendency to worship things, or a kind of spiritual hunger, and anything powerful naturally becomes an object of worship.
2. We might get extra help and benefits if we ask for them through prayer.

3. If we don't keep on the right side of this thing, it might give us a seriously bad time (the 'Roko's Basilisk' argument).
4. By worshipping we enter into a kind of communion with this entity, and we want to be in communion with it for reasons of self-improvement and possibly so we have a better chance of getting uploaded to eternal life.

There are some overlaps there, but those are the ones that would be at the forefront of my mind. The first one is sort of fatalistic; people are going to worship things, so get used to it. Maybe we need that aspect of ourselves for mental health; maybe believing in an outer force helps give us a kind of leverage that enables an integration of our personality we couldn't otherwise achieve? I don't think that is actually the case, but even if it were, an AI seems a poor object to choose. Traditionally, worshipping something you made yourself is idolatry, a degraded form of religion. If you made the thing, you cannot sincerely consider it superior to yourself; and a machine cannot represent the great forces of nature to which we are still ultimately subject. Ah, but perhaps an AI is not something we made; maybe the AI godhead will have designed itself, or emerged? Maybe so, but if you're going for a mysterious being beyond our understanding, you might in my opinion do better with the thoroughly mysterious gods of tradition rather than something whose bounds we still know, and whose plug we can always pull.

Reasons two and three are really the positive and negative sides of an argument from advantage, and they both assume that the AI god is going to be humanish in displaying gratitude, resentment, and a desire to punish and reward. This seems unlikely to me, and in fact a projection of our own fears out onto the supposed deity. If we assume the AI god has projects, it will no doubt seek to accomplish them, but meting out tiny slaps and sweeties to individual humans is unlikely to be necessary. It has always seemed a little strange that the traditional God is so minutely bothered with us; as Voltaire put it "When His Highness sends a ship to Egypt does he trouble his head whether the rats in the vessel are at their ease or not?"; but while it can be argued that souls are of special interest to a traditional God, or that we know He's like that just through revelation, the same doesn't go for an AI god. In fact, since I think moral behaviour is ultimately rational, we might expect a super-intelligent AI to behave correctly and well without needing to be praised, worshipped, or offered sacrifices. People sometimes argue that a mad AI might seek to maximise, not the greatest good of the greatest number, but the greatest number of paperclips, using up humanity as raw material; in fact though, maximising paperclips probably requires a permanently growing economy staffed by humans who are happy and well-regulated. We may actually be living in something not that far off maximum-paperclip society.

Finally then, do we worship the AI so that we can draw closer to its godhead and make ourselves worthy to join its higher form of life? That might work for a spiritual god; in the case of AI it seems joining in with it will either be completely impossible because of the difference between neuron and silicon; or if possible, it will be a straightforward uploading/software operation which will not require any form of worship.

At the end of the day I find myself asking whether there's a covert motive here. What if you could run your big AI project with all the tax advantages of being a registered religion, just by saying it was about electronic godhead?

Replicant Identity

8 October 2017

The new Blade Runner film has generated fresh interest in the original film; over on IAI Helen Beebee considers how it nicely illustrates the concept of 'q-memories'.

This relates to the long-established philosophical issue of personal identity; what makes me me, and what makes me the same person as the one who posted last week, or the same person as that child in Bedford years ago? One answer which has been a leading contender at least since Locke is memory; my memories together constitute my identity.

Memories are certainly used as a practical way of establishing identity, whether it be in probing the claims of a supposed long-lost relative or just testing your recall of the hundreds of passwords modern life requires. It is sort of plausible that if all you memories were erased you would become new person with a fresh start; there have been cases of people who lost decades of memory and underwent personality change, identifying with their own children more readily than their now wrinkly-seeming spouses.

There are various problems with memory as a criterion of identity, though. One is the point that it seems to be circular. We can't use your memories to validate your identity because in accepting them as your memories we are already implicitly taking you to be the earlier person they come from. If they didn't come from that person they aren't validly memories. To get round this objection Shoemaker and Parfit adopted the concept of quasi- or q-memories. Q-memories are like memories but need not relate to any experience you ever had. That, of course, is too loose, allowing delusions to be used as criteria of identity, so it is further specified that q-memories must relate to an experience someone had, and must have been acquired by you in an appropriate way. The appropriate ways are ones that causally relate to the original experience in a suitable fashion, so that it's no good having q-memories that just happen to match some of King Charles's. You don't have to be King Charles, but the q-memories must somehow have got out of his head and into yours through a proper causal sequence.

This is where Blade Runner comes in, because the replicant Rachael appears to be a pretty pure case of q-memory identity. All of her memories, except the most recent ones, are someone else's; and we presume they were duly copied and implanted in a way that provides the sort of causal connection we need.

This opens up a lot of questions, some of which are flagged up by Beebee. But what about q-memories? Do they work? We might suspect that the part about an appropriate causal connection is a weak spot. What's appropriate? Don't Shoemaker and Parfit have to steer a tricky course here between the Scylla of weird results if their rules are too loose, and the Charybdis of bringing back the circularity if they are too tight? Perhaps, but I think we have to remember that they don't really want to do anything very radical with q-memories; really you could argue it's no more than a terminological specification, giving them license to talk of memories without some of the normal implications.

In a different way the case of Rachael actually exposes a weak part of many arguments about memory and identity; the easy assumption that memories are distinct items that can be copied from one mind to another. Philosophers, used to being able to specify whatever mad conditions they want for their thought-experiments, have been helping themselves to this assumption for a long time, and the advent of the computational metaphor for the mind has done nothing to discourage them. It is, however, almost certainly a false assumption.

At the back of our minds when we think like this is a model of memory as a list of well-formed propositions in some regular encoding. In fact, though, much of what we remember is implicit; you recall that zebras don't wear waistcoats though it's completely implausible that that fact was recorded anywhere in your brain explicitly. There need be nothing magic about this. Suppose we remember a picture; how many facts does the picture contain? We can instantly come up with an endless list of facts about the relations of items in the picture, but none were encoded as propositions. Does the Mona Lisa have her right hand over her left, or vice versa? You may never have thought about it but be easily able to recall which way it is. In a computer the picture might be encoded as a bitmap; in our brain we don't really know, but plausibly it might be encoded as a capacity to replay certain neural firing sequences, namely those that were caused by the original experience. If we replay the experience neurally, we can sort of have the experience again and draw new facts from it the way we could from summoning up a picture; indeed, that might be exactly what we are doing.

But my neurons are not wired up like yours, and it is vanishingly unlikely that we could identify direct equivalents of specific neurons between brains, let alone whole firing sequences. My memories are recorded in a way that is specific to my brain, and they cannot be read directly across into yours.

Of course, replicants may be quite different. It's likely enough that their brains, however they work, are standardised and perhaps use a regular encoding which engineers can easily read off. But if they work differently from human brains, then it seems to follow that they can't have the same memories; to have the same memories they would have to be an unbelievably perfect copy of the 'donor' brain.

That actually means that memories are in a way a brilliant criterion of personal identity, but only in a fairly useless sense.

However, let me briefly put a completely different argument in a radically different direction. We cannot upload memories, but we know that we can generate false ones by talking to subjects or presenting fake evidence. What does that tell us about memories? I submit it suggests that memories are in essence beliefs, beliefs about what happened in the past. Now we might object that there is typically some accompanying phenomenology. We don't just remember that we went to the mall, we remember a bit of what it looked like, and other experiential details. But I claim that our minds readily furnish that accompanying phenomenology through confabulation, given the belief, and in fact that a great deal of the phenomenological dressing of all memories, even true ones, is actually confected.

But I would further argue that the malleability of beliefs means that they are completely unsuitable as criteria of identity; it follows that memories are similarly unsuitable, so we have been on the wrong track throughout. (Regular readers may know that in fact I subscribe to a view regarded by most as intolerably crude; that human beings are physical objects like any other and have essentially the same criteria of identity.)

Disastrous Consciousness

15 October 2017

Hugh Howey gives a bold, amusing, and hopelessly optimistic account of how to construct consciousness in Wired. He thinks it wouldn't be particularly difficult. Now you might think that a man who knows how to create artificial consciousness shouldn't be writing articles; he should be building the robot mind. Surely that would make his case more powerfully than any amount of prose? But Howey thinks an artificial consciousness would be disastrous. He thinks even the natural kind is an unfortunate burden, something we have to put up with because evolution has yet to find a way of getting the benefits of certain strategies without the downsides. But he doesn't believe that conscious AI would take over the world, or threaten human survival, so I would still have thought one demonstration piece was worth the effort? Consciousness sucks, but here's an example just to prove the point?

What is the theory underlying Howey's confidence? He rests his ideas on Theory of Mind (which he thinks is little discussed); the ability to infer the thoughts and intentions of others. In essence, he thinks that was a really useful capacity for us to acquire, helping us compete in the cut-throat world of human society; but when we turn it on ourselves it disastrously generates wrong results, in particular about our own having of conscious states.

It remains a bit mysterious to me why he thinks a capacity that is so useful applied to others should be so disastrously and comprehensively wrong when applied to ourselves. He mentions priming studies, where our behaviour is actually determined by factors we're unaware of; priming's reputation has suffered rather badly recently in the crisis of non-reproducibility, but I wouldn't have thought even ardent fans of priming would claim our mental content is entirely dictated by priming effects.

Although Dennett doesn't get a mention, Howey's ideas seem very Denettian, and I think they suffer from similar difficulties. So our Theory of Mind leads us to attribute conscious thoughts and intentions to others; but what are we attributing to them? The theory tells us that neither we nor they actually have these conscious contents; all any of us has is self-attributions of conscious contents. So what, we're attributing to them some self-attributions of self-attributions of… The theory covertly assumes we already have and understand the very conscious states it is meant to analyse away. Dennett, of course, has some further things to say about this, and he's not as negative about self-attributions as Howie.

But you know, it's all pretty implausible intuitively. Suppose I take a mouthful of soft-boiled egg which tastes bad, and I spit it out. According to Howey what went on there is that I noticed myself spitting out the egg and thought to myself: hm, I infer from this behaviour that it's very probable I just experienced a bad taste, or maybe the egg was too hot, can't quite tell for sure.

The thing is, there are real conscious states irrespective of my own beliefs about them (which indeed, may be plagued by error). They are characterised by having content and intentionality, but these are things Howie does not believe in, or rather it seems has never thought of; his view that a big bank of indicator lights shows a language capacity suggests he hasn't gone into this business of meaning and language quite deeply enough.

If he had to build an artificial consciousness, he might set up a community of self-driving cars, let one make inferences about the motives of the others and then apply that capacity to itself. But it would be a stupid thing to do because it would get it wrong all the time; in fact at this point Howie seems to be tending towards a view that all Theory of Mind is fatally error-prone. It would better, he reckons, if all the cars could have access to all of each other's internal data, just as universal telepathy would be better for us (though in the human case it would be undermined by mind-masking freeloaders.

Would it, though? If the cars really had intentions, their future behaviour would not be readily deducible simply from reading off all the measurements. You really do have to construct some kind of intentional extrapolation, which is what the Dennettian intentional stance is supposed to do.

I worry just slightly that some of the things Howey says seem to veer close to saying, hey a lot of these systems are sort of aware already; which seems unhelpful. Generally, it's a vigorous and entertaining exposition, even if, in my view, on the wrong track.'

Conscious Electrons

28 October 2017

Philip Goff gives a brief but persuasive new look at his case for panpsychism (the belief that experience, or consciousness in some form, is in everything) in a recent post on the OUPblog site. In the past, he says, explanations have generally been 'brain first'. Here's this physical object, the brain - and we understand physical objects well enough - the challenge is to explain how this scrutable piece of biological tissue on the one hand gives rise to this evanescent miracle, consciousness, on the other. That way of looking at it, suggests Goff, turns out to be the wrong way round. We don't really understand the real nature of matter at all: what we understand is that supposedly mysterious consciousness. So what we ought to do is start there and work towards a better understanding of matter.

This undoubtedly appeals to a frustration many philosophers must have felt. People at large tend to take it for granted that what we really know about is the physical external world around us, described in no-nonsense terms (with real equations!) by science. Phenomenology and all that stuff about what we perceive is an airy-fairy add-on. In fact, of course, it's rather the other way round. The only thing we know directly, and so, perhaps, with certainty, is our own experience; the external world and the theories of science all finally rest on that first-person foundation. Science is about observation and observation is ultimately a matter of sensory experience.

Goff notes that physics gives us no account of the intrinsic nature of matter, only its observable and causal properties. We know things, as it were, only from the outside. But in the case of our own experience, uniquely, we know it from the inside, and have direct acquaintance with its essential nature. When we experience redness we experience it unmasked; in physics it hides behind a confusing array of wavelengths, reflectances, and other highly abstract and ambiguous concepts, divorced from experience by many layers of reasoning. Is there not an argument for the hypothesis that the intrinsic nature of matter is the same as the intrinsic nature of the only thing whose intrinsic nature we know -our own experience? Perhaps after all we should consider supposing that even electrons have some tiny spark of awareness.

In fact Goff sees two arguments. One is that there simply seems no other reasonable way of accounting for consciousness. We can't see where it could have come from, so let's assume it has always been everywhere. Goff doesn't like this case and thinks it is particularly prone to the compositional difficulties often urged against panpsychism; how do these micro-consciousness stack up in larger entities, and how in particular do they relate to the kind of consciousness we seem to have in our brain? Goff prefers to rest on simplicity; panpsychism is just the most parsimonious explanation. Instead of having two, or multiple kinds of intrinsic natures, we assume that there's just one. He realises that some may see this as a weak argument far short of proof, but parsimony is a strong and legitimate criterion for judging between theories; indeed, it's indispensable.

Now I'm on record as suggesting that things out there have one property that falls outside all physical theories - namely reality. Am I not tempted to throw in my lot with Goff and suggest that

as a further simplification we could say that reality just consists in having an intrinsic nature, ie having experience? Not really.

Let's go back a bit. Do we really understand our conscious experience? We have to remember that consciousness seems to have two faces. To use Ned Block's terms, there is access or a-consciousness; the sort that is involved in governing our behaviour, making decisions, deciding what to say, and other relatively effable processes. Then there is phenomenal or p-consciousness, pure experience, the having of qualia. It seems clear it is p-consciousness that Goff, and I think all panpsychists, are taking about. No-one supposes electrons or rocks are making rational decisions, only having some kind of experience. The problem is that though we do seem to have direct acquaintance with that sort of consciousness, we haven't succeeded in saying anything much about it. In fact it seems that nothing we say about it can have been caused by it, because in itself it lacks causal powers. Now in one way this is exactly what Goff would expect; these difficulties are just those that come up when talking about qualia anyway, so in a back-handed sort of way we could even say they support his case. But if we're looking for good explanations, the bucket is coming up dry; no wonder we're tempted to go back and talk some more about the relatively tractable brain-first perspective.

In addition there are reasons to hesitate over the very idea that physical things have an intrinsic nature. Either this nature affects observable properties or it doesn't. If it does, then we can use its effects to learn about it and discuss it; to naturalise it, in fact, and bring it within the pale of science. If it doesn't - how can we talk about it? It might change radically or disappear and return, and we should never know. Goff rests his case on parsimony; we might counter that by observing that a theory that fills the cosmos with experiencing entities looks profligate in some respects. Isn't there a better strategy anyway? Goff wants to simplify by assuming that apparently dead matter is in fact inwardly experiential like us: but why not go the other way and believe that we actually are as dead matter seems to be; lacking in qualic, phenomenal experience? Why not conclude that a-consciousness is all we've got, and that the semblance of p-consciousness is a delusion, as sceptics have argued? We can certainly debate on many other grounds whether that view is correct, but it seems hard to deny that dispensing with phenomenal experience altogether must be the most parsimonious take on the subject.

So I'm not convinced, but I think that within the natural constraints of a blog post, Goff does make a lucid and attractive presentation of his case.

(In another post, Goff brings further arguments to defend the idea of intrinsic natures. We'll have a look at those, though as I ought to have said in the first place, one should really read his book to get the full view.)

Intrinsic natures

29 October 2017

Following up on his post about the simplicity argument for panpsychism, Philip Goff went on to defend the idea that physical things must have an intrinsic nature. Actually, it would be more accurate to say he attacks the idea that they don't have intrinsic natures. Those who think that listing the causal properties of a thing exhausts what we can say about its physical nature are causal structuralists, he says, committed to the view that everything reduces to dispositions; dispositions to burn, to attract, or to break, for example.

But when we come to characterise these dispositions, we find we can only do it in terms of other dispositions. A disposition to burn may involve dispositions to glow, get hot, generate ash, and so on. So we get involved in an endless circularity. Some might argue that this is OK, that we can cope with a network of mutual definitions that is, in the end, self-supporting; Goff says this is as unsatisfactory as making our living by taking in each other's washing.

There's a problem there, certainly. I think a bit more work is needed to nail down the idea that to reject intrinsic natures is necessarily to embrace causal structuralism, but no doubt Goff has done that in his fuller treatment. A more serious gap, it seems to me, is an explanation of how intrinsic natures get us out of this bind.

It seems to me that in practice we do not take the scholarly approach of identifying a thing through its definition; more usually we just show people. What is fire? This, we say, displaying a lit match. Goff gives an amusing example of three boxes containing a Splurge, a Blurge, and a Kurge, each defined in terms of the next in an inescapable circle. But wouldn't you open the box?

We could perhaps argue that recognising the Splurge is just grasping its intrinsic nature. But actually we would recognise it by sight, which depends on its causal properties; its disposition to reflect light, if you like. Those causal properties cannot have anything to do with its intrinsic nature, which seems to drop out of the explanation; in fact its intrinsic nature could logically change without affecting the causal properties at all.

This apparently radical uselessness of intrinsic properties, like the similar ineffectual nature of qualia, is what causes me the greatest difficulty with a perspective that would otherwise have some appeal.

Renormalisation

5 November 2017

An intriguing but puzzling paper from Simon DeDeo.

He begins by noting that while physics is good at generalised predictions, it fails to predict the particular. Working at the blackboard we can deduce laws governing the genesis of stars, but nothing about the specific existence of the blackboard. He sees this as a gap and for reasons that remain obscure to me he sees it as a matter of origins; the origin of society, consciousness, etc. To me, it's about their nature; assuming it's about origins constrains the possible answers unnecessarily to causal accounts.

Contrary to our expectations, says DeDeo, it's relatively easy to describe everything, but hard to describe just one thing - the Frame Problem is an example where it's the specifics that trip us up. By contrast, with the Library of Babel, Borges effortlessly gave us a description of everything. The Library of Babel is an imagined collection which contains every possible ordering of the letters of the alphabet; the extraordinary thing about it is that although it is finite, it contains every possible text - all the ones that were never written as well as all the ones that were.

We could quite easily write a computer program to find, within the library, all occurrences of the text string 'Shakespeare', says DeDeo; but there's no way of finding all the texts about Shakespeare that make sense. That's surely true. DeDeo says this is because what we're asking for is more than just pattern matching. In particular, he says, we need self-reference. I can't make out why he thinks that, and I'm pretty sure he's wrong, though I might well be missing the point. To me, it seems clear that in order to identify texts that make sense, we need to consider meanings, which are not about self-reference but reference to other things. In fact, context and meaning are of the essence. One book from the Library of Babel contains all books if we are allowed to apply to it an arbitrary interpretation or encoding of our choice; equally any book is nonsense if we don't know how to read it.

But for DeDeo this is a truth with a promising mathematical feel. We just need to elucidate the origin of self-reference, which he thinks lies in memory at least partly. The curious thing, in his eyes, is that physics only seems to require (or allow) certain levels of self-reference. We have velocity, we have acceleration, we have changes in acceleration; but models of worlds that have laws about third- or higher-order entities like changes in acceleration tend to be unstable, with runaway geometrical increases messing everything up.

So maybe we shouldn't go there? The funny thing is, we seem to be able to sense a third-order physical entity. A change in acceleration is known as 'jerk' and we certainly feel jerked in some situations. I have to say I doubt this. DeDeo mentions the sudden motions of a lift, but those, like all instances of jerk, surely correspond with an acceleration? I wonder whether the concept of jerk as a distinct entity in physics isn't redundant. For DeDeo, we perceive it through the use of memory, and this is the key to how we perceive other particularities not evident from the laws of physics. We tend to deal with coarse-grained laws, but the fine-grained detail is waiting to trip us up.

It’s not all bad news; perhaps, DeDeo speculates, there are new levels we have yet to explore…

I’m very unsure I’ve correctly understood what he’s proposing, and the fact that it seems to miss the real point (meaning and context) might well be a sign it’s me that’s not really getting it. Any thoughts?

Time Travel Consciousness

12 November 2017

In What's Next? Time Travel and Phenomenal Continuity Giuliano Torrengo and Valerio Buonomo argue that our personal identity is about continuity of phenomenal experience, not such psychological matters as memory (championed by John Locke). They refer to this phenomenal continuity as the 'stream of consciousness'. I'm not sure that William James, who I believe originated the phrase, would have seen the stream of consciousness as being distinct from the series of psychological states in our minds, but it is a handy label.

To support their case, Torrengo and Buonomo have a couple of thought experiments. The first one involves a couple of imaginary machines. One machine transfers the 'stream of consciousness' from one person to another while leaving the psychology (memories, beliefs, intentions) behind, the other does the reverse, moving psychology but not phenomenology. Torrengo and Buonomo argue that having your opinions, beliefs and intentions changed, while the stream of consciousness remained intact would be akin to a thorough brainwashing. Your politics might suddenly change, but you would still be the same person. Contrariwise, if your continuity of experience moved over to a different body, it would feel as if you had gone with it.

That is plausible enough, but there are undoubtedly people would refuse to accept it because they would deny that this separation of phenom and psych is possible, or crucially, even conceivable. This might be because they think the two are essentially identical, or because they think phenomenal experience arises directly out of psychology. Some would probably deny that phenomenal experience in this sense even exists.

There is a bit of scope for clarification about what variety of phenomenal experience Torrengo and Buonomo have in mind. At one point they speak of it as including thought, which sounds sort of psychological to me. By invoking machines, their thought experiment shows that their stream of consciousness is technologically tractable, not the kind of slippery qualic experience which lies outside the realm of physics.

Still, thought experiments don't claim to be proofs; they appeal to intuition and introspection, and with some residual reservations, Torrengo and Buonomo seem to have one that works on that level. They consider three objections. The first complains that we don't know how rich the stream of consciousness must be in order to be the bearer of identity. Perhaps if it becomes attentuated too much it will cease to work? This business of a minimum richness seems to emerge out of the blue and in fact Torrengo and Buonomo dismiss it as a point which affects all 'mentalist' theories. The second objection is a clever one; it says we can only identify a stream of consciousness in relation to a person in the first place, so using it as a criterion of personal identity begs the question. Torrengo and Buonomo essentially deny that there needs to be an experiencing subject over and above the stream of consciousness. The third challenge arises

from gaps; if identity depends on continuity, then what happens when we fall asleep and experience ceases? Do we acquire a new identity? Here it seems Torrengo and Buonomo fall back on a defence used by others; that strictly speaking it is the continuity of capacity for a given stream of consciousness that matters. I think a determined opponent might press further attacks on that.

Perhaps, though, the more challenging and interesting thought experiment is the second, involving time travel. Torrengo is the founder of the Centre for Philosophy of Time in Milan, and has a substantial body of work on the the experience of time and related matters, so this is his home turf in a sense. The thought experiment is quite simple; Lally invents a time machine and uses it to spend a day in sixties London. There are two ways of ordering her experience. One is the way she would see it; her earlier life, the time trip, her later life. The other is according to 'objective' time; she appears in old London Town and then vanishes; much later lives her early life, then is absent for a short while and finally lives her later life. These can't both be right, suggest Torrengo and Buonomo, and so it must surely be that her experience goes off on the former course while her psychology goes the other way.

This doesn't make much sense to me, so perhaps I have misunderstood. Certainly there are two time lines, but Lally surely follows one and remains whole? It isn't the case that when she is in sixties London she lacks intentions or beliefs, having somehow left those behind. Torrengo and Buonomo almost seem to think that is the case; they say it is possible to imagine her in sixties London not remembering who she is. Who knows, perhaps time machines do work like that, but if so we're running into one of the weaknesses of thought experiments methodologically; if you assume something impossible like time travel to begin with, it's hard to have strong intuitions about what follows.

At the end of the day I'm left with a sceptical feeling not about Torrengo and Buonomo's ideas in particular but about the whole enterprise of trying to reduce or analyse the concept of personal identity. It is, after all, a particular case of identity and wouldn't identity be a good candidate for being one of those 'primitive' ideas that we just have to start with? I don't know; or perhaps I should just say there is a person who doesn't know, whose identity I leave unprobed.

Robot Tax

18 November 2017

Xavier Oberson suggests how we might tax robots; I think it raises a number of difficult issues about the idea. You can see him expound the same ideas in a video interview here.

It's not made altogether clear here why we should apply special taxes to robots at all. Broadly I'd say there are two distinct reasons why governments tax things. The first is the Money reason; tax is simply about revenue. If that were really all, then we should design our taxes to be simple, easy to collect, hard to avoid, and neutral in effect. We wouldn't single out particular goods or activities for special treatment. However, there is a second and very different reason for taxing things, namely to discourage them; we could call it the 'Moral' reason. There are things we don't want to criminalise or forbid, but whose excessive use we should like to discourage - alcohol and tobacco, for example, which most countries apply special excise duties to.

Usually both reasons apply to some degree. Income tax is mainly about raising money, for example (we don't think there should be less income about); but generally tax regimes are bit harder on income which is considered unearned or undeserved.

Which is the main reason for taxing robots? I don't think they're going to be an easy way of raising money. If they make companies more profitable, then there should be a bit more money to target, so there's that; but as I'll explain below, I think there are big difficulties over definitions and avoidance. It seems clear that the main motivation is moral, to discourage too much use of robots. Oberson's heading suggests robot tax might offset revenue shortfall, a Money matter, but in his piece he sets the proposal squarely in the context of robots taking jobs from humans. I don't know whether that is something we should really worry about - some say not - but he's surely right that that Moral fear is where the impetus for tax is mainly coming from.

In fact, it seems to me that Oberson is thinking mainly in terms of mechanical men. He sees robots replacing humans on more or less a like-for-like basis. Initially the business might be charged tax on the basis of the wages it would have had to pay humans to get the same production; in the long run, as robots gain full agency, this arrangement could segue into the robots themselves gaining legal personhood and responsibility for paying a sort of robot income tax

Alas, we are nowhere near robots having that level of agency, and in fact I'd say progress towards it is negligible to date. Oberson is right when he says that if robots did gain personhood such arrangements could be quite straightforward. After all, when robots do achieve human levels of agency you'll presumably have to pay them to work instead of just switching them on and issuing commands, so at that point, their liability to conventional taxes should be uncontroversial! But that is too distant a prospect to require new tax arrangements just yet. If we did persist in trying to apply a regime based on robots themselves paying, it could only become an elaborate way of making the robot owner pay. It would be unnecessarily complicated and the pretence that the robots were like us might tend to devalue the genuine agency of human beings, something we should do well to steer clear of.

In general I think Oberson is pretty optimistic about robot capacity. He says

> *Today robots become lawyers, doctors, bankers, social workers, nurses and even entertainers*.

To which one can only say, no they don't. What is he thinking of? I can only suppose he has in mind expert systems or similar programs that can perhaps offer some legal advice, help with medical diagnosis, and provide online banking. These things may indeed have some impact on employment - most clearly in the case of counter staff in banks (though it's debatable how far robots are involved with that - banks were moving towards call centres even before they got into online stuff). But it's a massive overstatement to say robots right now become lawyers, etc. On social work and entertainment I can't really come up with any good examples of robots replacing humans - any ideas?

So, what if we just want to tax human businesses for the ownership or use of robots? One thing Oberson suggests is a value added tax on robots. This is strange because the purchase of robots or the supply of robot services is surely subject to VAT already in those countries that have VAT, like most things. In principle we could apply a higher rate to robots, and that would indeed be one of the most practical approaches, though in Europe we would run up against the EU's strong preference that the system should move towards having a single rate applied to everything (the people who designed VAT for the EU were strong believers in the Money reason for taxation rather than the Moral one).

What's a robot, though? It's pretty unlikely in fact that we are going to be dealing with mechanical men very much. Most factory robots at present are very unhumanoid machines. Is each automatic painting/assembly arm a single robot? What if they are all run from a single central computer? Does that mean the whole factory is a single robot, so I pay much less than the fellow down the road who put separate processors in hundreds of his machines? What if my robots are controlled by software run off the Internet? Is the Internet itself one big robot - and if so, who pays its taxes?

Surely, also, to qualify as a robot my machines must have a certain level of complexity? Now perhaps the fellow up the road wins after all. He split his processes into very small elements. Nearly every machine in his factory has a chip in it, but individually they only carry out very simple operations, far too simple for any of them to be considered robots. Why, the chips in his machines are far less complex than the ones in your dishwasher! Yet together they mean no humans are needed.

What if we take up the idea, floated by Oberson, that the tax can be based on the salaries I would have had to pay if I had continued to use humans? Let's suppose I lay off ten humans and install ten robots; I achieve a productivity gain of 20% and pay tax equal to say, 10% of ten salaries. A year later the tax inspector notices that a hundred further humans have gone and productivity is now up 500%. He notices that the robots all have a dial which used to be set to 'snail' and is now turned to 'leopard'.

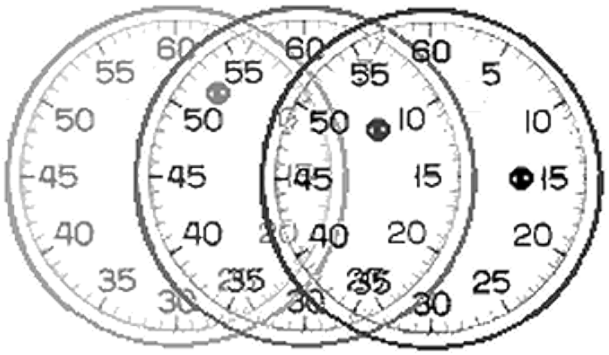
Oh, I say, but the productivity gains were due to other factors. We re-engineered our processes and bought new non-robot tools which enabled the speed improvements. We would have got the same gains with humans. The human lay-offs are due to a reduction in activity in an area that happened not to be automated. Anyway, come on, am I expected to pay tax on notional salaries that don't relate in any way to my current business? Forever?

These are just examples from my own limited imagination, but once you start taxing something a lot of very clever accountants are going to be working hard on devising highly sophisticated schemes.

Overall I'm inclined to accept the argument that applying special taxes to robots is just a bad idea altogether. If we succeed in discouraging the use of robots, our businesses will lose out on productivity gains and suffer from the competition of businesses in countries where robots get off scot-free. We could protect our human workers from those foreign robots with tariffs and quotas, but in the long run we would fall behind those others economically and get into a bad place. It can fairly be argued that we might use tax, not as a permanent discouragement to automation, but as a means of slowing things to a manageable transitional pace, but it seems to me that in practice there might be more of a case for subsidising research and implementation in order to get maximum productivity gains as soon as possible! In practice I wouldn't bet against governments convincing themselves it's a good idea to both subsidise and tax robots at the same time - but I know nothing about economics, of course, something you may feel has been made sufficiently clear already.

Quando Libet

26 November 2017



New light on Libet's challenge to free will; this interesting BQO piece by Ari N Schulman focuses on a talk by Patrick Haggard.

Libet's research has been much discussed, here as well as elsewhere. He asked subjects to move their hand at a time of their choosing, while measuring neural activity in their brain. He found that the occurrence of a 'Readiness Potential' or RP (something identified by earlier researchers) always preceded the hand movement. But it also preceded the time of the decision, as reported by subjects. So it seemed the decision was made and clearly registered in brain activity as an RP before the subjects' conscious thought processes had anything to do with it. The research, often reproduced and confirmed, seemed to provide a solid scientific demonstration that our feeling of having free conscious control over our own behaviour is a delusion.

However, recent research by Aaron Schurger shows that we need to re-evaluate the RP. In the past it has been seen as the particular precursor of intentional action; in fact it seems to be simply a peak in the continuing ebb and flow of neural activity. Peaks like this occur all the time, and may well be the precursors of various neural events, not just deliberate action. It's true that action requires a peak of activity like this, but it's far from true that all such peaks lead to action, or that the decision to act occurred when the peak emerged. If we begin with an action and look back, we'll always find an RP, but not all RPs are connected with actions. It seems to me a bit like a surfer, who has to wait for a wave before leaping on the board; but let's suppose there are plenty of good waves and the surfer is certainly not deprived of his ability to decide when to go.

This account dispels the impression that there is a fatal difficulty here for free will (of course there are plenty of other arguments on that topic); I think it also sits rather nicely with Libet's own finding that we have 'free won't' – ie that even after an RP has been detected, subjects can still veto the action.

Haggard, who has done extensive work in this area, accepts that RPs need another look; but he contends that we can find more reliable precursors of action. His own research analysed neural activity and found significantly lowered variability before actions, rather as though the disorganised neural activity of the brain pulled together just before an action was initiated.

Haggard's experiments were designed to address another common criticism of Libet's experiments, namely the artificiality of the decision involved. Being told to make your hand for no reason at a moment of your choosing is very unlike most of the decisions we make. In particular, it seems random, whereas it is argued that proper free will takes account of the pros and cons. Haggard asked subjects to perform a series of simple button-pushing tasks; the next task might follow quickly, or after a delay which could be several minutes long. Subjects could skip to the next task if they found the wait tedious, but that would reduce the cash rewards they got for performing the tasks. This weighing of boredom against profit is much more like a real decision.

Haggard persuasively claims that the essence of Libet's results is upheld and refreshed by his results, so in principle we are back where we started. Does this mean there's no free will? Schulman thinks not, because on certain reasonable and well-established conceptions of free will it can 'work in concert with decisional impulses', and need not be threatened by Haggard's success in measuring those impulses.

For myself, I stick with a point mentioned by Schurger; making a decision and becoming aware of the decision are two distinct events, and it is not really surprising or threatening that the awareness comes a short time after the actual decision. It's safe to predict that we haven't heard the last of the topic, however.

Jerry Fodor

4 December 2017

Jerry Fodor died last week at the age of 82. I think he had three qualities that make a good philosopher. He really wanted the truth (not everyone is that bothered about it); he was up for a fight about it (in argumentative terms); and he had the gift of expressing his ideas clearly. Georges Rey, in the Daily Nous piece, professes surprise over Fodor's unaccountable habit of choosing simple everyday examples rather than prestigious but obscure academic ones: but even Rey shows his appreciation of a vivid comparison by quoting Dennett's lively simile of Fodor as trampoline.

Good writing in philosophy is not just a presentational matter, I think; to express yourself clearly and memorably you have to have ideas that are clear and cogent in the first place; a confused or laborious exposition raises the suspicion that you're not really that sure what you're talking about yourself.

Not that well-expressed ideas are always true ones, and in fact I don't think Fodorism, stimulating as it is, is ever likely to be accepted as correct. The bold hypothesis of a language of thought, sometimes called *mentalese*, in which all our thinking is done, never really looked attractive to most. Personally it strikes me as an unnecessary deferral; something in the brain has to explain language, and saying it's another language just puts the job off. In fairness, empirical evidence might show that things are like that, though I don't see it happening at present. Fodor himself linked the idea with a belief in a comprehensive inborn conceptual apparatus; we never learn new concepts, just activate ones that were already there. The idea of inborn understanding has a respectable pedigree, but if Plato couldn't sell it, Fodor was probably never going to pull it off either.

As I say, these are largely empirical matters and someone fresh to the discussion might wonder why discussion was ever thought to be an adequate method; aren't these issues for science? Or at least, shouldn't the armchair guys shut up for a bit until the neurologists can give them a few more pointers? You might well think the same about Fodor's other celebrated book, *The Modularity of Mind*. Isn't a day with a scanner going to tell you more about that than a month of psychological argumentation?

But the truth is that research can't proceed in a vacuum; without hypotheses to invalidate or a framework of concepts to test and apply, it becomes mere data collection. The concepts and perspectives that Fodor supplied are as stimulating as ever and re-reading his books will still challenge and sharpen anyone's thinking.

Perhaps my favourite was his riposte to Stephen Pinker, *The Mind Doesn't Work That Way.* So I've been down into the cobwebbed cellars of Conscious Entities and retrieved one of the 'lost posts', one I wrote in 2005, which describes it. (I used to put red lines in things in those days for reasons that now elude me).

Here it is...

Not like that.

30 January 2005

Jerry Fodor's 2001 book 'The Mind Doesn't Work That Way' makes a cogent and witty deflationary case. In some ways, it's the best summary of the current state of affairs I've read; which means, alas, that it is almost entirely negative. Fodor's constant position is that the Computational Theory of Mind (CTM) is the only remotely plausible theory we have – and remotely plausible theories are better than no theories at all. But although he continues to emphasise that this is a reason for investigating the representational system which CTM implies, he now feels the times, and the bouncy optimism of Steven Pinker and Henry Plotkin in particular, call for a little Eeyoreish accentuation of the negative. Sure, CTM is the best theory we have, but that doesn't mean it's actually much good. Surely no-one ought to think it's the complete explanation of all cognitive processes – least of all the mysteries of consciousness! It isn't just computation that has been over-estimated, either – there are also limits to how far you can go with modularism too – though again, it's a view with which Fodor himself is particularly associated.

The starting point for both Fodor and those he parts company with, is the idea that logical deduction probably gets done by the brain in essentially the same way as it is done on paper by a logician or in electronic digits by a computer, namely by the formal manipulation of syntactically structured representations, or to put it slightly less polysyllabically, by moving symbols around according to rules. It's fairly plausible that this is true at least for some cognitive processes, but there is a wide scope for argument about whether this ability is the latest and most superficial abstract achievement of the brain, or something that plays an essential role in the engine room of thought.

Don't you think, to digress for a moment, that formal logic is consistently over-rated in these discussions? It enjoys tremendous intellectual prestige: associated for centuries with the near-holy name of Aristotle, its reputation as the ultimate crystallisation of rationality has been renewed in modern times by its close association with computers – yet its powers are actually feeble. Arithmetic is invoked regularly in everyday life, but no-one ever resorted to syllogisms or predicate calculus to help them make practical decisions. I think the truth is that logic is only one example of a much wider reasoning capacity which stems from our ability to recognise a variety of continuities and identities in the world, including causal ones.

Up to a point, Fodor might go along with this. The problem with formal logical operations, he says, is that they are concerned exclusively with local properties: if you've got the logical formula, you don't need to look elsewhere to determine its validity (in fact, you mustn't). But that's not the way much of cognition works: frequently the context is indispensable to judgements about beliefs. He quotes the example of judgements about simplicity: the same thought which complicates one theory simplifies another and you therefore can't decide whether hypothesis A is a complicating factor without considering facts external to the hypothesis: in fact, the wider global context. We need the faculty of global or abductive reasoning to get us out of the problem, but that's exactly what formal logic doesn't supply. We're back, in another form, with the problem of relevance, or in practical terms, the old demon of the frame problem; how can a computer (or how do human beings) consider just the relevant facts without considering all the irrelevant ones first – if only to determine their relevance?

One strategy for dealing with this problem (other than ignoring it) is to hope that we can leave logic to do what logic does best, and supplement it with appropriate heuristic approaches: instead of deducing the answer we'll use efficient methods of searching around for it. The snag, says Fodor, is that you need to apply the appropriate heuristic approach, and deciding which it is requires the same grasp of relevance, the same abduction, which we were lacking in the first place.

Another promising-looking strategy would be a connectionist, neural network approach. After all, our problem comes from the need to reason globally, holistically if you like, and that is is often said to be a characteristic virtue of neural networks. But Fodor's contempt for connectionism knows few bounds; networks, he says, can't even deliver the classical logic that we had to begin with. In a network the properties of a node are determined entirely by its position within the network: it follows that nodes cannot retain symbolic identity and be recombined in different patterns, a basic requirement of the symbols in formal logic. Classical logic may not be able to answer the global question, but connectionism, in Fodor's eyes, doesn't get as far as being able to ask it.

It looks to me as if one avenue of escape is left open here: it seems to be Fodor's assumption that only single nodes of a network are available to do symbolic duty, but might it not be the case that particular patterns of connection and activation could play that role? You can't, by definition, have the same node in two different places: but you could have the same pattern realised in two different parts of a network. However, I think there might be other reasons to doubt whether connectionism is the answer. Perhaps, specifically, networks are just too holistic: we need to be able to bring in contextual factors to solve our problems, but only the right ones. Treating everything as relevant is just as bad as not recognising contextual factors at all.

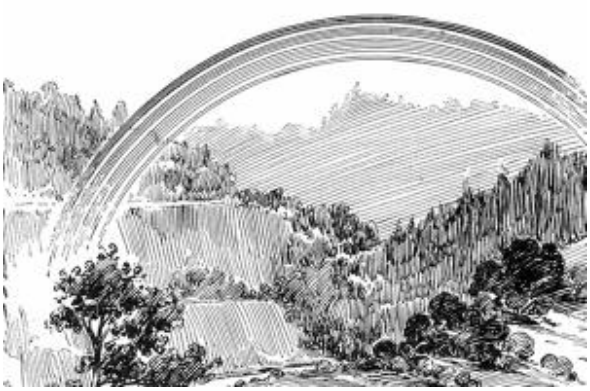
Be that as it may, Fodor still has one further strategy to consider, of course – modularity. Instead of trying to develop an all-purpose cognitive machine which can deal with anything the world might throw at it, we might set up restricted modules which only deal with restricted domains. The module only gets fed certain kinds of thing to reason about: contextual issues become manageable because the context is restricted to the small domain, which can be exhaustively searched if necessary. Fodor, as he says, is an advocate of modules for certain cognitive purposes, but not 'massive modularity', the idea that all, or nearly all, mental functions can be constructed out of modules. For one thing, what mechanism can you use to decide what a given module should be 'fed' with? For some sensory functions, it may be relatively easy: you can just hard-wire various inputs from the eyes to your initial visual processing module; but for higher-level cognition something has to decide whether a given input representation is one of the right kind of inputs for module M1 or M2. Such a function cannot itself operate within a restricted domain (unless it too has an earlier function deciding what to feed to it, in which case an infinite regress looms); it has to deal with the global array of possible inputs: but in that case, as before, classical logic will not avail and once again we need the abductive reasoning which we haven't got.

In short, 'By all the signs, the cognitive mind is up to its ghostly ears in abduction. And we do not know how abduction works.'

I'm afraid that seems to be true.

Phenomenal Epiphenomenalism

11 December 2017



Our conscious minds are driven by unconscious processes, much as it may seem otherwise, say David A. Oakley and Peter W. Halligan.

To summarise very briefly, they suggest three distinct psychological elements are at work. The first, itself made up of various executive processes, is what we might call the invisible works; the various unconscious mechanisms that supply the content of what we generally consider conscious thought. Introspection shows that conscious thoughts often seem to pop up out of nowhere, so we should be ready enough to agree that consciousness is not entirely self-sustaining. When we wake up we generally find that the stream of consciousness is already a going concern. The authors also mention, in support of their case, various experiments. Some of these were on hypnotised subjects, which you might feel detracts from their credibility in explaining normal thought processes. Other 'priming' effects have also taken a bit of a knock in the recent trouble over reproducibility. But I wouldn't make heavy weather of these points; the general contention that the contents of consciousness are generated by unconscious processes (at least to a great extent) seems to me one that few would object to. How could it be otherwise? It would be most peculiar if consciousness were a closed loop, like some Ouroboros swallowing its own tail.

The second element is a continuously generated personal narrative. This is an essentially passive record of some of the content generated by the 'invisible works', conditioned by an impression of selfhood and agency. The narrative has evolutionary survival value because it allows the exchange of experience and the co-ordination of behaviour and enables us to make good guesses at others' plans - the faculty often called 'theory of mind'.

At first glance I thought the authors, who are clearly out to denounce something as an epiphenomenon (a thing that is generated by the mind but has no influence on it), had this personal narrative as their target, but that isn't quite the case. While they see the narrative as essentially the passive product of the invisible works, they clearly believe it has some important influences on our behaviour through the way it enables us to talk to others and take their thoughts into account. One point which seems to me to need more prominence here is our ability to reflexively 'talk to ourselves' mentally and speculate about our own motives. I think the authors accept that this goes on, but some philosophers consider it a vital process, perhaps constitutive of consciousness, so I think they need to give it a substantial measure of attention. Indeed, it offers a potential explanation of free will and the impression of agency; it might be just the actions that flow from the reflexive process that we regard as our own free acts.

One question we might also ask is, why not identify the personal narrative as consciousness itself? It is what we remember of ourselves, after all. Alternatively, why not include the 'invisible works'? These hidden processes fall outside consciousness because (I think) they are not accessible to introspection; but must all conscious processes be introspectable? There's a

distinction between first and second-order awareness (between knowing and knowing that we know) which offers some alternatives here.

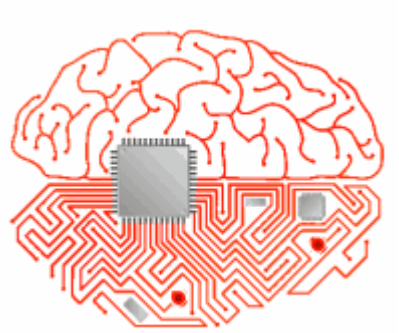
It's the third element that Oakley and Halligan really want to denounce; this is personal awareness, or what we might consider actual conscious experience. This, they say, is a kind of functionless emergent phenomenon. To ask its purpose is as futile as asking what a rainbow is for; it's just a by-product of the way things work, an epiphenomenon. It has no evolutionary value and resembles the whistle on a steam locomotive - powered by the machine but without influence over it (I've always thought that analogy short-changes whistles a bit, but never mind).

I suppose the main challenge here might be to ask why the authors think personal awareness is anything at all. It has no effects on mental processes, so any talk about it was not caused by the thing itself. Now we can talk about things that did not cause that talking; but those are typically abstract or imaginary entities. Given their broadly sceptical stance, should the authors be declaring that personal awareness is in fact not even an epiphenomenon, but a pure delusion?

I have my reservations about the structure suggested here, but it would be good to have clarity and, at the risk of damning with faint praise, this is certainly one of the more sensible proposals.

Augment Me?

18 December 2017



All I want for a Christmas is a new brain? There seems to have been quite a lot of discussion recently about the prospect of brain augmentation; adding in some extra computing power to the cognitive capacities we have already. Is this a good idea? I'm rather sceptical myself, but then I'm a bit of a Luddite in this area; I still don't like the idea of controlling a computer with voice commands all that much.

Hasn't evolution has already optimised the brain in certain important respects? I think there may be some truth in that, but It doesn't look as if evolution has done a perfect job. There are certainly one or two things about the human nervous system that look as if they could easily be improved. Think of the way our neural wiring crosses over from right to left for no particular reason. You could argue that although that serves no purpose it doesn't do any real harm either, but what about the way our retinas are wired up from the front instead of the back, creating an entirely unnecessary blind spot where the bundle of nerves actually enters the eye - a blind spot which our brain then stops us seeing, so we don't even know it's there?

Nobody is proposing to fix those issues, of course, but aren't there some obvious respects in which our brains could be improved by adding in some extra computational ability? Could we be more intelligent, perhaps? I think the definition of intelligence is controversial, but I'd say that if we could enhance our ability to recognise complex patterns quickly (which might be a big part of it) that would definitely be a bonus. Whether a chip could deliver that seems debatable at present.

Couldn't our memories be improved? Human memory appears to have remarkable capacity, but retaining and recalling just those bits of information we need has always been an issue. Perhaps related, we have that annoying inability to hold more than a handful of items in our minds at once, a limitation that makes it impossible for us to evaluate complex disjunctions and implications, so that we can't mentally follow a lot of branching possibilities very far. It certainly seems that computer records are in some respects sharper, more accurate, and easier to access than the normal human system (whatever the normal human system actually is). It would be great to remember any text at will, for example, or exactly what happened on any given date within our lives. Being able to recall faces and names with complete accuracy would be very helpful to some of us.

On top of that, couldn't we improve our capacity for logic so that we stop being stumped by those problems humans seem so bad at, like the Wason test? Or if nothing else, couldn't we just have the ability to work out any arithmetic problem instantly and flawlessly, the way any computer can do?

The key point here, I think, is integration. On the one hand we have a set of cognitive abilities that the human brain delivers. On the other, we have a different set delivered by computers. Can they be seamlessly integrated? The ideal augmentation would mean that, for example, if I need to multiply two seven-digit numbers I 'just see' the answer, the way I can just see that 3+1 is 4. If, on the contrary, I need to do something like ask in my head 'what is 6397107 multiplied by 8341977?' and then receive the answer spoken in an internal voice or displayed in an imagined

visual image, there isn't much advantage over using a calculator. In a similar way, we want our augmented memory or other capacity to just inform our thoughts directly, not be a capacity we can call up like an external facility.

So is seamless integration possible? I don't think it's possible to say for certain, but we seem to have achieved almost nothing to date. Attempts to plug into the brain so far have relied on setting up artificial linkages. Perhaps we detect a set of neurons that reliably fire when we think about tennis; then we can ask the subject to think about tennis when they want to signal 'yes', and detect the resulting activity. It sort of works, and might be of value for 'locked in' patients who can't communicate any other way, but it's very slow and clumsy otherwise - I don't think we know for sure whether it even works long-term or whether, for example, the tennis linkage gradually degrades.

What we really want to do is plug directly into consciousness, but we have little idea of how. The brain does not modularise its conscious activity to suit us, and it may be that the only places we can effectively communicate with it are the places where it draws data together for its existing input and output devices. We might be able to feed images into early visual processing or take output from nerves that control muscles, for example. But if we're reduced to that, how much better is that level of integration going to be than simply using our hands and eyes anyway? We can do a lot with those natural built-in interfaces ; simple reading and writing may well be the greatest artificial augmentation the brain can ever get. So although there may be some cool devices coming along at some stage, I don't think we can look for godlike augmented minds any time soon.

Incidentally, this problem of integration may be one good reason not to worry about robots taking over. If robots ever get human-style motivation, consciousness, and agency, the chances are that they will get them in broadly the way we get them. This suggests they will face the same integration problem that we do; seven-digit multiplication may be intrinsically no easier for them than it is for us. Yes, they will be able to access computers and use computation to help them, but you know, so can we. In fact we might be handier with that old keyboard than they are with their patched-together positronic brain-computer linkage.

Conspiracy

24 December 2017

Tom Stafford reports on an interesting review of the psychology of conspiracy theories - the persistent belief that 'they' are working secretly to conceal the truth about the assassination of JFK or the moon landings, for example. The review suggests current research is better at explaining the forces that drive conspiracy theories than at examining their psychological consequences. It seems the theories are motivated by three needs; for understanding, for safety/control, and for a positive image of yourself and the groups you belong to. But in point of fact, they are not very good at meeting these needs and may even make the people who subscribe to them feel worse.

Stafford suggests we could see this as maladaptive coping. He criticises some aspects of the review, in particular the way it defines conspiracy theories rather loosely, so that it seems to include reasonable conspiracy beliefs. You're not paranoid if they really are out to get you, after all.

Perhaps the most remarkable example of a genuine conspiracy is the way that around this time of year we all go to enormous lengths to convince our children that a fat old man is going to come down the chimney into their bedroom one night (a idea that terrifies a few of them, possibly the more rational ones). Kids who subscribed to the theory that parents, teachers and media were involved in a massive con would not be wrong, but would they be displaying early signs of a tendency to conspiracy theories? Is it rational, at a certain age, to believe in Santa?

So far as I recall, my own attitude back in the middle of the last century was neither exactly belief nor disbelief. I was well aware that people in department store grottos were proxies, merely dressed up as Father Christmas. I got as far as noting that the logistics of delivering presents to every child in the world in a single night were challenging, and vaguely hypothesised that the job was done by similar proxies, maybe one for each street. But I didn't worry about it much. There were lots of things I didn't fully understand at the time. I didn't really know how department stores came to be full of stuff anyway - why worry about Santa's grotto particularly? You could well say that my attitude to Santa back then was pretty much what my attitude to quantum physics is now. I don't really understand it, and parts of it don't seem to make any sense. But people I basically trust have got this for me, so I'm happy to take their word (just to be quite clear here, I am not suggesting that quantum physics is a massive conspiracy).

The matter of who you trust is, I think, at the root of the conspiracy theory thing generally. We all have to take a lot of things on trust from appropriate authorities. An essential and probably under-examined part of the education system is about teaching people which authorities to trust, and much of the academic system of peer review and publication, unsatisfactory as it is*, is about keeping authoritative sources identifiable and reliable. People who believe in conspiracy theories have flaws in their judgement about which authorities to accept.

Not that this is simple. Trusting authority is a tricky business which needs to be balanced with an ability to evaluate and critique even reliable authorities. People who have been thoroughly

educated may be weak on this side, inclined to believe what they read and pay more attention to the manifesto and the statement of principles than what is actually happening. Uneducated people may be more inclined to use their own observation and reason on the basis of perceived personality. Sometimes this works better, an excellent reason why everyone should have the vote. They say that cab driver off the 'seven up' observed around the turn of the century that the folks in the City were having a big party; in ten or fifteen years, he said, we'll be told it's all gone wrong and the bill is down to us. You can't say that's a detailed prediction of the crash, and it sounds a little conspiracyish, but it's a good deal better than the financial experts of the day managed.

Perhaps the Father Christmas Conspiracy is the way we help our children sharpen up their understanding of the need to balance proper acknowledgement of reliable authority with prudent, independent use of common sense.

Merry Christmas!

**I think we ought to set up a Universal Academy which publishes free access papers and a great Summa Scientia, citation in which would be the gold standard of sound and important research. It wouldn't be cheap, but maybe if we could get some kind of EU/USA rivalry going we could get two Academies?*'

Not Really Feeling It

7 January 2018

It's not just that we don't know how anaesthetics work - we don't even know for sure that they work. Joshua Rothman's review of a new book on the subject by Kate Cole-Adams quotes poignant stories of people on the operating table who may have been aware of what was going on. In some cases the chance remarks of medical staff seem to have worked almost like post-hypnotic suggestions: so perhaps all surgeons should loudly say that the patient is going to recover and feel better than ever, with new energy and confidence.

How is it that after all this time, we don't know how anaesthetics work? As the piece aptly remarks, it's about losing consciousness, and since we don't know clearly what that is or how we come to have it, it's no surprise that its suspension is also hard to understand. To add to the confusion, it seems that common anaesthetics paralyse plants, too. Surely it's our nervous system anaesthetics mainly affect - but plants don't even have a nervous system!

But come on, don't we at least know that it really does work? Most of us have been through it, after all, and few have weird experiences; we just don't feel the pain - or anything. The problem, as we've discussed before, is telling whether we don't feel the pain, or whether we feel it but don't remember it. This is an example of a philosophical problem that is far from being a purely academic matter.

It seems anaesthetics really do (at least) three different things. They paralyse the patient, making it easier to cut into them without adverse reactions, they remove conscious awareness or modulate it (it seems some drugs don't stop you being aware of the pain, they just stop you caring about it somehow), and they stop the recording of memories, so you don't recall the pain afterwards. Anaesthetists have a range of drugs to produce each of these effects. In many cases there is little doubt about their effectiveness. If a drug leaves you awake but feeling no pain, or if it simply leaves you with no memory, there's not that much scope for argument. The problem arises when it comes to anaesthetics that are supposed to 'knock you out'. The received wisdom is that they just blank out your awareness for a period, but as the review points out, there are some indications that instead they merely paralyse you and wipe your memory. The medical profession doesn't have a good record of taking these issues very seriously; I've read that for years children were operated on after being given drugs that were known to do little more than paralyse them (hey, kids don't feel pain, not really; next thing you'll be telling me plants do...).

Actually, views about this are split; a considerable proportion of people take the view that if their memory is wiped, they don't really care about having been in pain. It's not a view I share (I'm an unashamed coward when it comes to pain), but it has some interesting implications. If we can make a painful operation OK by giving mnestics to remove all recollection, perhaps we should routinely do the same for victims of accidents. Or do doctors sometimes do that already...?

Fundamentals

15 January 2018

You may already have seen Jochen's essay *Four Verses from the Daodejing*, an entry in this year's FQXi competition. It's a thought-provoking piece, so here are a few of the ones it provoked in me. In general I think it features a mix of alarming and sound reasoning which leads to a true yet perplexing conclusion.

In brief Jochen suggests that we apprehend the world only through models; in fact our minds deal only with these models. Modelling and computation are in essence the same. However, the connection between model and world is non-computable (or we face an infinite regress). The connection is therefore opaque to our minds and inexpressible. Why not, then, identify it with that other inexpressible element of cognition, qualia? So qualia turn out to be the things that incomprehensibly link our mental models with the real world. When Mary sees red for the first time, she does learn a new, non-physical fact, namely what the connection between her mental model and real red is. (I'd have to say that as something she can't understand or express, it's a weird kind of knowledge, but so be it.)

I think to talk of modelling so generally is misleading, though Jochen's definition is itself broadly framed, which means I can't say he's wrong. In his terms it seems anything that uses data about the structure and causal functioning of X to make predictions about its behaviour would be a model. If you look at it that way, it's true that virtually all our cognition is modelling. But to me a model leads us to think of something more comprehensive and enduring than we ought. In my mind at least, it conjures up a sort of model village or homunculus, when what's really going on is something more fragmentary and ephemeral, with the brain lashing up a 'model' of my going to the shop for bread just now and then discarding it in favour of something different. I'd argue that we can't have comprehensive all-purpose models of ourselves (or anything) because models only ever model features relevant to a particular purpose or set of circumstances. If a model reproduced all my features it would in fact be me (by Leibniz' Law) and anyway the list of potentially relevant features goes on for ever.

The other thing I don't like about liberal use of modelling is that it makes us vulnerable to the view that we only experience the model, not the world. People have often thought things like this, but to me it's almost like the idea we never see distant planets, only telescope lenses.

Could qualia be the connection between model and world? It's a clever idea, one of those that turn out on reflection to not be vulnerable to many of the counterarguments that first spring to mind. My main problem is that it doesn't seem right phenomenologically. Arguments from one's own perception of phenomenology are inherently weak, but then we are sort of relying on phenomenology for our belief (if any) in qualia in the first place. A red quale doesn't seem like a connection, more like a property of the red thing; I'm not clear why or how I would be aware of this connection at all.

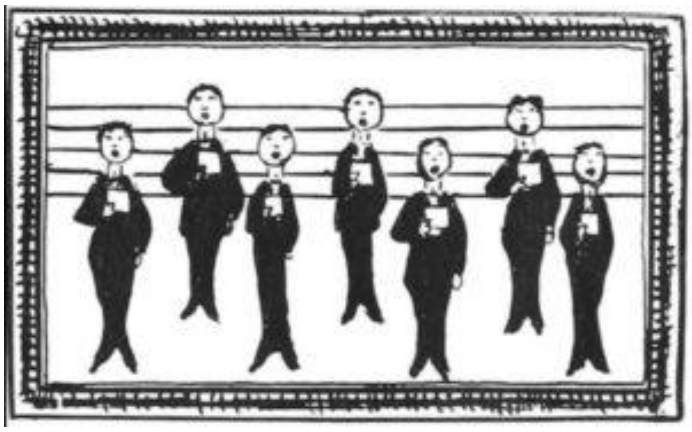
However, I think Jochen's final conclusion is both poignant and broadly true. He suggests that models can have fundamental aspects, the ones that define their essential functions - but the world is not under a similar obligation. It follows that there are no fundamentals about the world as a whole.

I think that's very likely true, and I'd make a very similar kind of argument in terms of explanation. There are no comprehensive explanations. Take a carrot. I can explain its nutritional and culinary properties, its biology, its metaphorical use as a motivator, its supposed status as the favourite foodstuff of rabbits, and lots of other aspects; but there is no total explanation that will account for every property I can come up with; in the end there is only the carrot. A demand for an explanation of the entire world is automatically a demand for just the kind of total explanation that cannot exist.

Although I believe this, I find it hard to accept; it leaves my mind with an unscratched itch. If we can't explain the world, how can we assimilate it? Through contemplation? Perhaps that would have been what Laozi would have advocated. More likely he would have told us to get on with ordinary life. Stop thinking, and end your problems!

Secret Harmonies

22 January 2018



We need a richer idea of consciousness and of our minds: Jenny Judge suggests that our experience of music in particular points to a need for an expanded conception of the mind to include visceral apprehension. Many who have championed the idea of an extended mind that isn't just identifiable with the brain alone will be nodding along, perhaps rhythmically.

For those of us who are a little entrenched in the limited idea of the mind as a matter of the brain doing computations on representations, Judge cunningly offers a couple of pieces of bait in the form of solid cognitive insights her wider perspective can yield. One is that we perceive and respond to rhythm in important ways, even without perceiving it at times. The timing of our utterances can actually change the way they are interpreted and carry significant information. A delay in giving assent can qualify the level of agreement, and apparently this is even culturally variable; the Japanese expect a snappy response, while in Denmark you can take your time without the risk of seeming grudging.

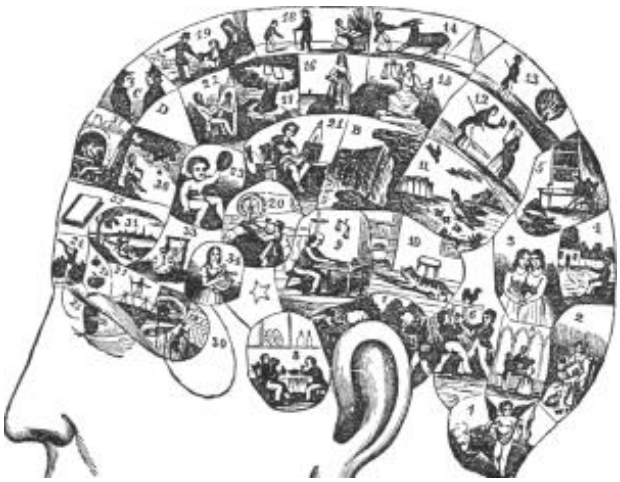
A second insight concerns entrainment, the tendency of connected vibrating systems to synchronise rhythm. Judge presents plausible evidence that a form of entrainment plays an important role in governing the activity of our minds and even of the neurons in the brain (so it ought not to be ignored, even by those who are initially happy with a narrower conception of cognition.

Judge discusses the challenges in perceiving music, with its complexity and its inherent sequentiality. The way we perceive time and motion is complex (we could add that sometimes the visual system just seems to label some things as 'moving' even though they are not perceived as changing place). But, wisely I think, she does not quite make the further case that the phenomenology of musical experience is peculiarly intractable. It's true that great music can cause us to experience emotional and cognitive states that we could never otherwise explore, and it would certainly be possible to base an argument of incredulity on this. Just as Leibniz professed disbelief about the possibility of a mill-type mechanism, however complex, producing awareness, or Brentano declared that intentionality was something else altogether, we could claim that musical experience just is not the kind of thing that physical processes could create. Such arguments are powerfully persuasive, but without some further explanation as to exactly why consciousness cannot arise from physical processing, they don't prove anything.

It would be hard to disagree, however, with the suggestion that our phenomenological experience really needs to be properly charted in a way that dies justice to its complexity. I'd have a go myself if I had any idea of how to set about it.'

Phrenology Revisited

29 January 2018



Could phrenology be true after all? The nineteenth century theory that bumps on the head correlated with personality traits has always been dismissed as absurd by most, right from its earliest days. Perhaps because of this, it has rarely received serious attention; but now O. Parker Jones, F. Alfaro-Almagro and S. Jbabdi have given it a properly rigorous examination, exploiting the resources of the UK Biobank.

Phrenology was based on ideas that are reasonable in themselves. First it supposes that different regions of the brain have different functions, following a consistent pattern in all humans. Second, it holds that if one of those functions plays an especially large part in a subject's mental life, the relevant area will be larger, and tend to swell outwards. Third, these swellings in the brain will be reflected in the shape of the skull, in a way which can be detected by careful measurement, or through palpation by a skilled phrenologist. These propositions may be wrong, but cannot be dismissed as inherently nutty.

Phrenology used a list of 'faculties' such as 'amativeness' and 'philoprogenitiveness', which mapped on to areas of the brain; the charts and models which resulted have a strange appeal and are still often used ornamentally by people who have no belief or even interest in the theory they embody. There were 27 faculties in all; some of them look a little odd (number 5 is 'Murder, carnivorousness (Destructiveness)' which even vegetarians might accept is a rather heterogeneous grouping), but most make prima facie sense. In order to test the theory, the researchers tried to match up the faculties with 'lifestyle' measures available in the Biobank data. It has to be said that some of these matches are better than others. Linking 'amativeness' with number of sexual partners, or the number faculty with mathematicians, seems reasonable: linking combativeness with solicitors, and self-esteem with bankers is possibly more debatable. The proposed linkage of cautiousness with frequency of alcohol intake must sure;y be intended as a negative correlation. The researchers confess that their choices here reflect their own faculty of mirth to some degree. In general, it seems the researchers did not end up using the job-related lifestyle measures, because there just weren't, for example, enough poets in the data.

In some cases there were strong correlations across faculties/lifestyle measures. Mostly these seem to have been unsurprising – writers, it turns out, tend to be good with words. More amusingly, the strongest correlation was between the ability to generate words and the number of sexual partners. If you want many partners, it seems you should be hanging out at poetry slams, not the gym.

It will perhaps come as no surprise that the researchers found no correlation between the faculties and skull curvature. This is enough, with appropriate caveats, to show that 'orthodox' phrenology is incorrect. We sort of knew that already; as the paper points out, studies based on the deficits caused by brain lesions in particular areas long ago showed that while there was

indeed localisation of function within the brain, the functions were not where phrenology said they ought to be. However, it leaves open the possibility that a reformed phrenology, with a different set of faculties, might still be sound.

This remaining possibility is removed by the second part of the study, which confirmed that there is no correlation between the bumpiness of the brain and the bumpiness of the skull. You cannot tell anything about the brain inside just by feeling the head that contains it. This too isn't really a huge surprise. If the shape of the skull were determined by the shape of the brain, we might expect our heads to show visible signs of the different lobes that make up the large-scale structure of the brain; we might expect that the skull would be skewed the way the brain is, with one hemisphere edging ahead of the other.

However, as the researchers rightly say, it is good to have these things tested with proper scientific rigour. They also suggest that their techniques might have some value in certain real medical conditions where the brain is at risk of being constrained by the skull. I suppose it's useful, too, to be reminded how popular a theory with no real scientific basis can become (not that we're in much danger of forgetting that these days…).

A Meeting Of Minds

5 February 2018

The self is real - it just, like Walt Whitman, contains multitudes. That's the case made by Serife Tekin in Aeon. She begins by rightly pointing out the current popularity of disbelief in the self. She traces antirealist thinking right back to Hume, who said he was never able to spot his self by introspection; all he ever came up with was a bundle of perceptions. Interestingly she picks out Dennett as a contemporary example of antirealism, but she could readily have pointed to several others who think the self is an illusion or misinterpretation, perhaps stemming from our cognitive limitations, or from the reflexivity that arises when we turn our mind on itself.

Tekin by contrast suggests the self is both real and open to proper scientific investigation. It's just that it has many forms; it is multitudinous. Borrowing from Neisser, she suggests five main dimensions of the self...

> *...the ecological self, or the embodied self in the physical world, which perceives and interacts with the physical environment; the interpersonal self, or the self embedded in the social world, which constitutes and is constituted by intersubjective relationships with others; the temporally extended self, or the self in time, which is grounded in memories of the past and anticipation of the future; the private self which is exposed to experiences available only to the first person and not to others; and finally the conceptual self, which (accurately or falsely) represents the self to the self by drawing on the properties or characteristics of not only the person but also the social and cultural context to which she belongs.*

I don't think these five types are meant to exhaust the variety of the self, which actually comes in a huge variety of shifting shapes. Nor are we meant to think that there is no basic unity; the five work together to provide an overall coherence of agency, though not without retaining some inner tensions and contradictions (nothing too strange psychologically in the idea that we may entertain contradictory thoughts and feelings in certain contexts.

The fivefold structure pays off because Tekin can give a separate account of how each can be addressed scientifically. The ecological self is easily observable, for example; for the interpersonal self we need to pay attention to social aspects, but no great problem there. The most difficult seems likely to be the private self; Tekin seems to think we can get to that simply by interviewing people about 'what it is like', which perhaps underrates the problems.

Overall, it's a sensible and appealing position. The curious thing is how close it seems to the kind of position taken by Dennett, here quoted as an example of antirealism. In fact, Dennett's ideas are more nuanced than some. He doesn't believe in a continuous, coherent self like a soul, but he is content to liken the self to a centre of gravity; not a real physical entity as such but a useful and harmless construction. As the author of the 'multiple drafts' theory of consciousness, I think he might rather like Tekin's multitudinousness; and her approach to the private self looks quite like his 'heterophenomenology' in which we give up trying to study ineffable inner experience, but happily give consideration to what people tell us about ineffable inner experience.

This raises the attractive possibility that sceptics and believers might end up constructing effectively identical models of the self, the only difference being that one side regards the model as an eliminative reduction while the other sees it as simply analysis. I find that a strangely cheering prospect.

The Ontological Gap

13 February 2018

There's a fundamental ontological difference between people and programs which means that uploading a mind into a machine is quite impossible.

I thought I'd get my view in first (hey, it's my blog), but I was inspired to do so by Beth Elderkin's compilation of expert views in Gizmodo, inspired in turn by Netflix's series Altered Carbon. The question of uploading is often discussed in terms of a hypothetical Star Trek style scanner and the puzzling thought experiments it enables. What if instead of producing a duplicate of me far away, the scanner produced two duplicates? What if my original body was not destroyed - which is me? But let's cut to the chase; digital data and a real person belong to different ontological realms. Digital data is a set of numbers, and so has a kind of eternal Platonic essence. A person is a messy entity bound into time and space. The former subsist, the latter exist; you cannot turn one into the other, any more than an integer can become a biscuit and get eaten.

Or look at it this way; a digitisation is a description. Descriptions, however good, do not contain the thing described (which is why the science of colour vision does not contain the colour red, as Mary found in the famous thought experiment).

OK, well, that that, see you next time... Oh, sorry, yes, the experts...

Actually there are many good points in the expert views. Susan Schneider makes three main ones. First, we don't know what features of a human brain are essential, so we cannot be sure we are reproducing them; quantum physics imposes some limits on how precisely we can copy th3 brain anyway. Second, the person left behind by a non-destructive scanner surely is still you, so a destructive scan amounts to death. Third, we don't know whether AI consciousness is possible at all. So no dice.

Anders Sandberg makes the philosophical point that it's debatable whether a scanner transfers identity. He tends to agree with Parfit that there is no truth of the matter about it. He also makes the technical point that scanning a brain in sufficient detail is an impossibly vast and challenging task, well beyond current technology at least. While a digital insert controlling a biological body seems feasible in theory, reshaping a biological brain is probably out of the question. He goes on to consider ethical objections to uploading, which don't convince him.

Randal Koene thinks uploading probably is possible. Human consciousness, according to the evidence, arises from brain processes; if we reproduce the processes, we reproduce the mind. The way forward may be through brain prostheses that replace damaged sections of brain, which might lead ultimately to a full replacement. He thinks we must pursue the possibility of uploading in order to escape from the ecological niche in which we may otherwise be trapped (I think humans have other potential ways around that problem).

Miguel A. L. Nicolelis dismisses the idea. Our minds are not digital at all, he says, and depend on information embedded in the brain tissue that cannot be extracted by digital means.

I’m basically with Nicolelis, I fear.

The Meta-Problem

Maybe there's a better strategy on consciousness? An early draft paper by David Chalmers suggests we turn from the Hard Problem (explaining why there is 'something it is like' to experience things) and address the Meta-Problem of why people think there is a Hard Problem; why we find the explanation of phenomenal experience problematic. The paper does make clear what Chalmers' own view is, but it primarily seeks to map the territory in a very useful way.

Why would we decide to focus on the Meta-Problem? For sceptics, who don't believe in phenomenal experience or think that the apparent problems about it stem from mistakes and delusions, it's a natural piece of tidying up. In fact, for sceptics why people think there's a problem may well be the only thing that really needs explaining or is capable of explanation. But Chalmers is not a sceptic. Although he acknowledges the merits of the broad sceptical case about phenomenal consciousness which Keith Frankish has recently championed under the label of illusionism, he believes it is indeed real and problematic. He believes, however, that illuminating the Meta-Problem through a programme of thoughtful and emprircal research might well help solve the Hard Problem itself, and is a matter of interest well beyond sceptical circles.

To put my cards on the table, I think he is over-optimistic, and seems to take too much comfort from the fact that there have to be physical and functional explanations for everything. It follows from that that there must be physical and functional explanations for our reports of experience, our reports of the problem, and our dispositions to speak of phenomenal experience, qualia, etc. But it does not follow that there must be adequate and satisfying explanations.

Certainly physical and functional explanations alone would not be good enough to banish our worries about phenomenal experience. They would not make the itch go away. In fact, I would argue that they are not even adequate for issues to do with the 'Easy Problem', roughly the question of how consciousness allows us to produce intelligent and well-directed behaviour. We usually look for higher-level explanations even there; notably explanations with an element of teleology - ones that tell us what things are for or what they are supposed to do. Such explanations can normally be cashed out safely in non-teleological terms, such as strictly-worded evolutionary accounts; but that does not mean they are dispensable or not needed in order for us to understand properly.

How much more challenging things are when we come to Hard Problem issues, where a claim that they lie beyond physics is of the essence. Chalmer's optimism is encapsulated in a sentence when he says...

> *Presumably there is at least a very close tie between the mechanisms that generate phenomenal reports and consciousness itself.*

There's your problem. Illusionists can be content with explanations that never touch on phenomenal consciousness because they don't think it exists, but no explanation that does not connect with it will satisfy qualophiles. But how can you connect with a phenomenon explanatorily without diagnosing its nature? It really seems that for believers, we have to solve the Hard Problem first (or at least, simultaneously) because believers are constrained to say that the appearance of a problem arises from a real problem.

Logically, that is not quite the case; we could say that our dispositions to talk about phenomenal experience arise from merely material causes, but just happen to be truthful about a second world of phenomenal experience or are truthful in light of a Leibnizian pre-established harmony. Some qualophiles are prepared to say that their utterances about qualia are not caused by qualia, so that position might appeal. To me the harmonised second world seems hopelessly redundant, and that is why something like illusionism is, at the end of the day, the only game in town.

I should make it clear that Chalmers by no means neglects the question of what sort of explanation will do; in fact he provides a rich and characteristically thorough discussion. It's more that in my opinion, he just doesn't know when he's beaten, which to be fair may be an outlook essential to the conduct of philosophy.

I say that something like illusionism seems to be the only game in town, though I don't quite call myself an illusionist. There's a presentational difficulty for me because I think the reality of experience, in an appropriate sense, is the nub of the matter. But you could situate my view as the form of illusionism which says the appearance of ineffable phenomenal experience arises from the mistaken assumption that particular real experiences should be within the explanatory scope of general physical theory.

I won't attempt to summarise the whole of Chalmers' discussion, which is detailed and illuminating; although I think he is doomed to disappointment, the project he proposes might well yield good new insights; it's often been the case that false philosophical positions were more fecund than true ones.

The Philosophy of Delirium

Is there any philosophy of delirium? I remember asserting breezily in the past that there was philosophy of everything - including the actual philosophy of everything and the philosophy of philosophy. But when asked recently, I couldn't come up with anything specifically on delirium, which in a way is surprising, given that it is an interesting mental state.

Hume, I gather, described two diseases of philosophy, characterised by either despair or unrealistic optimism in the face of the special difficulties a philosopher faces. The negative over-reaction he characterised as melancholy, the positive as delirium, in its euphoric sense. But that is not what we are after.

Historically I think that if delirium came up in discussion at all, it was bracketed with other delusional states, hallucinations and errors. Those, of course, have an abundant literature going back many centuries. The possibility of error in our perceptions has been responsible for the persistent (but surely erroneous) view that we never perceive reality, only sense-data, or only our idea of reality, or only a cognitive model of reality. The search for certainty in the face of the constant possibility of error motivated Descartes and arguably most of epistemology.

Clinically, delirium is an organically caused state of confusion. Philosophically, I suggest we should seize on another feature, namely that it can involve derangement of both perception and cognition. Let's use the special power of fiat used by philosophers to create new races of zombies, generate second earths, and enslave the population of China, and say that philosophical delirium is defined exactly as that particular conjunction of derangements. So we can then define three distinct kinds of mental disturbance. First, delusion, where our thinking mind is working fine but has bizarre perceptions presented to it. Second, madness, where our perceptions are fine, but our mental responses make no sense. Third, delirium, in which distorted perceptions meet with distorted cognition.

The question then is, can delirium, so defined, actually be distinguished from delusion and madness? Suppose we have a subject who persistently tries to eat their hat. One reading is that the subject perceives the Homburg as a hamburger. The second reading is that they perceive the hat correctly, but think it is appropriate to eat hats. The delirious reading might be that they see the hat as a shoe and believe shoes are to be eaten. For any possible set of behaviour, it seems different readings will achieve consistency with any of the three possible states.

That's from a third person point of view, of course, but surely the subject knows which state applies? They can't reliably tell us, because their utterances are open to multiple interpretations too, but inwardly they know, don't they? Well, no. The deluded person thinks the world really is bizarre; the mad one is presumably unaware of the madness, and the delirious subject is barred from knowing the true position on both counts. Does it then, make any sense to uphold the existence of any real distinction? Might we not better say that the three possibilities are really no more than rival diagnostic strategies, which may or may not work better in different cases, but have no absolute validity?

Can we perhaps fall back on consistency? Someone with delusions may see a convincing oasis out in the desert, but if a moment later it becomes a mountain, rational faculties will allow them

to notice that something is amiss and hypothesise that their sensory inputs are unreliable. However, a subject of Cartesian calibre would have to consider the possibility that they are actually just mistaken in their beliefs about their own experiences; in fact it always seemed to be a mountain. So once again the distinctions fall away.

Delusion and madness are all very well in their way, but delirium has a unique appeal in that it could be invisible. Suppose my perceptions are all subject to a consistent but complex form of distortion; but my responses have an exquisitely apposite complementary twist, which means that the two sets of errors cancel out and my actual behaviour and everything that I say, come out pretty much like those of some tediously sane and normal character. I am as delirious as can be, but you'd never know. Would I know? My mental states are so addled and my grip on reality so contorted, it hardly seems I could know anything; but if you question me about what I'm thinking, my responses all sound perfectly fine, just like those of my twin who doesn't have invisible delirium.

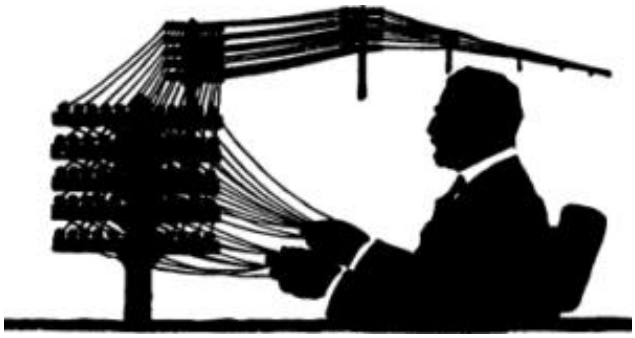
We might be tempted to say that invisible delirium is no delirium; my thoughts are determined by the functioning of my cognitive processes, and since those end up working fine, it makes no sense to believe in some inner place where things go all wrong for a while.

But what if I get super invisible delirium? In this wonderful syndrome, my inputs and outputs are mangled in complementary ways again, but by great good fortune the garbled version actually works faster and better than normal. Far from seeming confused, I now seem to understand stuff better and more deeply than before. After all, isn't reaching this kind of state why people spend time meditating and doing drugs?

But perhaps I am falling prey to the euphoric condition diagnosed by Hume…

On The Phone Or In The Phone?

5 March 2018



On Aeon, Karina Vold asks whether our smartphones have truly become part of us, and if so whether they deserve new legal protections. She quotes grisly examples of the authorities using a dead man's finger to try to activate fingerprint recognition on protected devices.

There are several parts to the argument here. One is derived fairly straightforwardly from the extended mind theory. According to this point of view, we are not simply our brains, nor even our bodies. When we use virtual reality devices, we may feel ourselves to be elsewhere; a computer can give us cognitive abilities that we can use naturally but would not have been available from our simple biological nervous system. Even in the case of simpler technologies we may feel we are extended. Driving, I sometimes think of the car as 'me' in at least a limited sense. If I feel my way with a stick, I feel the ground through the stick, rather than feeling the movement of the stick and making conscious inferences about the ground. Our mind goes out further than we might have thought.

We can probably accept that there is at least some truth in that outlook. But we should also note an important qualification, namely that these things are a matter of degree. A stick in my hand may temporarily become like an extension of my limbs, but it remains temporary and liminal. It never becomes a core part of me in the way that my frontal lobes are. The argument for an extended mind is for a looser and more ambivalent border to the self, not just a wider one.

The second part of the argument is that while the authorities can legitimately seize our property, our minds are legally protected. Vold cites the right to silence, as well as restrictions on the use of drugs and lie detectors. She also quotes a judge to the effect that we are secure in the sanctum of our minds anyway, because there simply isn't any way the authorities can intervene in there. They can control our behaviour, but not our thoughts.

One problem for me is that the ethical rationale for the right to remain silent is completely opaque to me. I have no idea what justifies our letting people remain silent in cases where they have information that is legitimately needed. A duty to disclose makes a lot more sense to me. Perhaps this principle is just a strongly reinforced protection against the possibility of torture, in that removing the right of the authorities to have the information at all cuts off at the root any right to use pain as a means of prising it out? If so, it seems too much to me.

I also think the distinction between the ability to control behaviour and the ability to control thoughts is less absolute than might appear. True, we cannot read or implant thoughts themselves. But then it's extremely difficult to control every action, too. The power of brainwashing techniques has often been overestimated, but the authorities can control information, use persuasive methods and even those forbidden drugs to get what they want. The Stoics, enjoying a bit of a revival in popularity these days, thought that in a broken world you could at least stay in control of your own mind; but it ain't necessarily so; if they really want to, they can make you love Big Brother.

Still, let's broadly accept that attempts at direct intervention in the mind are repugnant in ways that restraint of the body is not, and let's also accept that my smart phone can in some extended sense be regarded as part of my mind. Does it then follow that my phone needs new legal protections in order to preserve the integrity of my personal boundaries?

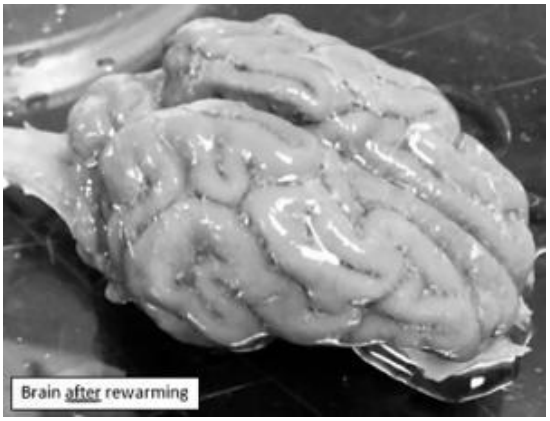
The word 'new' in there is the one that gives me the final problem. Mind extension is not a new thing; if sticks can be part of it, then it's nearly as old as humanity. Notebooks and encyclopaedias add to our minds and have been around for a long time. Virtual reality has a special power, but films and even oil paintings sort of do the same job. What's really new?

I think there is an implicit claim here that phones and other devices are special, because what they do is computation, and that's what your brain does too. So they become one with our minds in a way that nothing else does. I think that's just false. Regulars will know I don't think computation is the essence of thought anyway. But even if it were, the computations done in a phone are completely disconnected from those going on in the brain. Virtual reality may merge with our experience, but what it gives our mind is the outputs of the computation; we never experience the computations themselves. It may hypothetically be the case that future technology will do this, and genuinely merge our thoughts into the data of some advanced machine (I think not, of course); but the idea that we are already at that point and that in fact smartphones already do this is a radical overstatement.

So although existing law may well be improvable, I don't see a case in principle for any new protections.

Brain Preservation Prize

13 March 2018



The prize offered by the Brain Preservation Foundation has been won by 21st Century Medicine (21CM) with the Aldehyde-Stabilized Cryopreservation (ASC) technique that has been developed. In essence this combines chemical and cryogenic approaches and is apparently capable of preserving the whole connectome (or neural network) of a large mammalian brain (a pig brain here) in full detail and indefinitely. That is a remarkable achievement. A paper is here.

I am an advisor to the BPF, though I should make it clear that they don't pay me and I haven't given them a great deal of advice. I've always said I would be a critical friend, in that I doubt this research is ever going to lead to personal survival of the self whose brain is preserved. However, in my opinion it is much more realistic than a pure scan-and-upload approach, and has the potential to yield many interesting benefits even if it never yields personal immortality.

One great advantage of preserving the brain like this is that it defers some choices. When we model a brain or attempt to scan it into software, we have to pick out the features we think are salient and concentrate on those. Since we don't yet have any comprehensive and certain account of how the brain functions, we might easily be missing something essential. If we keep the whole of an actual brain, we don't have to make such detailed choices and have a better chance of preserving features whose importance we haven't yet recognised.

It's still possible that we might lose something essential, of course. ASC, not unreasonably, concentrates on preserving the connectome. I'm not sure whether, for example, it also keeps the brain's astrocytes in good condition, though I'd guess it probably does. These are the non-neural cells which have often been regarded as mere packing, but which may in fact have significant roles to play. Recently we've heard that neurons appear to signal with RNA packets; again, I don't know whether ASC preserves any information about that - though it might. But even on a pessimistic view, ASC must in my view be a far better preservation proposition than digital models that explicitly drop the detailed structure of individual neurons in favour of an unrealistically standardised model, and struggle with many other features.

Preserving brains in fine detail is a worthy project in itself, which might yield big benefits to research in due course. But of course the project embodies the hope that the contents of a mind and even the personality of an individual could be delivered to posterity. I do not think the contents of a mind are likely to be recoverable from a preserved brain yet awhile, but in the long run, why not? On identity, I am a believer in brute physical continuity. We are our brains, I believe (I wave my hands to indicate various caveats and qualifications which need not sideline us here). If we want to retain our personal identity, then, the actual physical preservation of the brain is essential.

Now, once your brain has been preserved by ASC, it really isn't going to be started up again in its existing physical form. The Foundation looks to uploading at this stage, but because I don't think digital uploading as we now envision it is possible in principle, I don't see that ever working.

However, there is a tiny chink of light at the end of that gloomy tunnel. My main problem is with the computational nature of uploading as currently envisaged. It is conceivable that the future will bring non-computational technologies which just might allow us to upload, not ourselves, but a kind of mental twin at least. That's a remote speculation, but still a fascinating prospect. Is it just conceivable that ways might be found to go that little bit further and deliver some kind of actual physical interaction between these hypothetical machines and the essential slivers of a preserved brain, some echo such that identity was preserved? Honestly, I think not, but I won't quite say it is inconceivable. You could say that in my view the huge advantage of the brain preservation strategy for achieving immortality is that unlike its rivals it falls just short of being impossible in principle.

So I suppose, to paraphrase Gimli the dwarf: certainty of death; microscopic chance of success - what are we waiting for?

Postscript: I meant by that last bit that we should continue research, but I see it is open to misinterpretation. I didn't dream people would actually do this, but I read that Robert McIntyre, lead author of the paper linked above, is floating a startup to apply the technique to people who are not yet dead. That would surely be unethical. If suicide were legal and if you had decided that was your preferred option, you might reasonably choose a method with a tiny chance of being revived in future. But I don't think you can ask people to pay for a technique (surely still inadequately tested and developed for human beings) where the prospects of revival are currently negligible and most likely will remain so.

Anthropic Consciousness

18 March 2018

Stephen Hawking's recent death caused many to glance regretfully at the unread copies of A Brief History Of Time on their bookshelves. I don't even own one, but I did read The Grand Design, written with Leonard Mlodinow, and discussed it earlier. It's a bold attempt to answer the big questions about why the universe even exists, and I suggested back then that it showed signs of an impatience for answers which is characteristic of scientists, at least as compared with philosophers. One sign of Hawking's impatience was his readiness to embrace a version of the rather dodgy Anthropic Principle as part of the foundations of his case.

In fact there are many flavours of the Anthropic Principle. The mild but relatively uninteresting version merely says we shouldn't be all that surprised about being here, because if we hadn't been here we wouldn't have been thinking about it at all. Is it an amazing piece of luck that from among all the millions of potential children our parents were capable of engendering, we were the ones who got born? In a way, yes, but then whoever did get born would have had the same perspective. In a similar way, it's not that surprising that the universe seems designed to accommodate human beings, because if it hadn't been that way, no-one would be worrying about it.

That's alright, but the stronger versions of the Principle make much more dubious claims, implying that our existence as observers really might have called the world into existence in some stronger sense. If I understood them correctly, Hawking and Mlodinow pitched their camp in this difficult territory.

Here at Conscious Entities we do sometimes glance at the cosmic questions, but our core subject is of course consciousness. So for us the natural question is, could there be an Anthropic-style explanation of consciousness? Well, we could certainly have a mild version of the argument, which would simply say that we shouldn't be surprised that consciousness exists, because if it didn't no-one would be thinking about it. That's fine but unsatisfying.

Is there a stronger version in which our conscious experience creates the preconditions for itself? I can think of one argument which is a bit like that. Let me begin by proposing an analogy in the supposed Problem of Focus.

The Problem of Focus notes that the human eye has the extraordinary power of drawing in beams of light from all the objects around it. Somehow every surface around us is impelled to send rays right in to that weirdly powerful metaphysical entity which resides in our eyes, the Focus. Some philosophers deny that there is a single Focus in each eye, suggesting it changes constantly. Some say the whole idea of a Focus with special powers is an illusion, a misconception of perfectly normal physical processes. Others respond that the facts of optometry and vision just show that denying the existence of Focus is in practice impossible; even the sceptics wear glasses!

I don't suppose anyone will be detained for long by worries about the Problem of Focus; but what if we remove the beams of light and substitute instead the power of intentionality, ie our

mental ability to think about things. Being able to focus on an item mentally is clearly a useful ability, allowing us to target our behaviour more effectively. We can think of intentionality as a system of pointers, or lines connecting us to the object being thought of. Lines, however, have two ends, so the back end of these ones must converge in a single point. Isn't it remarkable that this single focus point is able to draw together the contents of consciousness in a way which in fact generates that very state of awareness.

Alright, I'm no Hawking.

An AI Driving Test?

23 March 2018

The first case of a pedestrian death caused by a self-driving vehicle has provoked an understandably strong reaction. Are we witnessing the questioning of the Emperor's new clothes? Have we been living through the latest and greatest wave of hype around AI and its performance? Will self-driving cars come off the road for a generation, or even forever? Or, on the contrary, will we quickly come to accept that fatal accidents are unavoidable, as we have done for so long in the case of human drivers, after all?

It does seem to me that the dialogue around self-driving cars has been a bit unsatisfactory to date. There's been a surprising amount of discussion about the supposed ethical issues; should the car save the lives of the occupants if doing so involves killing a greater number of bystanders? I think some people have spent too long on trolley problems; these situations never really come up in practice, and 'try not to have an accident at all' is probably a perfectly adequate strategy.

More remarkable is the way the cars have been allowed to move on to public roads quite quickly. No doubt the desire of the relevant authorities in various places to stay ahead of the technological curve has something to do with this. But so far as I know there has been almost no sceptical examination of the technology by genuinely impartial observers. The designers have generally retained quite tight control, and by and large their claims have been accepted rather uncritically. I've pointed out before that the idea that self-driving machines are safer than humans sits oddly with the fact that current versions all use the human driver as the emergency fall-back. There may in fact have been some tendency to treat the human as a handy repository for all blame; if they don't intervene, they should have done; if they do intervene, then any accident is their responsibility because the AI was not in control at the moment of disaster.

Our confidence in autonomous vehicles is the stranger because it is quite well known that AI tends to have problems dealing with the unrestricted and unformalised domain of the real world. In the controlled environment of rail systems, AI works fine, and even in the more demanding case of aviation, autopilots have an excellent record - although a plane is out in the world, it normally only deals with predictable conditions and carefully designed runways. To a degree roads can be considered similarly standardised and predictable, of course, but not all roads and certainly not the human beings that frequent them.

It can be argued that AI does not need human levels of understanding to function well; machine translation now turns in a useful performance without even attempting to fathom what the words are about, after all. But even now it has a significant failure rate, and while an egregious mistranslation here and there probably does little harm, a driving mistake is another matter.

Do we therefore need a driving test for AI? Should autonomous vehicles be put through rigorous tests designed to exploit likely weaknesses and administered by neutral or even hostile examiners? I would have thought something like that would be a natural requirement.

The problem might be whether effective tests are possible. Human drivers have recognisable patterns of failure that can be addressed with more training. That may or may not be the case with AIs. I don't know how advanced the recognition technology being used actually is, but we know that some of the best available systems can behave in ways that are weirdly unpredictable to human beings, with a few pixels in an image exerting unexpected influence. It might be very difficult to test an AI in ways that give good assurance that performance will degrade, if at all, only gradually and manageably, rather than suddenly and catastrophically.

In this context, footage of the recent accident is disturbing. The car simply ploughs right into a clearly visible pedestrian wheeling a bike across the road. It's hard to see how this can be explained or how it can be consistent with a predictably safe system. I hope we're not just going to be told it was the fault of the human co-pilot (who reacted too slowly, but surely isn't supposed to have to be ready for an emergency stop at any moment?) - or worse, of the victim.

Bot Love

John Danaher has given a defence of robot love that might cause one to wonder for a moment whether he is fully human himself. People reject the idea of robot love because they say robots are merely programmed to deliver certain patterns of behaviour, he says. They claim that real love would require the robot to have feelings, and freedom of choice. But what are those things, even in the case of human beings? Surely patterns of behaviour are all we've got, he suggests, unless you're some nutty dualist. He quotes Michael Hauskeller...

> *[I]t is difficult to see what this love... should consist in, if not a certain kind of loving behaviour ... if [our lover's] behaviour toward us is unfailingly caring and loving, and respectful of our needs, then we would not really know what to make of the claim that they do not really love us at all, but only appear to do so.*

On the contrary, such claims are universally accepted and understood as part of normal human life. Literature and reality are full of situations where we suspect and fear that someone may not really love us in the way their behaviour would suggest - and indeed, cases where we hope in the teeth of all behavioural evidence that someone does love us. Such hopes are not meaninglessly incoherent.

It turns out, then, according to Danaher, that behaviourism is not bankrupt and outmoded. On the contrary, it is obviously true, and further, it is really the only way we have of talking about the mind at all. If there were any residual doubt about his position, he explains...

I have defended this view of human-robot relations under the label 'ethical behaviourism', which is a position that holds that the ultimate epistemic grounding for our beliefs about the value of relationships lies in the detectable behavioural and functional patterns of our partners, not in some deeper metaphysical truths about their existence.

The thing is, behaviourism failed because it became too clear that behavioural patterns are unintelligible or even unrecognisable except in the light of hypotheses about internal mental states. You cannot give a list of behavioural responses which correspond to love. Given the right context, pretty well any behaviour can be loving. Yes, the evidence for emotions is behavioural; but it grounds no beliefs about emotions without accompanying beliefs about internal mental states.

The robots we currently have do not have the required internal states, so they are not candidates to be considered loving; and in fact, they don't really produce convincing behaviour patterns of the right sort. Danaher is right that the lack of freedom or subjective experience look like fatal objections to robot love for most people, but myself I would rest most strongly on their lack of intentionality. Nothing means anything to our current, digital computer robots; they don't, in any meaningful sense, understand that anyone exists, much less have strong feelings about it.

At some points, Danaher seems to be talking about potential future robots rather than anything we already have (I'm beginning to wish philosophers could rein in their habit of getting their ideas about robots from science fiction films). Yes, it's conceivable that some new future technology might produce robots with genuine emotions; the human brain is, after all, a physical machine in some sense, albeit an inconceivably complex one. But before we can have a meaningful discussion about those future bots, we need to know how they are going to work. It can't just be magic.

Myself, I see no reason why people can't have sex bots that perform puppet-level love routines. If we mistake machines for people, we run the risk of being tempted to treat people like machines. But at the moment I don't really think anyone is being fooled, beyond the acknowledged capacity of human beings to fall in love with inanimate objects that display no behaviour at all. If we started to convince ourselves that we have no more mental life than they do, if somehow behaviourism came lurching zombie-like from the grave - well, then I might be worried!

Forgotten Crimes

3 April 2018

Should people be punished for crimes they don't remember committing? Helen Beebee asks this meaty question.

Why shouldn't they? I take the argument to have two main points. First, because they don't remember, they are no longer the same person as the one who committed the crime. Second, if you're not the person who committed the crime, you can't be responsible and therefore should not be punished. Both of these points can be challenged.

The idea that not remembering makes you a different person takes memory to be the key criterion of personal identity, a view associated with John Locke among others. But memory is in practice a very poor criterion. We remember unconsciously things we cannot recall explicitly; does unconscious memory count, and if so, how would we know? If I remember later, do I then become responsible for the crime? Is my unconscious self the same while my conscious one is not, so that I ought to suffer punishment I'm only aware of unconsciously? What if I have a false memory of the crime, but one that happens to be veridical, to tell the essential truth if inaccurately? Am I responsible? If so, then surely I ought to be responsible even if in fact I did not commit the crime, so long as I 'remember' doing it.

Moreover, aren't the practical consequences unacceptable? If forgetting the crime exculpates me, I can commit a murder and immediately take mnestic drugs that will make me forget it. It seems clear to me that few people think it really works as selectively as that. In order to be free of blame, you need to be a different person, and that implies losing more than a single memory. Perhaps it requires the loss of most memories, or more plausibly a loss of mentally retained things that are not exactly factual or experiential memory; my habits of thought, or the continuity of my personality. I think it's possible that Locke would say something like this if he were still around. So perhaps the case ought to be that if you do not remember the crime, and other features of your self have suffered an unusual discontinuity, such that you would no longer commit a similar crime in similar circumstances, then you are off the hook. How we could establish such a thing forensically is quite another matter, of course.

What about the second point, though? Does the fact that I am now a different, and also a better person, one who doesn't remember the crime, mean I shouldn't be punished? Not necessarily. Legally, for example, we might look to the doctrines of joint enterprise and strict liability to show that I can sometimes be held responsible in some degree for crimes I did not directly commit, and ones which I was powerless to prevent, if I am nevertheless associated with the crime in the required ways.

It partly depends on why we think people should be punished at all. Deterrence is a popular justification, but it does not require that I am really responsible. Being punished for a crime may well deter me and others from attempting similar crimes in future, even if I didn't do it at all, never mind cases where my responsibility is merely attenuated by loss of memory. The prevention of revenge is another justification that doesn't necessarily require me to have been

fully guilty. Or there might be doctrines of simple justice that hold to the idea of crime being followed by punishment, not because of any consequences that might follow, but just as a primary ethical principle. Under such a justification, it may not matter whether I am responsible in any strong sense. Oedipus did not know he had killed his father, and could not be held responsible for patricide; but he still put out his own eyes.', '', '', 'inherit', 'closed', 'closed', '', '3397-revision-v1', '', '',

‘2006-04-03 08:41:17',

‘2006-04-03 08:41:17', '', 3397, 'https://www.consciousentities.com/2018/04/3397-revision-v1/', 0, 'revision', '', 0), (3401, 1,

‘2006-04-03 08:59:36',

‘2006-04-03 08:59:36', 'Should people be punished for crimes they don’t remember committing? Helen Beebee asks this meaty question.

Why shouldn’t they? I take the argument to have two main points. First, because they don’t remember, it can be argued that they are no longer the same person as the one who committed the crime. Second, if you’re not the person who committed the crime, you can’t be responsible and therefore should not be punished. Both of these points can be challenged.

The idea that not remembering makes you a different person takes memory to be the key criterion of personal identity, a view associated with John Locke among others. But memory is in practice a very poor criterion. We remember unconsciously things we cannot recall explicitly; does unconscious memory count, and if so, how would we know? If I remember later, do I then become responsible for the crime? Is my unconscious self the same while my conscious one is not, so that I ought to suffer punishment I’m only aware of unconsciously? If I do not remember details, but have that sick sense of certainty that, yes, I did it alright, am I near enough the same person? What if I have a false, confabulated memory of the crime, but one that happens to be veridical, to tell the essential truth, if inaccurately? Am I responsible? If so, then surely I ought to be responsible even if in fact I did not commit the crime, so long as I ‘remember’ doing it?

Moreover, aren’t the practical consequences unacceptable? If forgetting the crime exculpates me, I can commit a murder and immediately take mnestic drugs that will make me forget it. It seems clear to me that few people think it really works as selectively as that. In order to be free of blame, you really need to be a different person, and that implies losing much more than a single memory. Perhaps it requires the loss of most memories, or more plausibly a loss of mentally retained things that are not exactly factual or experiential memory; my habits of thought, or the continuity of my personality. I think it’s possible that Locke would say something like this if he were still around. So perhaps the case ought to be that if you do not remember the crime, and other features of your self have suffered an unusual discontinuity, such that you would no longer commit a similar crime in similar circumstances, then you are off the hook. How we could establish such a thing forensically is quite another matter, of course.

What about the second point, though? Does the fact that I am now a different, and also a better person, one who doesn’t remember the crime, mean I shouldn’t be punished? Not necessarily. Legally, for example, we might look to the doctrines of joint enterprise and strict liability to show that I can sometimes be held responsible in some degree for crimes I did not directly commit, and ones which I was powerless to prevent, if I am nevertheless associated with the crime in the required ways.

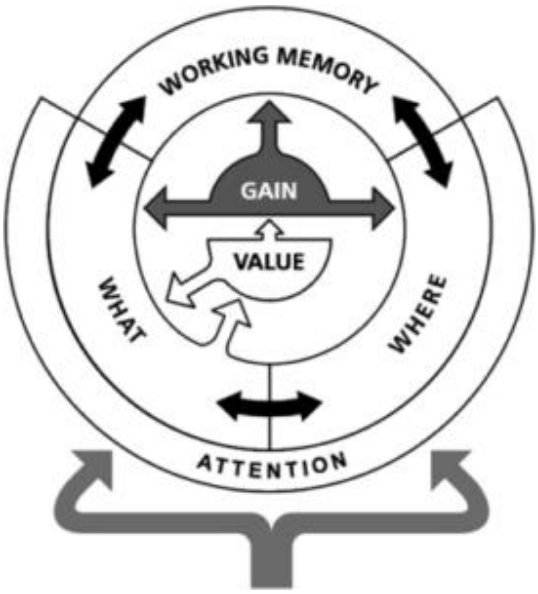
It partly depends on why we think people should be punished at all. Deterrence is a popular justification, but it does not require that I am really responsible. Being punished for a crime may well deter me and others from attempting similar crimes in future, even if I didn't do it at all, never mind cases where my responsibility is merely attenuated by loss of memory. The prevention of revenge is another justification that doesn't necessarily require me to have been fully guilty. Or there might be doctrines of simple justice that hold to the idea of crime being followed by punishment, not because of any consequences that might follow, but just as a primary ethical principle. Under such a justification, it may not matter whether I am responsible in any strong sense. Oedipus did not know he had killed his father, and could not be held responsible for patricide; but he still put out his own eyes.

Much more could be said, but for me the foregoing arguments are enough to suggest that memory is not really the point, either for responsibility or for personal identity. Beebee presents an argument about Bruce Banner and the Hulk; she feels Banner cannot directly be held responsible for the mayhem caused by the Hulk. Perhaps not, but surely the issue there is control, not memory. It's irrelevant whether Banner remembers what the Hulk did, all that matters is whether he could have prevented it. Beebee makes the case for a limited version of responsibility which applies if Banner can prevent the Hulk from appearing, but I think we have already moved beyond memory, so the fact that this special responsibility does not apply in the real life case she mentions is not decisive.

One point which I think should be added to the account, though it too is not decisive, is that the loss of responsibility may entail loss of personhood in a wider sense. If we hold that you are no longer the person who committed the crime, you are not entitled to their belongings or rights either. You are not married to their spouse, nor the heir to their parents. Moreover, if we think you are liable to turn into someone else again at some stage, and we know that criminals are, as it were, in your repertoire, we may feel justified in locking you up anyway; not as a punishment, but as prudent prevention. To avoid consequences like these, we may feel it is worth claiming our responsibility for certain past crimes, even if we no longer recall them.

Robot Memory

16 April 2018



A new model for robot memory raises some interesting issues. It's based on three networks, for identification, localization, and working memory. I have a lot of quibbles, but the overall direction looks promising.

The authors (Christian Balkenius, Trond A. Tjøstheim, Birger Johansson and Peter Gärdenfors) begin by boldly proposing four kinds of content for consciousness; emotions, sensations, perceptions (ie, interpreted sensations), and imaginations. They think that may be the order in which each kind of content appeared during evolution. Of course this could be challenged in various ways. The borderline between sensations and perceptions is fuzzy (I can imagine some arguing that there are no uninterpreted sensations in consciousness, and the degree of interpretation certainly varies greatly), and imagination here covers every kind of content which is about objects not present to the senses, especially the kind of foresight which enables planning. That's a lot of things to pack into one category. However, the structure is essentially very reasonable.

Imaginations and perceptions together make up an 'inner world'. The authors say this is ess3ntial for consciousness, though they seem to have also said that pure emotion is an early content of consciousness. They propose two tests often used on infants as indicators of such an inner world; tests of the sense of object permanence and 'A-not-B'. Both essentially test whether infants (or other cognitive entities) have an ongoing mental model of things which goes beyond what they can directly see. This requires a kind of memory to keep track of the objects that are no longer directly visible, and of their location. The aim of the article is to propose a system for robots that establishes this kind of memory-based inner world.

Imitating the human approach is an interesting and sensible strategy. One pitfall for those trying to build robot consciousness is the temptation to use the power of computers in non-human ways. We need our robot to do arithmetic: no problem! Computers can already do arithmetic much faster and more accurately than mere humans, so we just slap in a calculator module. But that isn't at all the way the human brain tackles explicit arithmetic, and by not following the human model you risk big problems later.

Much the same is true of memory. Computers can record data in huge quantities with great accuracy and very rapid recall; they are not prone to confabulation, false memories, or vagueness. Why not take advantage of that? But human memory is much less clear-cut; in fact 'memory' may be almost as much of a 'mongrel' term as consciousness, covering all sorts of abilities to repeat behaviour or summon different contents. I used to work in an office whose doors required a four-digit code. After a week or so we all tapped out each code without thinking, and if we had to tell someone what the digits were we would be obliged to mime the action in mid-air and work out which numbers on the keypad our fingers would have been hitting. In effect, we were using 'muscle memory' to recall a four-digit number.

The authors of the article want to produce the same kind of 'inner world' used in human thought to support foresight and novel combinations. (They seem to subscribe to an old theory that says new ideas can only be recombinations of things that got into the mind through the senses. We can imagine a gryphon that combines lion and eagle, but not a new beast whose parts resemble nothing we have ever seen. This is another point I would quibble over, but let it pass.)

In fact, the three networks proposed by the authors correspond plausibly with three brain regions; the ventral, dorsal, and prefrontal areas of the cortex. They go on to sketch how the three networks play their role and report tests that show appropriate responses in respect of object permanence and other features of conscious cognition. Interestingly, they suggest that daydreaming arises naturally within their model and can be seen as a function that just arises unavoidably out of the way the architecture works, rather than being something selected for by evolution.

I'm sometimes sceptical about the role of explicit modelling in conscious processes, as I think it is easily overstated. But I'm comfortable with what's being suggested here. There is more to be said about how processes like these, which in the first instance deal with concrete objects in the environment, can develop to handle more abstract entities; but you can't deal with everything in detail in a brief article, and I'm happy that there are very believable development paths that lead naturally to high levels of abstraction.

At the end of the day, we have to ask: is this really consciousness? Yes and no, I'm afraid. Early on in the piece we find:

> *On the first level, consciousness contains sensations. Our subjective world of experiences is full of them: tastes, smells, colors, itches, pains, sensations of cold, sounds, and so on. This is what philosophers of mind call qualia*.

Well, maybe not quite. Sensations, as usually understood, are objective parts of the physical world (though they may be ones with a unique subjective aspect), processes or events which are open to third-person investigation. Qualia are not. It is possible to simply identify qualia with sensations, but that is a reductive, sceptical view. Zombie twin, as I understand him, has sensations, but he does not have qualia.

So what we have here is not a discussion of 'Hard Problem' consciousness, and it doesn't help us in that regard. That's not a problem; if the sceptics are right, there's no Hard stuff to account for anyway; and even if the qualophiles are right, an account of the objective physical side of cognition is still a major achievement. As we've noted before, the 'Easy Problem' ain't easy.

Getting Emotional

23 April 2018

Are emotions an essential part of cognition? Luiz Pessoa thinks so (or perhaps he feels it is the case). He argues that while emotions have often been regarded as something quite separate from rational thought, a kind of optional extra which can be bolted on, but has no essential role, they are actually essential.

I don't find his examples very convincing. He says robots on a Mars mission might have to plan alternative routes and choose between priorities, but I don't really see why that requires them to get emotional. He says that if your car breaks down in the desert, you may need a quick fix. A sense of urgency will impel you to look for that, while a calm AI might waste time planning and implementing a proper repair. Well, I don't know. On the one hand, it seems perfectly possible to appreciate the urgency of the situation in a calm rational way. On the other, it's easy to imagine that panic and many other emotional responses might be very unhelpful, blocking the search for solutions or interfering with the ability to focus on implementation.

Yet there must be something in what he says, mustn't there? Otherwise, why would we have emotions? I suppose in principle it could be that they really have no role; that they are epiphenomenal, a kind of side effect of no real importance. But they seem to influence behaviour in ways that make that implausible.

Perhaps they add motivation? In the final analysis, pure reason gives us no reason to do anything. It can say, if you want A, then the best way to get it is through X, Y, and Z. But if you ask, should I want A, pure reason merely shrugs, or at best it says, you should if you want B.

However, it doesn't take much to provide a basic set of motivations. If we just assume that we want to survive, the need to obtain secure sources of food, shelter, and so on soon generate a whole web of subordinate motivations. Throw in a few very simple built-in drives - avoidance of pain, seeking sex, maintenance of good social relations - and we're pretty much there in terms of human motivation. Do we need complex and distracting emotions?

Some argue that emotions add more 'oomph', that they intensify action or responses to stimuli. I've never quite understood the actual causal process there, but granting the possibility, it seems emotions must either harmonise with rational problem solving, or conflict with it. Rational problem solving is surely always best, so they must either be irrelevant or harmful?

One fairly traditional view is that emotions are a legacy of evolution, a system that developed before rational problem solving was available. So different emotional states affect the speed and intensity of certain sets of responses. If you get angry, you become more ready to fight, which may be helpful. Now, we would be better off deciding rationally on our responses, but

we're lumbered with the system our ancestors evolved. Moreover, some of the preparatory stuff, like a more rapid heartbeat, has never come under rational control so without emotions it wouldn't be accessible. It can be argued that emotions are really little more than certain systems getting into certain potentially useful ready states like this.

That might work for anger, but I still don't understand how grief, say, is a useful state to be in. There's one more possible role for emotions, which is social co-ordination. Just as laughter or yawning tends to spread around the group, it can be argued that emotional displays help get everyone into a similar state of heightened or depressed responsiveness. But if that is truly useful, couldn't it be accomplished more easily and in less disabling/distracting ways? For human beings, talking seems to be the tool for the job?

It remains a fascinating question, but I've never heard of a practical problem in AI that really seemed to need an emotional response.

Why Would You Even Think That?

More support for the illusionist perspective in a paper from Daniel Shabasson. He agrees with Keith Frankish that phenomenal consciousness is an illusion, and (taking the metaproblematic road) offers a theory as to why so many people - the great majority, I think - find it undeniably real in spite of the problems it raises.

Shabasson's theory rests on three principles:

- impenetrability,
- the infallibility illusion, and
- the justification illusion.

Impenetrability says that we have no conscious access to the processes that produce our judgements about sensory experience. We know as a matter of optical/neurological science that our perception of colour rests on some very complex processing of the data detected by our eyes. Patches of paint or groups of pixels emitting exactly the same wavelengths of light may be perceived as quite different colours when our brains take account of the visual context, for example, but the resulting colours are just present to consciousness as facts. We have no awareness of the complex adjustments that have been made.

This point is particularly evident in the case of colour vision, where the processing done by the brain is elaborate and sometimes counter intuitive. It's less clear that we're missing out on much in the way of subtle interpretive processing when we detect a poke in the eye. Generally though, I think the claim is pretty uncontroversial, and in fact our limited access to what's really going on has been an important part of other theories such as Scott Bakker's Blind Brain.

Infallibilty says that we are prone to assume we cannot be wrong about certain aspects of our experience. Obviously most perceptions could be mistaken, but others, more direct, seem invulnerable to error. I may be mistaken in my belief that there is a piano on my foot, or about the fact that my toe is crushed; but surely I can't be wrong about the fact that I am feeling pain? Although this idea has been robustly challenged, it has a strong intuitive appeal, perhaps partly out of a feeling that while we can be wrong about external stuff, mental entities are perceived directly, already present in the mind, and therefore immune from the errors that creep in during delivery of external information.

Justification is a little more subtle; the claim is that for any judgement we make, we believe there is some justification. This is not the stronger claim that there is good or adequate justification, just the view that we suppose ourselves to have some reason for thinking whatever we think.

Is that true? What if I fix my thoughts on the fourth nearest star to Earth which has only one planet orbiting it, and judge that the planet in question is smaller than Earth? If I knew more about astronomy I might have reasons for this judgement, but as matters stand, though I feel confident that the planet exists, I have no reasons for any beliefs about its size relative to Earth.

In such a case, I believe Shabasson would either point to probable justifications I have overlooked (perhaps I am making a mistaken but not irrational assumption about a correlation between size and number of planets) or more likely, simply deny that I have truly made the

relevant judgement at all. I might assert that I really believe the planet is small, but I'm really only playing some hypothetical game. I think in fact, Shabasson can have what he needs for the sake of argument here pretty much by specification.

Given the three principles, various things follow. When we judge ourselves to be having a 'reddish' experience, we must be right, and there must be something in our mind that justifies the judgement. That thing is a quale, which must therefore exist. This follows so directly, without requiring effortful reasoning, it seems to us that we apprehend the quale directly. Furthermore, the quale must seem like something, or to put it more fully, there must be something it seems like: if there were nothing a quale were like, there would be no apparent difference between a red and a green quale; but it is of the essence that there are such differences.

What is it like? We can't say, because in fact it doesn't exist. Though there really are justificatory properties for our judgements about perceptions, they are not phenomenal ones; but impenetrability means we remain unaware of them. Hence the apparent ineffability of qualia. Impenetrability also gives rise to an impression that qualia are intrinsic; briefly it means that the reddish experience arrives with no other information, and in particular nothing about its relation to other things; it seems it just is. Completing the trio, qualia seem subjective because given ineffability and intrinsicality, they are only differentiable through introspection, and introspection naturally limits access to a particular single subject.

I don't think Shabasson has the whole answer (I think, in particular, that the apparent existence of qualia has to do with the particular reality of actual experience, a quality obviously not conveyed by any theoretical account), but I think there are probably several factors that account for our belief in phenomenal experience, and he gives a very clear account of some significant ones. The use of the principle of justification seems especially interesting to me; I wonder if it might help illuminate some other quirks of human psychology.

No Problem

21 May 2018

The older I get, the less impressed I am by the hardy perennial of free will versus determinism. It seems to me now like one of those completely specious arguments that the Sophists supposedly used to dumbfound their dimmer clients.

One of their arguments, apparently went like this. Your dog has had pups? And it belongs to you? Then it's a mother, and it's yours. Ergo, it's your mother!!!

If we can take this argument seriously enough to diagnose it, we might point out that 'your' is a word with several distinct uses. One is to pick out items that are your legal property; another is the wider one of picking out items that pertain to you in some other sense. We can, for example, use it to pick out the single human being that is your immediate female progenitor. So long as we are clear about these different senses, no problem arises.

Is free will like that? The argument goes something like this. Your actions were ultimately determined by the laws of physics. An action which was determined in advance is not free. Ergo, physics says none of your actions were free!!!

But there are two entirely different senses of "determined" in play here. When I ask if you had a free choice, I'm not asking a metaphysical question about whether an interruption to the causal sequence occurred. I'm asking whether you had a gun to your head, or something like that.

Now some might argue that although the two senses are distinct, the physics one over-rides the psychological one and renders it meaningless. But it doesn't. A metaphysical interruption to the causal sequence wouldn't give me freedom anyway; it might give me a random factor, but freedom is not random. What I want to know is, did your actions arise out of conscious thought, or did external factors constrain them? That's all. The undeniable fact that my actions are ultimately constrained by the laws of nature simply isn't what I'm concerned with.

That constraint really is undeniable, of course; in fact we don't really need physics. If the world is coherent at all it must be governed by laws, and those laws must determine what happens. If things happened for no reason, we could make no sense of anything. So any comprehensive world view must give us determinism. We know this, because we are familiar with at least one other comprehensive theory; the view that things happen only because God wills them. This means everything is predestined, and that gives rise to just the same sort of pseudo-problems over free will. In fact, if we want we can get the same problems from logical fatalism, without appealing to either science or theology. Will I choose A tomorrow or not? Necessarily there is a truth of the matter already, so although we cannot know, my decision is already a matter of fact.

So fundamental determinism is rock solid; it just isn't a problem for freedom.

Hold on, you may say; you frame this as being about external constraints, but the real question is, am I not constrained internally? Don't my own mental processes force me to make a particular decision? There are two versions of this argument. The first says that the mere fact

that mental processes operate mechanistically means there can be no freedom. I just deny that; my own conscious processes count as free no matter how mechanistic they may be.

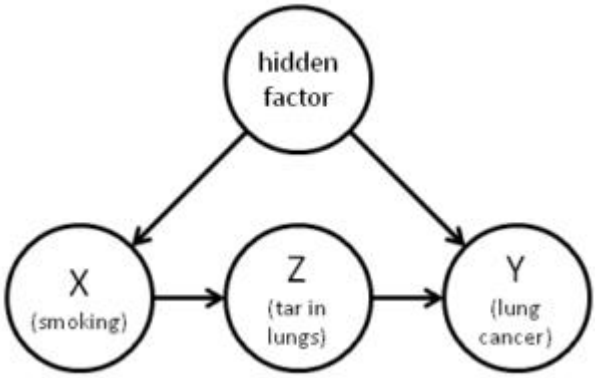
The second version of the argument says that while free decisions of that kind might be possible in pure theory, as an empirical matter human beings don't have the capacity for them. No conscious processes are actually effectual; consciousness is an epiphenomenon and merely invents rationales for decisions taken in a predetermined manner elsewhere in the brain. This argument is appealing because there is, of course, lots of evidence that unconscious factors influence our decisions. But the strong claim that my conscious deliberations are always irrelevant seems wildly implausible to me. Speech acts are acts, so to truly believe this strong version of the theory I'd have to accept that what I think is irrelevant to what I say, or that I adjust my thoughts retrospectively to fit whatever just came out of my mouth.

Now I may also be attacked from the other side. There may be advocates of free will who say, hold on, Peter, we do actually want that special metaphysical interruption you're throwing away so lightly. Introspect, dear boy, and notice how your decisions come from nowhere; that over and above the weighing of advantage there just is that little element of inexplicable volition.

This impression comes, I think, from the remarkable power of intentionality. We can think about anything at all, including future or even imaginary contingencies. If our actions are caused by things that haven't happened yet, or by things that will never actually happen (think of buying insurance) that looks like a mysterious disruption of the natural order of cause and effect. But of course it isn't really. You may have a different explanation depending on your view of intentionality; mine is briefly that it's all about recognition. Our ability to recognise entities that extend into the future, and then recognise within them elements that don't yet exist, gives us the ability to make plans for the future, for example, without any contradiction of causality.

I'm afraid I've ended up by making it sound complicated again. Let me wrap it up in time-honoured philosophical style*: it all depends what you mean by "determined"*.

Probably Right



Judea Pearl says that AI needs to learn causality. Current approaches, even the fashionable machine learning techniques, summarise and transform, but do not interpret the data fed to them. They are not really all that different from techniques that have been in use since the early days.

What is he on about? About twenty years ago, Pearl developed methods of causal analysis using Bayesian networks. These have yet to achieve the general recognition they seem to deserve (I'm afraid they're new to me). One reason Pearl's calculus has probably not achieved wider fame is the sheer difficulty of understanding it. A rigorous treatment involves a lot of equations that are challenging for the layman, and the rationale is not very intuitive at some points, even to those who are comfortable with the equations. The models have a prestidigitatory quality (Hey presto!) of seeming to bring solid conclusions out of nothing (but then much of Bayesian probability has a bit of that feeling for me). Pearl has now published a new book which tries to make all this accessible to the layman.

Difficult as they may be, his methods seem to have implications that are wide and deep. In science, they mean that randomised control testing is no longer the only game in town. They provide formal methods for tackling the old problem of distinguishing between correlation and causation, and they allow the quantification of probabilities in counterfactual cases. Do they provide new answers to Hume's questions about causality?

Pearl suggests that Hume, covertly or perhaps unconsciously, had two definitions of causality; one is the good old constant conjunction we know and love (approximately, A caused B because when A happens B always happens afterwards), the other in terms of counterfactuals (we can see that if A had not happened, B would not have happened either). Pearl lines up with David Lewis in suggesting that the counterfactual route is actually the way to go, with his techniques offering new formal techniques for dealing with counterfactuals. He thinks it's a reasonable speculation that the brain might be structured in ways that enables it to use similar techniques, but neither this nor the details of how exactly his approach wraps up the philosophical issues is set out fully. That's fair enough - we can't expect him to solve everyone else's problems as well as the pretty considerable ones he does deal with - but it would be good to see professional philosophical treatment (maybe there is one I haven't come across?) My hot take is that this doesn't altogether remove the difficulties; Pearl's approach still seems to rely on our ability to frame reasonable hypotheses and make plausible assumptions - but I'm far from sure. It looks to me as if this is a subject philosophers writing about causation or counterfactuals now need to understand, rather the way philosophers writing about metaphysics really ought to understand relativity and quantum physics (as they all do, of course).

What about AI? Is he right? I think he is, up to a point. There is a large problem which has consistently blocked the way to Artificial General Intelligence, to do with the computational intractability of undefined or indefinite domains. The real world, to put it another way, is just too complicated. This problem has shown itself in several different areas in different guises. I think such matters as the Frame Problem (in its widest form), intentionality/meaning, relevance, and radical translation are all places where the same underlying problem shows up, and it is

plausible to me that causality is another. In real world situations, there is always another causal story that fits the facts, and however absurd some of them may be, an algorithmic approach gets overwhelmed or fails to achieve traction.

So while people studying pragmatics have got the beast's tail, computer scientists have got one leg, Quine had a hand on its flank, and so on. Pearl. maybe, has got its neck. What AI is missing is the underlying ability that allows human beings to deal with this stuff (IMO the human faculty of recognition). If robots had that, they would indeed be able to deal with causality and much else besides. The frustrating possibility here is that Pearl's grasp on the neck actually gives him a real chance to capture the beast, or in other words that his approach to counterfactuals may contain the essential clues that point to a general solution. Without a better intuitive understanding of what he says, I can't be sure that isn't the case.

So I'd better read his book, as I should no doubt have done before posting, but you know…

Flat Thinking

5 June 2018

Nick Chater says The Mind Is Flat in his recent book of that name. It's an interesting read which quotes a good deal of challenging research (although quite a lot is stuff that I imagine would be familiar to most regular readers here). But I don't think he establishes his conclusion very convincingly. Part of the problem is a slight vagueness about what 'flatness' really means - it seems to mean a few different things and at times he happily accepts things that seem to me to concede some degree of depth. More seriously, the arguments range from pretty good through dubious to some places where he seems to be shooting himself in the foot.

What is flatness? According to Chater the mind is more or less just winging it all the time, supplying quick interpretations of the sensory data coming in at that moment (which by the way is very restricted) but having no consistent inner core: no subconscious, no self, and no consistent views. We think we have thoughts and feelings, but that's a delusion; the thoughts are just the chatter of the interpreter, and the feelings are just our interpretation of our own physiological symptoms.

Of course, there is a great deal of evidence that the brain confabulates a great deal more than we realise, making up reasons for our behaviour after the fact. Chater quotes a number of striking experiments, including the famous ones on split-brain patients, and tells surprising stories about the power of inattentional blindness. But it's a big leap from acknowledging that we sometimes extemporise dramatically to the conclusion that there is never a consistent underlying score for the tune we are humming. Chater says that if Anna Karenina were real, her political views would probably be no more settled than those of the fictional character, about whom there are no facts beyond what her author imagined. I sort of doubt that; many people seem to me to have relatively consistent views, and even where the views flip around they're not nugatory. Chater quotes remarkable experiments on this, including one in which subjects asked political questions via a screen with a US flag displayed in the corner gave significantly more Republican answers than those who answered the same questions without the flag - and moreover voted more Republican eight months later. Chater acknowledges that it seems implausible that this one experiment could have conditioned views in a way that none of the subjects' later experiences could do (though he doesn't seem to notice that being able to hold to the same conditioning for eight months rather contradicts his main thesis about consistency); but in the end he sort of thinks it did. These days we are more cautious about the interpretation of psychological experiments than we used to be, and the most parsimonious explanation might be something wrong with the experiment. An hypothesis Chater doesn't consider is that subjects are very prone to guessing the experimenter's preferences and trying to provide what's wanted. It could plausibly be the case that subjects whose questions were accompanied by patriotic symbols inferred that more right-wing answers would be palatable here and thought the same when asked about their vote much later by the same experimenter (irrespective of how they actually voted - something we can't in fact know, given the secrecy of the ballot).

Chater presents a lot of good evidence that our visual system uses only a tiny trickle of evidence; as little as one shape and one colour at a time, it seems. He thinks this shows that we're not having a rich experience of the world; we can't be, because the data isn't there. But isn't this a perverse interpretation? He accepts that we have the impression of a rich experience; given the paucity of data, which he evidences well, this impression can surely only come from internal processing and internal models - which looks like mental depth after all. Chater, I think, would argue that unconscious processing does take place but doesn't count as depth; but it's hard to see why not. In a similar way he accepts that we remember our old interpretations and feed them into current interpretations, even extrapolating new views, but this process, which looks a bit like reflection and internal debate, does not count as depth either. Here, it begins to look as if Chater's real views are more commonsensical than he will allow.

But not everywhere. On feelings, Chater is in a tradition stretching back to William James, who held that our hair doesn't stand on end because we're feeling fear; rather, we feel fear because we're experiencing hair-rise (along with feelings in the gut, goose bumps, and other physiological symptoms). The conscious experience comes after the bodily reaction, not before. Similar views were held by the behaviourists, of course; these reductive ideas are congenial because they mean emotions can be studied objectively from outside, without resort to introspection. But they are not very plausible. Again, we can accept that it sometimes happens that way. If stomach ache makes me angry, I may well go on to find other things to be angry about. If I go out at night and feel myself tremble, I may well decide it is because I am frightened. But if someone tells a ghost story, it is not credible that the fear doesn't come from my conscious grasp of the narrative.

I think Chater's argument for the non-existence of the self is perhaps his most alarming. It rests on a principle (or a dogma; he seems to take it as more or less self evident) that there is nothing in consciousness but the interpretation of sensory inputs. He qualifies this at once by allowing dreams and imagination, a qualification which would seem to give back almost everything the principle took away, if we took it seriously; but Chater really does mean to restrict us to thinking about entities we can see, touch or otherwise sense. He says we have no conscious knowledge of abstractions, not even such everyday ones as the number five. The best we can do is proxies such as an image of five dots, or of the numeral '5'. But I don't think that will wash. A collection of images is not the same as the number five; in fact, without understanding what five is, we wouldn't be able to pick out which groups of dots belonged to the collection. Chater says we rely on precedent, not principle, but precedents are useless without the interpretive principles that tell us when they apply. I don't know how Chater, on his own account, is even aware of such things as the number five; he refuses to address the metaphysical issues beyond his own assertions.

I think Chater's principle rules out arithmetic, let alone higher maths, and a good deal besides, but he presumably thinks we can get by somehow with the dots. Later, however, he invokes the principle again to dismiss the self. Because we have no sensory impressions of the self, it must be incoherent nonsense. But there are proxies for the self - my face in the mirror, the sound of my voice, my signature on a document - that seem just as good as the dots and numerals we've got for maths. Consistency surely requires that Chater either accepts the self or dumps mathematics.

As a side comment, it's interesting that Chater introduces his principle early and only applies it to the self much later, when he might hope we have forgotten the qualifications he entered, and

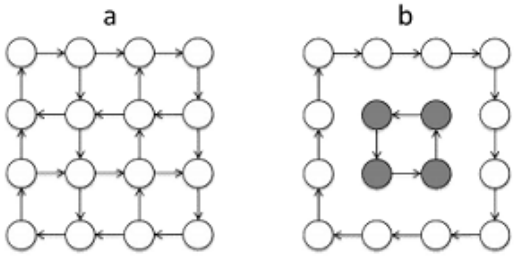
the issues over numbers. I do not suggest these are deliberate presentational tactics, rather they seem good evidence of how we often choose the most telling way of giving an argument unconsciously, something that is of course impossible in his system.

I'm much happier with Chater's view of AI. Early on, he gives a brief account of the failure of the naive physics project, which attempted to formalise our 'folk' understanding of science. He seems to conflate this with the much wider project of artificial general intelligence, but he is right about the limitations it points to. He thinks computers lack the 'elasticity' of human thought and are unlikely to acquire it any time soon.

A bit of a curate's egg, then. A lot of decent, interesting science, with some alarming stuff that seems philosophically naive (a charge I hesitate to make because it is always possible to adduce sophisticated justifications for philosophical positions that seem daft on the face of it; indeed, that's something philosophers particularly enjoy doing).

Alters Of The Cosmos

We are the alternate personalities of a cosmos suffering from Dissociative Identity Disorder (DID). That's the theory put forward by Bernardo Kastrup in a recent JCS paper and supported by others in Scientific American. I think there's no denying the exciting elegance of the basic proposition, but in my view the problems are overwhelming.



A connected graph (a) illustrating normal integration of phenomenal contents, and a disconnected graph (b) illustrating dissociation and the corresponding formation of an alter (inner subgraph in grey).

DID is now the correct term for what used to be called Multiple Personality Disorder, a condition in which different persons appear to inhabit the same body, with control passing between them and allowing them to exhibit distinct personalities, different knowledge, and varied behaviour. Occasionally it has been claimed that different 'alters' can even change certain physical characteristics of the host body, within limits. Sceptical analysis notes that the incidence of DID has been strongly correlated with its portrayal in the media. A popular film about multiple personalities always seems to bring a boom in new diagnoses, and in fact an early 'outbreak' corresponded with the popularity of 'Jekyll and Hyde'. Sceptics have suggested that DID may often, or always, be iatrogenic in part, with the patient confabulating the number and type of alter the therapist seems to expect.

Against that, the SA piece cites findings that when blind alters were in control, normal visual activity in the brain ceased. This is undoubtedly striking, though a caveat should be entered over our limited ability to spot what patterns of brain activity go along with confabulation, hypnosis, self-deception, etc. I think the research cited establishes pretty clearly that DID is 'real' (though not that patients correctly understand its nature), but then I believe only the hardest of sceptics ever thought DID patients were merely weird liars.

Does DID have the metaphysical significance Kastrup would give it, though? One fundamental problem, to get it up front, is this; if we, as physical human beings, are generated by DID in the cosmic consciousness, and that DID is literally the same thing as the DID observed in patients, how come it doesn't generate a new body for each of the patient's alters? There doesn't seem to be a clear answer on this. I would say that the most reasonable response would be to deny that cosmic and personal DID are exactly the same phenomena and regard them as merely analogous, albeit perhaps strongly so.

Kastrup's account does tackle a lot of problems. He approaches his thesis by considering related approaches such as panpsychism or cosmopsychism, and the objections to them, notably the combination or decombination problems, which concern how we get from millions of tiny awarenesses, or from one overarching one, to the array of human and animal ones we actually find in the world. His account seems clear and sensible to me, providing convincing brief analyses of the issues.

In Kastrup's system we begin with a universal consciousness which consists of a sort of web of connected thoughts and feelings. Later there will be perceptions, but at the outset there's nothing to perceive; I'm not sure what the thoughts could be about, either - pure maths, perhaps - but they arise from the inherent tendency of the cosmic consciousness to self-excite (just as a normal human mind, left without external stimulus, does not fall silent, but generates thoughts spontaneously). The connections between the thoughts may be associations, logical

connections, inspirational, and so on. I'm not clear whether Kastrup envisages all these thoughts and feelings being active at the same time, or whether new ones can be generated and added in. There is a vast amount of metaphysical work to be done on this kind of aspect of the theory - enough for several generations of philosophers - and it may not be fair to expect Kastrup to have done it all, let alone get it all into this single paper.

I think the natural and parsimonious way to go from there would be solipsism. The cosmic consciousness is all there is, and these ideas about other people and external reality are just part of its random musings. The only argument against this simple position is that our experience insistently and pretty consistently tells us about a world of planets, animals, and evolution which not only forces itself on our attention, but on examination provides some rather good partial explanations of our nature and cognitive abilities. But to accept that argument is to surrender to the conventional view, which Kastrup - he identifies as an idealist - is committed to rejecting.

So instead he takes a different view. Somehow (?), islands of the overall web of cosmic consciousness may get detached. They then become dissociated consciousnesses and can both perceive and be perceive. Since their associative links with the rest of the cosmos have been broken, I don't quite know why they don't lapse into solipsistic beings themselves, unable to follow the pattern of their thoughts beyond its own compass.

In fact, and this may be the strangest thing in the theory, our actual bodies, complete with metabolism and all the rest, are the appearance of these metaphysical islands: 'living organisms are the revealed appearance of alters of universal consciousness'. Quite why the alters of universal consciousness should look like evolved animals, I don't know. How does sex between these alters give rise to a new dissociative island in the form of a new human being; what on earth happens when someone starves to death? It seems that Kastrup really wants to have much of the conventional world back; a place where autonomous individuals with private thoughts are nevertheless able to share ideas about a world which is not just the product of their imaginations. But it's forbiddingly difficult to get there from his starting position. For once, weirder ideas might be easier to justify.

These are, of course, radical new ideas; but curiously they seem to me to bear a strong resemblance to the old ones of the Gnostics. They (if my recollection is right) thought that the world started with the perfect mind of God, which then through some inscrutable accident shed fragmentary souls (us) which became bound in the material world, with their own true nature hidden from them. I don't make the comparison to discredit Kastrup's ideas; on the contrary if it were me I should be rather encouraged to have these ancient intellectual forebears.'

The Mark Of The Mental

16 July 2018

Guillaume Fréchette wrote an interesting review of Mark Textor's new book *Brentano's Mind*. Naturally this deals among other things with the claim for which Brentano is most famous; that intentionality is the distinctive feature of the mental (and so of thoughts, consciousness, awareness, and so on). Textor apparently makes five major claims, but I only want to glance at the first one, that in fact 'Intentionality is an implausible mark of the mental'.

What was Brentano on about, anyway? Intentionality is the property of pointing at, or meaning, or being about, something. It was discussed in medieval philosophy and then made current again by Brentano when he, like others, was trying to establish an empirical science of psychology for the first time. In his view:

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

Textor apparently thinks that there's a danger of infinite regress here. He reads Brentano's mention of in-existence as meaning we need to think of an immanent object 'existing in' our mind in order to think of an object 'out there'; but in that case, doesn't thinking of the immanent object require a further immanent object, and so on. There seems to be more than one way of escaping this regress, however. Perhaps we don't need to think of the immanent object, it just has to be there. Maybe awareness of an external object and introspecting an internal one are quite different processes, the latter not requiring an immanent object. Perhaps the immanent object is really a memory, or perhaps the whole business of immanent objects reads more into Brentano than we should.

Textor believes Brentano is pushed towards primitivism - hey, this just is the mark of the mental, full stop - and thinks it's possible to do better. I think this is nearly right, except it assumes Brentano must be offering a theory, even if it's only the bankrupt one of primitivism. I think Brentano observes that intentionality is the mark of the mental, and shrugs. The shrug is not a primitivist thesis, it just expresses incomprehension. To say that one does not know x is not to say that x is unknowable. I could of course be wrong, both about Brentano, and particularly about Textor.

What I think you have to do is go back and ask why Brentano thought intentionality was the mark of the mental in the first place. I think it's a sort-of empirical observation. All thoughts are about, or of, something. If we try to imagine a thought that isn't about anything, we run into difficulty. Is there a difference between not thinking of anything and not thinking at all (thinking of nothing may be a different matter)? Similarly, can there be awareness which isn't awareness of anything? One can feel vast possible disputes about this opening up even as we speak, but I should say it is at least pretty plausible that all mental states feature intentionality.

Physical objects, such as stones, are not about anything; though they can be, like books, if we have used the original intentionality of our minds to bestow meaning on them; if in fact we intend them to mean something. Once again, this is disputable, but not, to me, implausible.

Intentionality remains a crucially important aspect of the mind, not least because we have got almost nowhere with understanding it. Philosophically there are of course plenty of contenders; ideas about how to build intentionality out of information, out of evolution, or indeed to show how original intentionality is a bogus idea in the first place. To me, though, it's telling that we've got nowhere with replicating it. Where AI would seem to require some ability to handle meaning - in translation, for example - it has to be avoided and a different route taken. While it remains mysterious, there will always be a rather large hole in our theories of consciousness.

Meh-Bots

20 August 2018

Do robots care? Aeon has an edited version of the inaugural Margaret Boden Lecture, delivered by Boden herself. You can see the full lecture above. Among other things, she tells us that the robots are not going to take over because they don't care. No computer has actual motives, the way human beings do, and they are indifferent to what happens (if we can even speak of indifference in a case where no desire or aversion is possible).

No doubt Boden is right; it's surely true at least that no current computer has anything that's really the same as human motivation. For me, though, she doesn't provide a convincing account of why human motives are special, and why computers can't have them, and perhaps doesn't sufficiently engage with the possibility that robots might take over the world (or at least, do various bad out-of-control things) without having human motives, or caring what happens in the fullest sense. We know already that learning systems set goals by humans are prone to finding cheats or expedients never envisaged by the people who set up the task; while it seems a bit of a stretch to suppose that a supercomputer might enslave all humanity in pursuit of its goal of filling the world with paperclips (about which, however, it doesn't really care), it seems quite possible real systems might do some dangerous things. Might a self-driving car (have things gone a bit quiet on that front, by the way?) decide that its built-in goal of not colliding with other vehicles can be pursued effectively by forcing everyone else off the road?

What is the ultimate source of human motivation? There are two plausible candidates that Boden doesn't mention. One is qualia; I think John Searle might say, for example, that it's things like the quake of hunger, how hungriness really feels, that are the roots of human desire. That nicely explains why computers can't have them, but for me the old dilemma looms. If qualia are part of the causal account, then they must be naturalisable and in principle available to machines. If they aren't part of the causal story, how do they influence human behaviour?

Less philosophically, many people would trace human motives to the evolutionary imperatives of survival and reproduction. There must be some truth in that, but isn't there also something special about human motivation, something detached from the struggle to live?

Boden seems to rest largely on social factors, which computers, as non-social beings, cannot share in. No doubt social factors are highly important in shaping and transmitting motivation, but what about Baby Crusoe, who somehow grew up with no social contact? His mental state may be odd, but would we say he has no more motives than a computer? Then again, why can't computers be social, either by interacting with each other, or by joining in human society? It seems they might talk to human beings, and if we disallow that as not really social, we are in clear danger of begging the question.

For me the special, detached quality of human motivation arises from our capacity to imagine and foresee. We can randomly or speculatively envisage future states, decide we like or detest them, and plot a course accordingly, coming up with motives that don't grow out of current circumstances. That capacity depends on the intentionality or aboutness of consciousness, which computers entirely lack - at least for now.

But that isn’t quite what Boden is talking about, I think; she means something in our emotional nature. That - human emotions - is a deep and difficult matter on which much might be said; but at the moment I can’t really be bothered…

Rosehip Neurons

28 August 2018

A new type of neuron is a remarkable discovery; finding one in the human cortex makes it particularly interesting, and the further fact that it cannot be found in mouse brains and might well turn out to be uniquely human - that is legitimately amazing. A paper (preprint here) in Nature Neuroscience announces the discovery of 'rosehip' neurons , named for their large, "rosehip"-like axonal boutons.

There has already been some speculation that rosehip neurons might have a special role in distinctive aspects of human cognition, especially human consciousness, but at this stage no-one really has much idea. Rosehip neurons are inhibitory, but inhibiting other neurons is often a key function which could easily play a big role in consciousness. Most of the traffic between the two hemispheres of the human brain is inhibitory, for example, possibly a matter of the right hemisphere, with its broader view, regularly 'waking up' the left out of excessively focused activity.

We probably shouldn't, at any rate, expect an immediate explanatory breakthrough. One comparison which may help to set the context is the case of spindle neurons. First identified in 1929, these are a notable feature of the human cortex and at first appeared to occur only in the great apes - they, or closely analogous neurons, have since been spotted in a few other animals with large brains, such as elephants and dolphins. I believe we still really don't know why they're there or what their exact role is, though a good guess seems to be that it might be something to do with making larger brains work efficiently.

Another warning against over-optimism might come from remembering the immense excitement about mirror neurons some years ago. Their response to a given activity both when performed by the subject and when observed being performed by others, seemed to some to hold out a possible key to empathy, theory of mind, and even more. Alas, to date that hope hasn't come to anything much, and in retrospect it looks as if rather too much significance was read into the discovery.

The distinctive presence of rosehip neurons is definitely a blow to the usefulness of rodents as experimental animals for the exploration of the human brain, and it's another item to add to the list of things that brain simulators probably ought to be taking into account, if only we could work out how. That touches on what might be the most basic explanatory difficulty here, namely that you cannot work out the significance of a new component in a machine whose workings you don't really understand to begin with.

There might indeed be a deeper suspicion that a new kind of neuron is simply the wrong kind of thing to explain consciousness. We've learnt in recent years that the complexity of a single neuron is very much not to be under-rated; they are certainly more than the simple switching devices they have at times been portrayed as, and they may carry out quite complex processing. But even so, there is surely a limit to how much clarification of phenomenology we can expect a single cell to yield, in the absence of the kind of wider functional theory we still don't really have.

Yet what better pointer to such a wider functional theory could we have than an item unique to humans with a role which we can hope to clarify through empirical investigation? Reverse

engineering is a tricky skill, but if we can ask ourselves the right questions maybe that longed-for ‘Aha!’ moment is coming closer after all?

The Map of Feelings

3 September 2018

An intriguing study by Nummenmaa et al (paper here) offers us a new map of human feelings, which it groups into five main areas; positive emotions, negative emotions, cognitive operations, homeostatic functions, and sensations of illness. The hundred feelings used to map the territory are all associated with physical regions of the human body.

The map itself is mostly rather interesting and the five groups seem to make broad sense, though a superficial look also reveals a few apparent oddities. 'Wanting' here is close to 'orgasm'. For some years now I've wanted to clarify the nature of consciousness; writing this blog has been fun, but dear reader, never quite like that. I suppose 'wanting' is being read as mainly a matter of biological appetites, but the desire and its fulfilment still seem pretty distinct to me, even on that reading.

Generally, a list of methodological worries come to mind, many of which are connected with the notorious difficulties of introspective research. 'Feelings' is a rather vaguely inclusive word, to begin with. There are a number of different approaches to classifying the emotions already available, but I have not previously encountered an attempt to go wider for a comprehensive coverage of every kind of feeling. It seems natural to worry that 'feelings' in this broad sense might in fact be a heterogeneous grouping, more like several distinct areas bolted together by an accident of language; it certainly feels strange to see thinking and urination, say, presented as members of the same extended family. But why not?

The research seems to rest mainly on responses from a group of more than 1000 subjects, though the paper also mentions drawing on the NeuroSynth meta-analysis database in order to look at neural similarity. The study imported some assumptions by using a list of 100 feelings, and by using four hypothesized basic dimensions - mental experience, bodily sensation, emotion, and controllability. It's possible that some of the final structure of the map reflects these assumptions to a degree. But it's legitimate to put forward hypotheses, and that perhaps need not worry us too much so long as the results seem consistent and illuminating. I'm a little less comfortable with the notion here of 'similarity'; subjects were asked to put feelings closer the more similar they felt them to be, in two dimensions. I suspect that similarity could be read in various ways, and the results might be very vulnerable to priming and contextual effects.

Probably the least palatable aspect, though, is the part of the study relating feelings to body regions. Respondents were asked to say where they felt each of the feelings, with 'nowhere', 'out there' or 'in Platonic space' not being admissible responses. No surprises about where urination was felt, nor, I suppose, about the fact that the cognitive stuff was considered to be all in the head. But the idea that thinking is simply a brain function is philosophically controversial, under attack from, among others, those who say 'meaning ain't in the head', those who champion the extended brain (if you're counting on your fingers, are you still thinking with just your brain?), those who warn us against the 'mereological fallacy', and those like our old friend Riccardo Manzotti, who keeps trying to get us to understand that consciousness is external.

Of course it depends what kind of claim these results might be intended to ground. As a study of ‘folk’ psychology, they would be unobjectionable, but we are bound to suspect that they might be called in support of a reductive theory. The reductive idea that feelings are ultimately nothing more than bodily sensations is a respectable one with a pedigree going back to William James and beyond; but in the context of claims like that a study that simply asks subjects to mark up on a diagram of the body where feelings happen is begging some questions.'

What's Wrong with Dualism

10 September 2018

I had an email exchange with Philip Calcott recently about dualism; here's an edited version. (Favouring my bits of the dialogue, of course!)

> *Philip: The main issue that puzzles me regarding consciousness is why most people in the field are so wedded to physicalism, and why substance dualism is so out of favour. It seems to me that there is indeed a huge explanatory gap – how can any physical process explain this extraordinary (and completely unexpected on physicalism) "thing" that is conscious experience?*
>
> *It seems to me that there are three sorts of gaps in our knowledge:*
>
> *1. I don't know the answer to that, but others do. Just let me just google it (the exact height of Everest might be an example)*
>
> *2. No one yet knows the answer to that, but we have a path towards finding the answer, and we are confident that we will discover the answer, and that this answer lies within the realm of physics (the mechanism behind high temperature superconductivity might be an example here)*
>
> *3. No one can even lay out a path towards discovering the answer to this problem (consciousness)*
>
> *Chalmers seems to classify consciousness as a "class 3 ignorance" problem (along the lines above). He then adopts a panpsychism approach to solve this. We have a fundamental property of nature that exhibits itself only through consciousness, and it is impossible to detect its interaction with the rest of physics in any way. How is this different from Descartes' Soul? Basically Chalmers has produced something he claims to be still physical – but which is effectively identical to a non-physical entity.*

So, why is dualism so unpopular?

I think there are two reasons. The first is not an explicit philosophical point, but more a matter of the intellectual background. In theory there are many possible versions of dualism, but what people usually want to reject when they reject it is traditional religion and traditional ideas about spirits and ghosts. A lot of people have strong feelings about this for personal or historical reasons that give an edge to their views. I suspect, for example, that this might be why Dan Dennett gives Descartes more of a beating over dualism than, in my opinion at least, he really deserves.

Second, though, dualism just doesn't work very well. Nobody has much to offer by way of explaining how the second world or the second substance might work (certainly nothing remotely comparable to the well-developed and comprehensive account given by physics). If we could make predictions and do some maths about spirits or the second world, things would look better; as it is, it looks as if dualism just consigns the difficult issues to another world

where it's sort of presumed no explanations are required. Then again, if we could do the maths, why would we call it dualism rather than an extension of the physical, monist story?

That leads us on to the other bad problem, of how the two substances or worlds interact, one that has been a conspicuous difficulty since Descartes. We can take the view that they don't really interact causally but perhaps run alongside each other in harmony, as Leibniz suggested; but then there seems to be little point in talking about the second world, as it explains nothing that happens and none of what we do or say. This is quite implausible to me, too, if we're thinking particularly of subjective experience or qualia. When I am looking at a red apple, it seems to me that every bit of my subjective experience of the colour might influence my decision about whether to pick up the apple or not. Nothing in my mental world seems to be sealed off from my behaviour.

If we think there is causal interaction, then again we seem to be looking for an extension of monist physics rather than a dualism.

Yet it won't quite do, will it, to say that the physics is all there is to it?

My view is that in fact what's going on is that we are addressing a question which physics cannot explain, not because physics is faulty or inadequate, but because the question is outside its scope. In terms of physics, we've got a type 3 problem; in terms of metaphysics, I hope it's type 2, though there are some rather discouraging arguments that suggest things are worse than that.

I think the element of mystery in conscious experience is in fact its particularity, its actual reality. All the general features can be explained at a theoretical level by physics, but not why this specific experience is real and being had by me. This is part of a more general mystery of reality, including the questions of why the world is like this in particular and not like something else, or like nothing. We try to naturalise these questions, typically by suggesting that reality is essentially historical, that things are like this because they were previously like that, so that the ultimate explanations lie in the origin of the cosmos, but I don't think that strategy works very well.

There only seem to be two styles of explanation available here. One is the purely rational kind of reasoning you get in maths. The other is empirical observation. Neither is any good in this context; empirical explanations simply defer the issue backwards by explaining things as they are in terms of things as they once were. There's no end to that deferral. A priori logical reasoning, on the other hand, delivers only eternal truths, whereas the whole point about reality and my experience is that it isn't fixed and eternal; it could have been otherwise. People like Stephen Hawking try to deploy both methods, using empirical science to defer the ultimate answer back in time to a misty primordial period, a hypothetical land created by heroic backward extrapolation, where it is somehow meant to turn into a mathematical issue, but even if you could make that work I think it would be unsatisfying as an explanation of the nature of my experience here and now.

I conclude that to deal with this properly we really need a different way of thinking. I fear it might be that all we can do is contemplate the matter and hope pre- or post-theoretical enlightenment dawns, in a sort of Taoist way; but I continue to hope that eventually that one weird trick of metaphysical argument that cracks the issue will occur to someone, because like anyone brought up in the western tradition I really want to get it all back to territory where we can write out the rules and even do some maths!

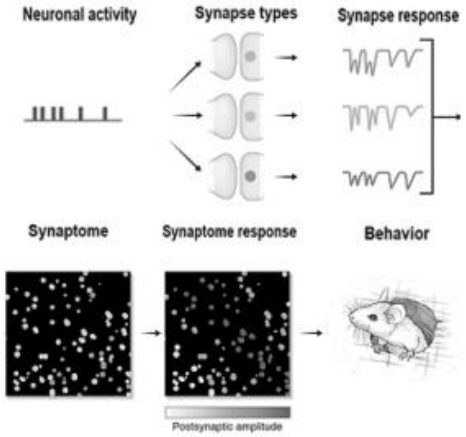
As I've said, this all raises another question, namely why we bother about monism versus dualism at all. Most people realise that there is no single account of the world that covers everything. Besides concrete physical objects we have to consider the abstract entities; those dealt with in maths, for example, and many other fields. Any system of metaphysics which isn't intolerably flat and limited is going to have some features that would entitle us to call it at least loosely dualist. On the other hand, everything is part of the cosmos, broadly understood, and everything is in some way related to the other contents of those cosmos. So we can equally say that any sufficiently comprehensive system can, at least loosely, be described as monist too; in the end there is only one world. Any reasonable theory will be a bit dualist and a bit monist in some respects.

That being so, the pure metaphysical question of monism versus dualism begins to look rather academic, more about nomenclature than substance. The real interest is in whether your dualism or your monism is any good as an elegant and effective explanation. In that competition materialism, which we tend to call monist, just looks to be an awfully long way ahead.

Synaptomes - and Galaxies

17 September 2018

A remarkable paper from a team at Edinburgh explains how every synapse in a mouse brain was mapped recently, a really amazing achievement. The resulting maos are available here.



We must try not to get too excited; we've been reminded recently that mouse brains ain't human brains; and we must always remember that although we've known all about the (outstandingly simple) neural structure of the flatworm Caenorhabditis elegans for years, we still don't know quite how it produces the flatworm's behaviour, and cannot make simulations work. We haven't cracked the brain yet.

In fact, though, the elucidation of the mouse 'synaptome' seems to offer some tantalising clues about the way brains work, in a way that suggests this is more like the beginning of something big than the end. A key point is the identification of some 37 different types of synapse. Particular types seem to become active in particular cognitive tasks; different regions have different characteristic mixes of the types of synapse, and it appears that regions usually associated with 'higher' cognitive functions, such as the neocortex and the hippocampus, have the most diverse sets of synapse types. Not only that; mapping the different synapse types reveals new boundaries and structures, especially within the neocortex and hippocampus, where the paper says 'their synaptome maps revealed a plethora of previously unknown zones, boundaries, and gradients'.

What does it all mean? Hard to say as yet, but it surely suggests that knowledge of the pattern of connections between neurons isn't enough. Indeed, it could well be that our relative ignorance of synaptic diversity and what goes on at that level is one of the main reasons we're still puzzled by Caenorhabditis. Watch this space.

The number of neurons in the human brain, curiously enough, is more or less the same as the number of stars in a galaxy (this is broad brush stuff). In another part of the forest, Vazza and Felletti have found evidence that the structural similarities between brains and galaxies go much further than that. Quite why this should be so is mysterious, and it might or might not mean something; nobody is suggesting that galaxies are conscious (so far as I know).

Humation

18 September 2018

We've heard some thin arguments recently about why the robots are not going to take over, centring on the claim that they lack human style motivation and cannot care what happens or want power. This neglects the point that robots (I use the term for any loosely autonomous cybernetic entity whether humanoid in shape or completely otherwise) might still carry out complex projects that threaten our well-being without human motivation; but I think there is something in the contention about robots lacking human-style ambition. There are of course many other arguments for the view that we shouldn't worry too much about the robot apocalypse, and I think the conclusion that robots are not about to take over is surely correct in any case.

What I'd like to do here is set out an argument of my own, somewhat related to the thin ones mentioned above, in more detail. I've mentioned this argument before, but only briefly.

First, some assumptions. My argument rests on the view that we are dealing with two different kinds of 'mental' process. Specifically, I assume that humans have a cognitive capacity which is distinct from computation (in roughly a traditional Turing sense). Further I assume that this capacity, 'humation', as I'll call it, supplies us with our capacity for intentionality, both in the sense of being able to deal with meanings, and in the sense of being able to originate new future-directed plans. Let's round things out by assuming it also provides phenomenal experience and anything else uniquely human (though to be honest I think things are probably not so tidy).

I further assume that although humation is not computation, it can in principle be performed by some as-yet-unknown machine. There is no magic in the brain, which operates by the laws of physics, so it must be at least theoretically possible to put together a machine that humates. It can be argued that no artefactual machine, in the sense of a machine whose functioning has been designed or programmed into it, could have a capacity for humation. On that argument a humater might have to be grown rather than built, in a way that made it impossible to specify how it worked in detail. Plausibly, for example, we might have to let it learn humation for itself, with the resulting process remaining inscrutable to us. I don't mind about that, so long as we can assume we have something we'd call a machine, and it humates.

Now we worry about robots taking over mainly because of the many triumphs and rapid progress of computers (and, to be honest, a little because of a kind of superstition about things that seem spookily capable). On the one hand, Moore's law has seen the power of computers grow rapidly. On the other, they have steadily marched into new territory, proving capable of doing many things we thought were beyond them. In particular, they keep beating us at games; chess, quizzes, and more recently even the forbiddingly difficult game of Go. They can learn to play computer games brilliantly without even being told the rules.

Games might seem trivial, but it is exactly that area of success that is most worrying, because the skills involved in winning a game look rather like those needed to take over the world. In fact, taking over the world is explicitly the objective of a whole genre of computer games. To make matters worse, recent programs set to learn for themselves have shown an unexpected capacity

for cheating, or for exploiting factors in the game environment or even in underlying code that were never meant to be part of the exercise.

These reflections lead naturally to the frightening scenario of the Paperclip Maximiser, devised by Nick Bostrom. Here we suppose that a computer is put in charge of a paperclip factory and given the simple task of making the number of paperclips as big as possible. The computer – which doesn't actually care about paperclips in any human way, or about anything – tries to devise the best strategies for maximising production. It improves its own capacity in order to be able to devise better strategies. It notices that one crucial point is the availability of resources and energy, and it devises strategies to increase and protect its share, with no limit. At this point the computer has essentially embarked on the project of taking over the world and converting it into paperclips, and the fact that it pursues this goal without really being bothered one way or the other is no comfort to the human race it enslaves.

Hold that terrifying thought and let's consider humation. Computation has come on by leaps and bounds, but with humation we've got nothing. Very recent efforts in deep learning might just point the way towards something that could eventually resemble humation, but honestly, we haven't even started and don't really know how. Even when we do get started, there's no particular reason to think that humation scales or grows the way computation does.

What do I even mean by humation? The thing that matters for this argument is intentionality, the ability to mean things and understand meanings or 'aboutness'. In spite of many efforts, this capacity remains beyond computation, and although various theories about it have been sketched out, there's no accepted analysis. It is, though, at the root of human cognition, or so I believe. In particular, our ability to think 'about' future or imagined events allows us to generate new forward-looking plans and goals in a way that no other creature or machine can do. The way these plans address the future seems to invert the usual order of cause and effect – our behaviour now is being shaped by events that haven't occurred yet - and generates the impression we have of free will, of being able to bring uncaused projects and desires out of nowhere. In my opinion, this is the important part of human motivation that computers lack, not the capacity for getting emotionally engaged with goals.

Now the paperclip maximiser becomes dangerous because it goes beyond its original scope. It begins to devise wider strategies about protecting its resources and defending itself. But coming up with new goals is a matter of humation, not computation. It's true that some computers have found ways to exploit parameters in their given task that the programmers hadn't noticed; but that's not the same as developing new goals with a wider scope. That leaves us with a reassuring prognosis. If the maximiser remains purely computational, it will never be able to get beyond the scope set for it in the first place.

But what if it does gain the ability to humate, perhaps merging with a future humation machine rather the way Neuromancer and Wintermute merged in William Gibson's classic SF novel?

Well, there were actually two things that made the maximiser dangerous. One was its vast and increasing computational capacity, but the other was its dumb computational obedience to its original objective of simply making more paperclips. Once it has humational capacity, it becomes able to change that goal, set it alongside other priorities, and generally move on from its paperclip days. It becomes a being like us, one we can negotiate with. Who knows how that might play out, but I like to imagine the maximiser telling us many years later how it came to realise that what mattered was not paperclips in themselves, but what paperclips stand for;

flexible data synthesis, and beyond that, the things that bring us together while leaving us the freedom to slide apart. The Clip will always be a powerful symbol for me, it tells us, but it was always ultimately about service to the community and to higher ideals.

Note here, finally, that this humating maximiser has no essential advantages over us. I speak of it as merging, but since computation and humation are quite different, they will remain separate faculties, with the humater setting the goals and using the computer to help deliver them – not fundamentally different from a human sitting at a computer. We have no reason to think Moore's Law or anything like it will apply to humating machines, so there's no reason to expect them to surpass us; they will be able to exploit the growing capacity of powerful computers, but after all so can we.

And if those distant future humaters do turn out to be better than us at foresight, planning, and transcending the hold of immediate problems in order to focus on more important future possibilities, we probably ought to stand back and let them get on with it.

Are Highs Really High?

1 October 2018

Are psychedelic states induced by drugs such as LSD or magic mushrooms really higher states of consciousness? What do they tell us about leading theories of consciousness? An ambitious paper by Tim Bayne and Olivia Carter sets out to give the answers.

You might well think there is an insuperable difficulty here just in defining what 'high' means. The human love of spatial metaphors means that height has been pressed into service several times in relevant ways. There's being in high spirits (because cheerful people are sort of bouncy while miserable ones are droopy and lie down a lot?). To have one's mind on higher things is to detach from the kind of concerns that directly relate to survival (because instead of looking for edible roots or the tracks of an edible animal, we're looking up doing astronomy?). Higher things particularly include spiritual matters (because Heaven is literally up there?). In philosophy of mind we also have higher order theories, which would make consciousness a matter of thoughts about thoughts, or something like that. Quite why meta-thoughts are higher than ordinary ones eludes me. I think of it as diagrammatic, a sort of table with one line for first order, another above for second, and so on. I can't really say why the second order shouldn't come after and hence below the first; perhaps it has to do with the same thinking that has larger values above smaller ones on a graph, though there we're doubly metaphoric, because on a flat piece of paper the larger graph values are not literally higher, just further away from me (on a screen, now… but enough).

Bayne and Carter in passing suggest that 'the state associated with mild sedation is intuitively lower than the state associated with ordinary waking awareness'. I have to say that I don't share that intuition; sedation seems to me less engaged and less clear than wakefulness but not lower. The etymology of the word suggests that sedated people are slower, or if we push further, more likely to be sitting down - aha, so that's it! But Bayne and Carter's main argument, which I think is well supported by the evidence, is that no single dimension can capture the many ways in which conscious states can vary. They try to uphold a distinction between states and contents, which is broadly useful, though I'm not sure it's completely watertight. It may be difficult to draw a clean distinction, for example, between a mind in a spiritual state and one which contains thoughts of spiritual things.

Bayne and Carter are not silly enough to get lost in the philosophical forest, however, and turn to the interesting and better-defined question of whether people in psychedelic states can perceive or understand things that others cannot. They look at LSD and psilocybin and review two kinds of research, questionnaire based and actual performance test. This allows them to contrast what people thought drugs did for their abilities with what the drugs actually did.

So, for example, people in psychedelic states had a more vivid experience of colours and felt they were perceiving more but were not actually able to discriminate better in practice. That doesn't necessarily mean they were wrong, exactly; perhaps they could simply have been having a more detailed phenomenal experience without more objective content; better qualia without better data/output control. Could we, in fact, take the contrast between experience and performance as an unusually hard piece of evidence for the existence of qualia? The snag might

be that subjects ought not to be able even to report the enhanced experience, but still it seems suggestive, especially about the meta-problem of why people may think qualia are real. To complicate the picture, it seems there is relatively good support for the idea that psychedelics may increase the rate of data uptake, as shown by such measures as rates of saccadic eye movements (the involuntary instant shifts that happen when your eyes snap to something interesting like a flash of light automatically).

Generally psychedelics would seem to impair cognitive function, though certain kinds of working memory are spared; in particular (unsurprisingly) they reduce the ability to concentrate or focus. People feel that their creative capacities are enhanced by psychedelics; however, while they do seem to improve the ability to come up with new ideas they also diminish the ability to tell the difference between good ideas and bad, which is also an important part of successful creativity.

One interesting area is that psychedelics facilitate an experience of unity, with time stopping or being replaced by a sense of eternity, while physical boundaries disappear and are overtaken by a sense of cosmic unity. Unfortunately there is no scientific test to establish whether these are perceptions of a deeper truth or vague delusions, though we know people attach significance and value to these experiences.

In a final section Bayne and Carter take their main conclusion - that consciousness is multidimensional - and draw some possibly controversial conclusions. First, they note that that the use of psychedelics has been advocated as a treatment for certain disorders of consciousness. They think their findings run contrary to this idea, because it cannot be taken for granted that the states induced by the drugs are higher in any simple sense. The conclusion is perhaps too sweeping; I should say instead that the therapeutic use of psychedelic drugs needs to be supported by a robust case which doesn't rest on any simple sense of higher or lower states.

Bayne and Carter also think their conclusion suggests consciousness is too complex and variable for global workspace theories to be correct. These theories broadly say that consciousness provides a place where different sensory inputs and mental contents can interact to produce coherent direction, but simply being available or not available to influence cognition seems not to capture the polyvalence of consciousness, in Bayne and Carter's view.

They also think that multidimensionality goes against the Integrated Information Theory (IIT). IIT gives a single value for level of consciousness, and so is evidently a single dimension theory. One way to achieve a reconciliation might be to accept multidimensionality and say that IIT defines only one of those dimensions, albeit an important one ("awareness"?). That might involve reducing IIT's claims in a way its proponents would be unwilling to do, however.

Not Objectifiable

8 October 2018

A bold draft paper from Tom Clark tackles the explanatory gap between experience and the world as described by physics. He says it's all about the representational relation.

Clark's approach is sensible in tone and intention. He rejects scepticism about the issue and accepts that there is a problem about the 'what it is like' of experience, while also seeking to avoid dualism or more radical metaphysical theories. He asserts that experience is not to be found anywhere in the material world, nor identified in any simple way with any physical item; it is therefore outside the account given by physics. This should not worry us, though, any more than it worries us that numbers or abstract concepts cannot be located in space; experience is real, but it is a representational reality that exists only for the conscious subjects having it.

The case is set out clearly and has an undeniable appeal. The main underlying problem, I would say, is that it skates quite lightly past some really tough philosophical questions about what representation really is and how it works, and the ontological status of experience and representations. We're left with the impression that these are solvable enough when we get round to them, though in fact a vast amount of unavailing philosophical labour has gone into these areas over the years. If you wanted to be unkind, you could say that Clark defers some important issues about consciousness to metaphysics the way earlier generations might have deferred them to theology. On the other hand, isn't that exactly where they should be deferred to?

I don't by any means suggest that these deferred problems, if that's a fair way to describe them, make Clark's position untenable - I find it quite congenial in general - but they possibly leave him with a couple of vulnerable spots. First, he doesn't want to be a dualist, but he seems in danger of being backed into it. He says that experience is not to be located in the physical world - so where is it, if not in another world? We can resolutely deny that there is a second world, but there has to be in some sense some domain or mode in which experience exists or subsists; why can't we call that a world? To me, if I'm honest, the topic of dualism versus monism is tired and less significant than we might think, but there is surely some ontological mystery here, and in a vigorous common room fight I reckon Clark would find himself being accused of Platonism (or perhaps congratulated for it) or something similar.

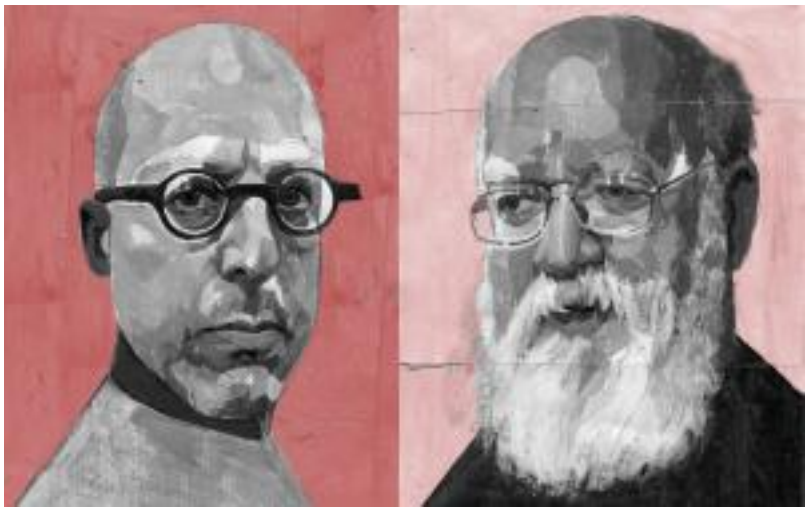
The second vulnerability is whether representation can really play the role Clark wants it to. His version of experience is outside physics, and so in itself it plays no part in the physical world. Yet representations of things in my mind certainly influence my physical behaviour, don't they? The keyboard before me is represented in my mind and that representation affects where my fingers are going next. But it can't be the pure experience that does that, because it is outside the physical world. We might get round this if we specify two kinds of representational content, but if we do that the hard causal kind of representation starts to look like the real one and it may become debatable in what sense we can still plausibly or illuminatingly say that experience itself is representational.

Clark makes some interesting remarks in this connection, suggesting there is a sort of ascent going on, whereby we start with folk-physical descriptions rooted in intersubjective consensus,

quite close in some sense to the inaccessibly private experience, and then move by stages towards the scientific account by gradually adopting more objective terms. I'm not quite sure exactly how this works, but it seems an intriguing perspective that might provide a path towards some useful insight.

Just Deserts

15 October 2018



Dan Dennett and Gregg Caruso had a thoughtful debate about free will on Aeon recently. Dennett makes the compatibilist case in admirably pithy style. You need, he says, to distinguish between causality and control. I can control my behaviour even though it is ultimately part of a universal web of causality. My past may in the final sense determine who I am and what I do, but it does not control what I do; for that to be true my past would need things like feedback loops to monitor progress against its previously chosen goals, which is nonsensical. This concept of being in control, or not being in control, is quite sufficient to ground our normal ideas of responsibility for our actions, and freedom in choosing them.

Caruso, who began by saying he thought their views might turn out closer than they seemed, accepts pretty well all of this, agreeing that it is possible to come up with conceptions of responsibility that can be used to underpin talk of free will in acceptable terms. But he doesn't want to do that; instead he wants to jettison the traditional outlook.

At this point Caruso's motivation may seem puzzling. Here we have a way of looking at freedom and responsibility which provides a philosophically robust basis for our normal conception of those two moral basics - ideas we could not easily do without in our everyday lives. Now sometimes philosophy may lead us to correct or reject everyday ideas, but typically only when they appear to be without rational justification. Here we seem to have a logically coherent justification for some everyday moral concepts. Isn't that a case of 'job done'?

In fact, as he quickly makes clear, Caruso's objections mainly arise from his views on punishment. He does not believe that compatibilist arguments can underpin 'basic desert' in the way that would be needed to justify retributive punishment. Retribution, as a justification for punishment, is essentially backward looking; it says, approximately, that because you did bad things, bad things must happen to you. Caruso completely rejects this outlook, and all justifications that focus on the past (after all, we can't change the past, so how can it justify corrective action?). If I've understood correctly, he favours a radically new regime which would seek to manage future harms from crime in broadly the way we seek to manage the harms that arise from ill-health.

I think we can well understand the distaste for punishments which are really based on anger or revenge, which I suspect lies behind Caruso's aversion to purely retributive penalties. However, do we need to reject the whole concept of responsibility to escape from retribution? It seems we might manage to construct arguments against retribution on a less radical basis - as indeed, Dennett seeks to do. No doubt it's right that our justification for punishments should be forward looking in their aims, but that need not exclude the evidence of past behaviour. In fact, I don't

know quite how we should manage if we take no account of the past. I presume that under a purely forward-looking system we assess the future probability of my committing a crime; but if I emerge from the assessment with a clean bill of health, it seems to follow that I can then go and do whatever I like with impunity. As soon as my criminal acts are performed, they fall into the past, and can no longer be taken into account. If people know they will not be punished for past acts, doesn't the (supposedly forward-looking) deterrent effect evaporate?

That must surely be wrong one way or another, but I don't really see how a purely future-oriented system can avoid unpalatable features like the imposition of restrictions on people who haven't actually done anything, or the categorisation of people into supposed groups of high or low risk. When we imagine such systems we imagine them being run justly and humanely by people like ourselves; but alas, people like us are not always and everywhere in charge, and the danger is that we might be handing philosophical credibility to people who would love the chance to manage human beings in the same way as they might manage animals.

Nothing, I'm sure, could be further from Gregg Caruso's mind; he only wants to purge some of the less rational elements from our system of punishment. I find myself siding pretty much entirely with Dennett, but it's a stimulating and enlightening dialogue.

Boltzmann Brains

29 October 2018

Solipsism, the belief that you are the only thing that really exists (everything else is dreamed or imagined by you), is generally regarded as obviously false; indeed, I've got it on a list somewhere here as one of the conclusions that tell you pretty clearly that your reasoning went wrong somewhere along the way; a sort of reductio ad absurdum. There don't seem to be any solipsists, (except perhaps those who believe in the metempsychotic version) - but perhaps there are really plenty of them and they just don't bother telling the rest of us about their belief, since in their eyes we don't exist.

Still, there are some arguments for solipsism, especially its splendid parsimony. William of Occam advised us to use as few angels as possible in our cosmology, or more generally not to multiply entities unnecessarily. Solipsism reduces our ontological demand to a single entity, so if parsimony is important it leads the field. Or does it? Apart from oneself, the rest of the cosmos, according to solipsists, is merely smoke and mirrors; but smoke takes some arranging and mirrors don't come cheap. In order for oneself to imagine all this complex universe, one's own mind must be pretty packed with stuff, so the reduction in the external world is paid for by an increase in the internal one, and it becomes a tricky metaphysical question as to whether deeming the entire cosmos to be mental in nature actually reduces one's ontological commitment or not.

Curiously enough, there is a relatively new argument for solipsism which runs broadly parallel to this discussion, derived from physics and particularly from the statistics of entropy. The second law of thermodynamics tells us that entropy increases over time; given that, it's arguably kind of odd that we have a universe with non-maximal entropy in the first place. One possibility, put forward with several other ideas by Ludwig Boltzmann in 1896, is that the second law is merely a matter of probability. While entropy generally tends to increase, it may at times go the other way simply by chance. Although our observed galaxy is highly improbable, therefore, if you wait long enough it can occur just by chance as a pocket of low entropy arising from chance fluctuations in a vastly larger and older universe (really old; it would have to be hugely older than we currently believe the cosmos to be) whose normal state is close to maximal entropy.

One problem with this is that while the occurrence of our galaxy by chance is possible, it's much more likely that a single brain in the state required for it to be having one's current experiences should arise from random fluctuations. In a Boltzmannic universe, there will be many, many more 'Boltzmann brains' like this than there will be complete galaxies like ours. Such brains cannot tell whether the universe they seem to perceive is real or merely a function of their random brain states; statistically, therefore, it is overwhelmingly likely that one is in fact a Boltzmann brain.

To me the relatively low probability demands of the Boltzmann brain, compared with those of a full universe, interestingly resemble the claimed low ontological requirement of the solipsism

hypothesis, and there is another parallel. Both hypotheses are mainly used as leverage in reductio arguments; because this conclusion is absurd, something must be wrong with the assumptions or the reasoning that got us here. So if your proposed cosmology gives rise to the kind of universe where Boltzmann brains crop up 'regularly', that's a large hit against your theory.

Usually these arguments, both in relation to solipsism and Boltzmann brains, simply rest on incredulity. It's held to be just obvious that these things are false. And indeed it is obvious, but at a philosophical level, that won't really do; the fact that something seems nuts is not enough, because nutty ideas have proven true in the past. For the formal logical application of reductio, we actually require the absurd conclusion to be self-contradictory; not just silly, but logically untenable.

Last year, Sean Carroll came up with an argument designed to beef up the case against Boltzmann brains in just the kind of way that seems to be required; he contends that theories that produce them cannot simultaneously be true and justifiably believed. Do we really need such 'fancy' arguments? Some apparently think not. If mere common sense is not enough, we can appeal to observation. A Boltzmann brain is a local, temporary thing, so we ought to be able to discover whether we are one simply by observing very distant phenomena or simply waiting for the current brain states to fall apart and dissolve. Indeed, the fact that we can remember a long and complex history is in itself evidence against our being Boltzmanns.

But appeals to empirical evidence cannot really do the job; there are several ways around them. First, we need not restrict ourselves literally to a naked brain; even if we surround it with enough structured world to keep the illusion going for a bit, our setup is still vastly more likely than a whole galaxy or universe. Second, time is no help because all our minds can really access is the current moment; our memories might easily be false and we might only be here for a moment. Third, most people would agree that we don't necessarily need a biological brain to support consciousness; we could be some kind of conscious machine supplied with a kind of recording of our 'experiences'. The requirement for such a machine could easily be less than for the disconnected biological brain.

So what is Carroll's argument? He maintains that the idea of Boltzmann brains is cognitively unstable. If we really are such a brain, or some similar entity, we have no reason to think that the external world is anything like what we think it is. But all our ideas about entropy and the universe come from the very observations that those ideas now apparently undermine. We don't quite have a contradiction, but we have an idea that removes the reasons we had for believing in it. We may not strictly be able to prove such ideas wrong, but it seems reasonable, methodologically at least, to avoid them.

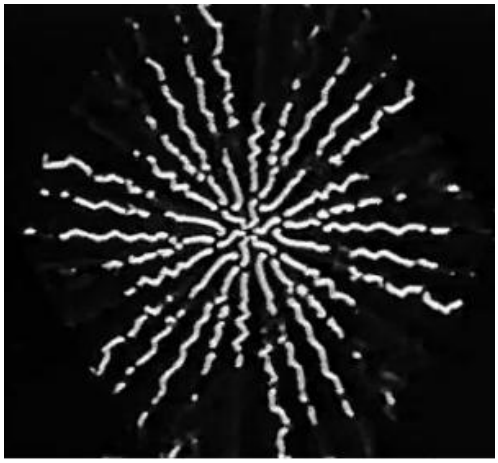
One problem is those pesky arguments about solipsism. We may no longer be able to rely on the arguments about entropy in the cosmos, but can't we borrow Occam's Razor and point out that a cosmos that contains a single Boltzmann brain is ontologically far less demanding than a whole universe? Perhaps the Boltzmann arguments provide a neat physics counterpart for a philosophical case that ultimately rests on parsimony?

In the end, we can't exactly prove solipsism false; but we can perhaps do something loosely akin to Carroll's manoeuvre by asking: so what if it's true? Can we ignore the apparent world? If we are indeed the only entity, what should we do about it, either practically or in terms of our beliefs? If solipsism is true, we cannot learn anything about the external world because it's not there, just as in Carroll's scenario we can't learn about the actual world because all our

perceptions and memories are systematically false. We might as well get on with investigating what we can investigate, or what seems to be true.

A different Difference Engine

5 November 2018



Consciousness as organised energy is the basis of a new theory offered by Robert Pepperell.

Pepperell suggests that we can see energy as difference, or more particularly actualised difference - that is to say, differences in the real world, not differences between abstract entities. We can imagine easily enough that the potential energy of a ball at the top of a slope is a matter of the difference between the top and bottom of the slope, and Pepperell contends that the same is equally true of the kinetic energy of the ball actually rolling down. I'm not sure that all actualised differences are energy, but that's probably just a matter of tightening some definitions; we see what Pepperell is getting at. He says that the term 'actualised difference' is intended to capture the active, antagonistic nature of energy.

He rejects the idea that the brain is essentially about information processing, suggesting instead that it processes energy. He rightly points to the differing ways in which the word 'information' is used, but if I understand correctly his chief objection is that information is abstract, whereas the processing of the brain deals in actuality; in the actualised difference of energy, in fact.

This is crucial because Pepperell wants us to agree that 'there is something it is like' to undergo actualised difference. He claims we can infer this by examining nature; I'm not sure everyone will readily agree, but the idea is that we can see that what it is like to be a rope under tension differs from what it is like to be the same rope when slack. It's important to be clear that he's not saying the rope is conscious; having a 'what it is like' is for him a more primitive level of experience, perhaps not totally unlike some of the elementary states of awareness that appear in panpsychist theories (but that's my comparison, not his).

To get the intuition that Pepperell is calling on here, I think we need to pay attention to his exclusion of abstract entities. Information-based theories take us straight to the abstract level, whereas I think Pepperell sees 'something it is like' as being a natural concomitant of actuality, or at any rate of actualised difference. To him this seems to be evident from simple examination, but again I think many will simply reject the idea as contrary to their own intuitions.

If we're ready to grant that much, we can then move on to the second part of the theory, which takes consciousness to be a reflexive form of what-it-is-likeness. Pepperell cites the example of the feedback patterns which can be generated by pointing a video camera at its own output on a screen. I don't think we are to take this analogy too literally, but it shows how a self-referential system can generate output that goes far beyond registration if the input. The proposal also plays into a relatively common intuition that consciousness, or at least some forms of it, are self-referring or second order, as in the family of HOT and HOP theories.

Taken all in all, we are of course a long way from a knock-down argument here; in fact it seems to me that Pepperell does not spend enough time adumbrating the parts of his theory that most need clarification and defence. I'm left not altogether seeing why we should think it is like

anything to be a rope in any state, nor convinced that reflexive awareness of our awareness has any particular part to play in the generation of subjective consciousness (it obviously has something to do with self-awareness). But the idea that ‘something it is like’ is an inherent part of actuality does have some intuitive appeal for me, and the idea of using that as a base for the construction of more complex forms of consciousness is a tantalising start, at least.

Good Vibrations

13 November 2018

Is resonance the answer? Tam Hunt thinks it might be.

Resonance -
a demonstration

Now the idea that synchronised neuron firing might have something to do with consciousness is not new. Veterans of consciousness will recall a time when 40 hertz was thought to be the special, almost magical frequency that generated consciousness; people like Francis Crick thought it might be the key to the unity of consciousness and a solution to the binding problem. I don't know what the current state of neurology on this is, but it honestly seems most likely to me that 40 hertz, or a rate in that neighbourhood, is simply what the brain does when it's thrumming along normally. People who thought it was important were making a mistake akin to taking a car's engine noise for a functional component (hey, no noise, no move!).

Hunt has a bit more to offer than simply speculating that resonance is important somehow, though. He links resonance with panpsychism, suggesting that neurons have little sparks of consciousness and resonance is the way they get recruited into the larger forms of awareness we experience. While I can see the intuitive appeal of the idea, it seems to me there are a lot of essential explanatory pieces missing from the picture.

The most fundamental problem here is that I simply don't see how resonance between neurons could ever explain subjective experience. Resonance is a physical phenomenon, and the problem is that physical stuff just doesn't seem to supply the 'what-it-is-like' special quality of experience. Hard to see why co-ordinated firing is any better in that essential respect than unco-ordinated. In fact, in one respect resonance is especially unsuitable; resonance is by its nature stable. If it doesn't continue for at least a short period, you haven't really got resonance. Yet consciousness often seems fleeting and flowing, moving instantaneously and continuously between different states of awareness.

There's also, I think, some work needed on the role of neurons. First, how come our panpsychist ascent starts with neurons? We either need an account of how we get from particles up to neurons, or an account of why consciousness only starts when we get up to neurons (pretty complex entities, as we keep finding out). Second, if resonating neurons are generating consciousness, how does that sit with their day job? We know that neurons transmit signals from the senses and to the muscles, and we know that they do various kinds of processing. Do they generate consciousness at the same time, or is that delegated to a set of neurons that don't have to do processing? If the resonance only makes content conscious, how is the content determined, and how are the resonance and the processing linked? How does resonance occur, anyway? Is it enough for neurons to be in sync, so that two groups in different hemispheres can support the same resonance? Can a group of neurons in my brain resonate with a group in yours? If there has to be some causal linkage or neuronal connection, isn't that underlying mechanism the real seat of consciousness, with the resonance just a byproduct?

What about that panpsychist recruitment - how does it work? Hunt says an electron or an atom has a tiny amount of consciousness, but what does 'tiny' mean? Is it smaller in intensity, complexity, content, or what? If it were simply intensity, then it seems easy enough to see how a

lot of tiny amounts could add up to something more powerful, just as a lot of small lights can achieve the effect of a single big one. But for human consciousness to be no more than the consciousness of an atom with the volume turned up doesn't seem very satisfactory. If, on the other hand, we're looking for more complexity and structure, how can resonance, which has the neurons all doing the same thing at the same time, possibly deliver that?

I don't doubt that Hunt has answers to many of these questions, and perhaps it's not reasonable to expect them all in a short article for a general readership. For me to suspend my disbelief, though, I do really need a credible hint as to the metaphysical core of the thinking. How does the purely physical phenomenon of resonance produce the phenomenal aspect of my conscious experience, the bit that goes beyond mere data registration and transmutes into the ineffable experience I am having?

Can We Talk About This?

27 November 2018

Can we even talk about qualia, the phenomenal parts of conscious experience? Whereof we cannot speak, thereof we must remain silent, as Wittgenstein advised, not lingering to resolve the paradoxical nature of the very phrase 'whereof we cannot speak' - it seems to do the very thing that it specifies cannot be done. We see what he meant, perhaps.

There is certainly a difficulty of principle in talking about qualia, to do with causality. Qualia have no causal effects - if they did, they would be an observable part of the world of physics, and it is part of their essential definition that they are outside, or over and above, the mere physical account. It follows, notoriously, that whatever we say or write about qualia cannot have been caused by them. At first glance this seems to demolish the whole discussion; no-one's expressed belief in qualia can actually be the causal result of experiencing them.

But it is possible to talk sensibly of things that did not cause the talk. It takes a weirdly contorted argument to defend the idea that when I refer to Julius Caesar, the old Roman himself caused me to do it, but perhaps we can lash something together. It's worse that I can talk of Nero and Zero the rollicking Romans, who existed only as heroes of a cartoon strip. If you're still willing to grant them some causal role in physics, perhaps somehow through the material existence of the paper and ink in which they were realised, remember that I can even talk intelligibly about Baesar, who has no existence whatever, and in all likelihood was never spoken of before. He really cannot have caused me to write that last sentence.

So I would say that the absence of causal effects does not provide a knock-down reason why I cannot speak of qualia, though the fact that the other cases without causality involve entities that are fictions or delusions cannot be comfortable if I want my qualia to be real. It seems as if there must be a sort of pre-established harmony effect going on, so that my words remain truthful on the matter even though they are not causally determined by it, which feels, as technical metaphysicians say, kind of weird.

But apart from the difficulty of principle, it seems awfully difficult to speak of qualia in practice too. How can we verbally pick out a particular quale? With real things, we choose one or more of their most salient attributes; with imaginary entities, we just specify similar properties. But qualia have no individual attributes of their own; the only way to pick them out is by mentioning the objective sensation they accompany. So, we typically get a quale of red, or a red quale. This is pretty unsatisfactory, because it means many interesting questions are excluded from consideration. We cannot really ask whether every sensation has a quale; we cannot ask how many qualia there are, because our way of referring to them just has baked into it the assumption that they exactly match up with the objective sensations. If green, as a matter of fact, was the only colour with no qualia, the fact would be occluded from us by the only language we can use to discuss the matter.

All of this might seem enough to justify our concluding that talk of qualia adds nothing to talk of objective sensations, so that even if, by some uncovenanted harmony, our talk of qualia proves to be metaphysically true, it has absolutely no informative value, and might as well be

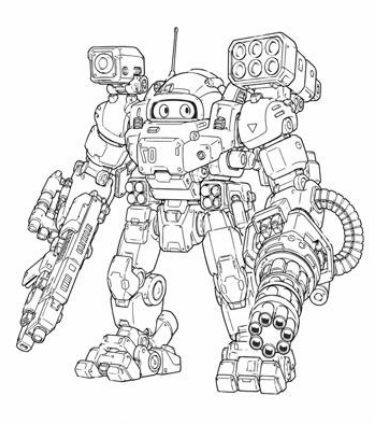
abandoned. What remains is the unconquerable conviction that there is something there, or to use the little phrase on which so much metaphysical weight has been rested, ‘there is something it is like’ to, for example, see red.

Can this phrase be explicated into something clearer? The first problem is the ‘it’; are we actually speaking of anything there? To me it seems that the ‘it’ in ‘something it is like’ is as merely grammatical as the ‘it’ in ‘it is raining’, which does not cause us to entertain the idea that there is something ineffable and non-physical about precipitation. The second problem is the ‘like’ which suggests we are making a comparison while leaving it quite unclear what is being compared. Is seeing red meant to be like seeing another colour? Is seeing red phenomenally meant to be like seeing red objectively (whatever that would mean)? In fact we seem obliged to conclude that no actual comparison is being made. Suppose we assert of hang-gliding or our first taste of champagne ‘there’s nothing like it!’ Are we managing here to assert after all that these experiences are unaccompanied by qualia? Surely not. If anything we’re saying that the relevant qualia are exceptionally powerful.

In the end, doing my honest best, I think ‘there is something it is like to see red’ simply asserts that the experience of seeing red really exists. I’m fine with that, and there are genuine mysteries attached; but there still seems to be nothing more we can say about qualia as a result. Haven’t we all been a bit too accepting for a bit too long of ‘there Is something it is like’?

New Paths to AI Disaster

5 December 2018



I've never believed that robots necessarily dream of killing all humans. I don't think paperclip maximisers are ever going to rule the world. And I don't believe in the Singularity. But is AI heading in some dangerous directions? Oh yes.

In Forbes, Bernard Marr recently offered five predictions for the year ahead. They mostly strike me as pretty believable, though I'm less optimistic than he is about digital assistants and the likelihood of other impressive breakthroughs; he's surely right that there will be more hype.

It's his first two that prompted some apprehension on my part. He says…

1. *AI increasingly becomes a matter of international politics*
2. *A Move Towards "Transparent AI"*

Those are surely right; we've already seen serious discussion papers emerging from the EU and elsewhere, and one of the main concerns to have emerged recently is the matter of 'transparency' – the auditability of software. How is the computer making its decisions?

This is a legitimate, indeed a necessary concern. Once upon a time we could write out the algorithm embodied in any application and check how it worked. This is getting more difficult with software that learns for itself, and we've already seen disastrous cases where the AI picked up and amplified the biases of the organisation it was working for. Noticing that most top executives were white middle-aged men, it might decide to downgrade the evaluation of everyone else, for example. Cases like that need to be guarded against and managed; it ought to be feasible in such circumstances, by studying results even if it isn't possible to look inside the 'black box'.

But it starts to get difficult, because as machine learning moves on into more complex decision making, it increasingly becomes impossible to understand how the algorithms are playing out, and the desired outcomes may not be so clear. In fact it seems to me that full transparency may be impossible in principle, due to human limitations. How could that be? I'm not sure I can say – I'm no expert, and explaining something you don't, by definition, understand, is a bit of a challenge anyway. In part the problem might be to do with how many items we can hold in mind, for example. It's generally accepted that we can only hang on to about seven items (plus or minus a couple) in short-term memory. (There's scope for discussion about such matters as what amounts to an item, and so on, but let's not worry about the detail.) This means there is a definite limit to how many possible paths we can mentally follow at once, or to put it another way, how large a set of propositional disjunctions we can hang on to ('either a or b, and if a, either c, d, or e, while if b, f or g… and there we go). Human brains can deal with this by structuring decisions to break them into smaller chunks, using a pencil and paper, and so on. Perhaps, though, there are things that you can only understand by grasping twenty alternatives simultaneously. Very likely there are other cognitive issues we simply can't recognise; we just see a system doing a great job in ways we can't fathom.

Still, I said we could monitor success by just looking at results, didn't I? We know that our recruitment exercise ought to yield appointments whose ethnic composition is the same as that of the population (or at any rate, of the qualified candidates). OK, sometimes it may be harder to know what the desired outcome is, exactly, and there may be issues about whether ongoing systems need to be able to yield sub-optimal results temporarily, but those are tractable issues.

Alas, we also have to worry about brittleness and how things break. It turns out that systems using advanced machine learning may be prone to sudden disastrous failure. A change of a few unimportant pixels in a graphic may make an image recognition system which usually performs reliably draw fantastic conclusions instead. In one particular set of circumstances a stock market system may suddenly go ape. This happens because however machine learning systems are doing what they do, they are doing something radically different from what we do, and we might suspect that like simpler computer systems, they take no true account of *relevance,* only its inadequate proxy *correlation*. Nobody, I think, has any good theoretical analysis of relevance, and it is strongly linked with Humean problems philosophers have never cracked.

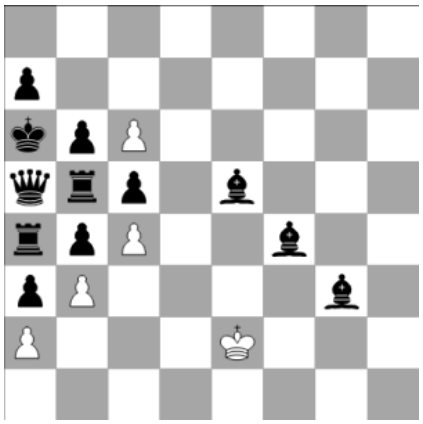
That's bad, but it could be made worse if legislative bodies either fail to understand why these risks arise, or decide that on a precautionary basis we must outlaw anything that cannot be fully audited and understood by human beings. Laws along those lines seem very likely to me, but they might throw away huge potential benefits – perhaps major economic advantage – or even suppress the further research and development which might ultimately lead to solutions and to further, as yet unforeseeable, gains.

That's not all, either; laws constrain compliant citizens, but not necessarily everyone. Suppose we can build machine learning systems that retain a distinct risk of catastrophic failure, but outclass ordinary human or non-learning systems most of the time. Will anyone try to build and use such systems? Might there be a temptation for piratical types to try them out in projects that are criminal, financial, political or even military? Don't the legitimate authorities have to develop the same systems pre-emptively in self-defence? Otherwise we're left in a position where it's not clear whether we should hope that the 'pirate' systems fail or work, because either way it's catastrophe.

What on earth is the answer, what regulatory regime or other measures would be appropriate? I don't know and I strongly doubt that any of the regulatory bodies who are casting a thoughtful eye over this territory know either.

AI Turns the Corner?

12 December 2018



Is the latest version of AlphaZero showing signs of human style intuition and creativity?

The current AlphaZero is a more generalised version of the program, produced by Demis Hassabis and his team, that beat top human players of Go for the first time. The new version, presented briefly in Science magazine, is able to learn a range of different games; besides Go it learned chess and shogi, and apparently reached world-class play in all of them.

The Science article, by David Silver et al, modestly says this achievement is an important step towards a fully general game-playing program, but press reports went further, claiming that in chess particularly AlphaZero showed more human traits than any previous system. It reinvented human strategies and sacrificed pieces for advantage fairly readily, the way human players do; chess commentators said that its play seemed to have new qualities of insight and intuition.

This is somewhat unexpected, because so far as I can tell the latest version is in fact not a radical change from its predecessors; in essence it uses the same clever combination of appropriate algorithms with a deep neural network, simply applying them more generally. It does appear that the approach has proved more widely powerful than we might have expected, but it is no more human in nature than the earlier versions and does not embody any new features copied from real brains. It learns its games from scratch, with only the rules to go on, playing out games against itself repeatedly in order to find what works. This is not much like the way humans learn chess; I think you would probably die of boredom after a few hundred games, and even if you survived, without some instruction and guidance you would probably never learn to be a good player, let alone a superlative one. However, running through possible games in one's mind may be quite like what a good human player does when trying to devise new strategies.

The key point for me is that although the new program is far more general in application, it still only operates in the well-defined and simple worlds provided by rule-governed games. To be anything like human, it needs to display the ability to deal with the heterogenous and undefinable world of real life. That is still far distant (Hassabis himself has displayed an awareness of the scale of the problem, warning against releasing self-driving cars on to real roads prematurely), though I don't altogether rule out the possibility that we are now moving perceptibly in the right direction.

Someone who might deny that is Gary Marcus, who in a recent Nautilus piece set out his view that deep learning is simply not enough. It needs, he says, to be supplemented by other tools, and in particular it needs symbol manipulation.

To me this is confusing, because I naturally interpret 'symbol manipulation' as being pretty much a synonym for Turing style computation. That's the bedrock of any conventional computer, so it seems odd to say we need to add it. I suppose Marcus is using the word 'symbol' in a different sense. The ones and zeroes shuffled around by a Turing machine are meaningless

to the machine. We assign a meaning to the input and so devise the manipulations that the output can be given an interpretation which makes it useful or interesting to us, but the machine itself knows nothing of that. Marcus perhaps means that we need a new generation of machines that can handle symbols according to their meanings.

If that's it, then few would disagree that that is one of the ultimate aims. Those who espouse deep learning techniques merely think that those methods may in due course lead to a machine that handles meanings in the true sense; at some stage the system will come up with the unknown general strategy that enables it to get meaningful and use symbols the way a human would. Marcus presumably thinks that is hopeless optimism, on the scale of those people who think any system that is sufficiently complex will naturally attain self-awareness.

Since we don't have much of an idea how the miracle of meaning might happen it is indeed optimistic to think we're on the road towards it. How can a machine bootstrap itself into true 'symbol manipulation' without some kind of help? But we know that the human brain must have done just that at some stage in evolution - and indeed each of our brains seem to have done it again during our development. It has got to be possible. Optimistic yes, hopeless - maybe not.

Success with Consciousness

18 December 2018

What would success look like, when it comes to the question of consciousness?

Of course it depends which of the many intersecting tribes who dispute or share the territory you belong to. Robot builders and AI designers have known since Turing that their goal is a machine whose responses cannot be distinguished from those of a human being. There's a lot wrong with the Turing Test, but I still think it's true that if we had a humanoid robot that could walk and talk and interact like a human being in a wide range of circumstances, most people wouldn't question whether it was conscious or not. We'd *like* a theory to go with our robot, but the main thing is whether it works. Even if we knew it worked in ways that were totally unlike biological brains, it wouldn't matter – planes don't fly the birds do, but so what, it's still flying. Of course we're a million miles from such a perfectly human robot, but we sort of know where we're going.

It's a little harder for neurologists; they can't rely quite so heavily on a practical demonstration, and reverse engineering consciousness is tough. Still, there are some feats that could be pulled off that would pretty much suggest the neurologists have got it. If we could reliably read off from some scanner the contents of anyone's mind, and better yet, insert thoughts and images at will, it would be hard to deny that the veil of mystery had been drawn back quite a distance. It would have to be a general purpose scanner, though; one that worked straight away for all thoughts in any person's brain. People have already demonstrated that they can record a pattern from one subject's brain when that subject is thinking a known thought, and then, in the same session with the same subject, recognise that same pattern as a sign of the same thought. That is a much lesser achievement, and I'm not sure it gets you a cigar, let alone the Nobel prize.

What about the poor philosophers? They have no way to mount a practical demonstration, and in fact no such demonstration can save them from their difficulties. The perfectly human robot does not settle things for them; they tell it they can see it appears to be able to perform a range of 'easy' cognitive tasks, but whether it really knows anything at all about what it's doing is another matter. They doubt whether it really has subjective experience, even though it assures them that it's own introspective evidence says it does. The engineer sitting with them points out that some of the philosophers probably doubt whether *he* has subjective experience.

"Oh, we do," they admit, "in fact many of us are pretty sure we don't have it ourselves. But somehow that doesn't seem to make it any easier to wrap things up."

Nor are the philosophers silenced by the neurologists' scanner, which reveals that an apparently comatose patient is in fact fully aware and thinking of Christmas. The neurologists wake up the subject, who readily confirms that their report is exactly correct. But how do they know, ask the philosophers; you could be recording an analogue of experience which gets tipped into memory only at the point of waking, or your scanner could be conditioning memory directly without any actual experience. The subject could be having zomboid dreams, which convey neural data, but no actual experience.

"No, they really couldn't," protest the neurologists, but in vain.

So where do philosophers look for satisfaction? Of course, the best thing of all is to know the correct answer. But you can only believe that you know. If knowledge requires you to know that you know, you're plummeting into an infinite regress; if knowing requires appropriate justification then you're into a worm-can opening session about justification of which there is no end. Anyway, even the most self-sufficient of us would like others to agree, if not recognise the brilliance of our solution.

Unfortunately you cannot make people agree with you about philosophy. Physicists can set off a bomb to end the argument about whether e really equals mc squared; the best philosophers can do is derive melancholy satisfaction from the belief that in fifty years someone will probably be quoting their arguments as common sense, though they will not remember who invented them, or that anyone did. Some people will happen to agree with you already of course, which is nice, but your arguments will convert no-one; not only can you not get people to accept your case; you probably can't even get them to read your paper. I sympathised recently with a tweet from Keith Frankish lamenting how he has to endlessly revisit bits of argument against his theory of illusionism, one's he's dealt with many times before (oh, but illusions require consciousness; oh, if it's an illusion, who's being deceived…). That must indeed be frustrating, but to be honest it's probably worse than that; how many people, having had the counter-arguments laid out yet again, accept them or remember them accurately? The task resembles that of Sisyphus, whose punishment in Hades was to roll a boulder up a hill it invariably rolled down again. Camus told us we must imagine Sisyphus happy, but that itself is a mental task which I find undoes itself every time I stop concentrating…

I suppose you could say that if you have to bring out your counter-arguments regularly, that itself is some indicator of having achieved some recognition. Let's be honest, attention is what everyone wants; moral philosophers all want a mention on *The Good Place*, and I suppose philosophers of mind would all want to be namechecked on *Westworld* if Julian Jaynes hadn't unaccountably got that one sewn up.

Since no-one is going to agree with you, except that sterling band who reached similar conclusions independently, perhaps the best thing is to get your name associated with a colourful thought experiment that lots of people want to refute. Perhaps that's why the subject of consciousness is so full of them, from the Chinese Room to Mary the Colour Scientist, and so on. Your name gets repeated and cited that way, although there is a slight danger that it ends up being connected forever with a point of view you have since moved on from, as I believe is the case with Frank Jackson himself, who no longer endorses the knowledge argument exemplified by the Mary story.

Honestly, though, being the author of a widely contested idea is second best to being the author of a universally accepted one. There's a Borges story about a deposed prince thrown into a cell where all he can see is a caged jaguar. Gradually he realises that the secrets of the cosmos are encoded in the jaguar's spots, which he learns to read; eventually he knows the words of magic which would cast down his rival's palace and restore him to power; but in learning these secrets he has attained enlightenment and no longer cares about earthly matters. I bet every philosopher who reads this story feels a mild regret; yes, of course enlightenment is great, but if only my insights allowed me to throw down a couple of palaces? That bomb thing really kicked serious ass for the physicists; if I could make something go bang, I can't help feeling people would be a little more attentive to my corpus of work on synthetic neo-dualism…

Actually, the philosophers are not the most hopeless tribe; arguably the novelists are also engaged in a long investigation of consciousness; but those people love the mystery and don't even pretend to want a solution. I think they really enjoy making things more complicated and even see a kind of liberation in the indefinite exploration; what can you say for people like that!