Conscious Entities

The Archive

2019 to 2021

Welcome to Conscious Entities, the Archive

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 A few posts have been omitted, and quite a few are rather out of date, but I still enjoy reading them.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

Unconscious and Conscious

1 January 2019

What if consciousness is just a product of our non-conscious brain, ask Peter Halligan and David A Oakley? But how could it be anything else? If we take consciousness to be a product of brain processes at all, it can only come from non-conscious ones. If consciousness had to come from processes that were themselves already conscious, we should have a bit of a problem on our hands. Granted, it is in one sense remarkable that the vivid light of conscious experience comes from stuff that is at some level essentially just simple physical matter - it might even be that incredulity over that is an important motivator for panpsychists, who find it easier to believe that everything is at least a bit conscious. But most of us take that gap to be the evident core fact that we're mainly trying to explain when we debate consciousness.

Of course, Halligan and Oakley mean something more than that. Their real point is a sceptical, epiphenomenalist one; that is, they believe consciousness is ineffective. All the decisions are really made by unconscious processes; consciousness notices what is happening and, like an existentialist, claims the act as its own after the fact (though of course the existentialist does not do it automatically). There is of course a lot of evidence that our behaviour is frequently influenced by factors we are not consciously aware of, and that we happily make up reasons for what we have done which are not the true ones. But 'frequently' is not 'invariably' and in fact there seem to be plenty of cases where, for example, our conscious understanding of a written text really does change our behaviour and our emotional states. I would find it pretty hard to think that my understanding of an email with important news from a friend was somehow not conscious, or that my conscious comprehension was irrelevant to my subsequent behaviour. That is the kind of unlikely stuff that the behaviourists ultimately failed to sell us. Halligan and Oakley want to go quite a way down that same unpromising path, suggesting that it is actually unhelpful to draw a sharp distinction between conscious and unconscious. They will allow a kind of continuum, but to me it seems clear that there is a pretty sharp distinction between the conscious, detached plans of human beings and the instinct-driven, environment-controlled behaviour of animals, one it is unhelpful to blur or ignore.

One distinction that I think would be helpful here is between conscious and what I'll call self-conscious states. If I make a self-conscious decision I'm aware of making it; but I can also just make the decision; in fact, I can just act. In my view, cases where I just make the decision in that unreflecting way may still be conscious; but I suspect that Halligan and Oakley (like Higher Order theorists) accept only self-conscious decisions as properly conscious ones.

Interesting to compare the case put by Peter Carruthers in Scientific American recently; he argues that the whole idea of conscious thought is an error. He introduces the useful idea of the Global Workspace proposed by Bernard Baars and others; a place where data from different senses can be juggled and combined. To be in the workspace is to be among the contents of consciousness, but Carruthers believes only sensory items get in there. He's happy to allow bits of 'inner speech' or mental visualisation, deceptive items that mislead us about our own thoughts; but he won't allow completely abstract stuff (again, you may see some resemblance

to the ‘mentalistic’ entities disallowed by the behaviourists). I don’t really know why not; if abstractions never enter consciousness or the workspace, how is it that we’re even talking about them?

Carruthers thinks our ‘Theory Of Mind’ faculty misleads us; as a quick heuristic it’s best to assume that others know their own mental states accurately, and so, it’s natural for us to think that we do too: that we have direct access and cannot be wrong about whether we, for example, feel hungry. But he thinks we know much less than we suppose about our own motives and mental states. On this he seems a little more moderate than Halligan and Oakley, allowing that conscious reflection can sometimes affect what we do.

I think the truth is that our mental, and even our conscious processes, are in fact much more complex and multi-layered than these discussions suggest. Let’s consider the causal efficacy of conscious thought in simple arithmetic. When I add two to two in my head and get four, have the conscious thoughts about the sum caused the conscious thought of the answer, or was there an underlying non-conscious process which simply waved flags at a couple of points?

Well, I certainly can do the thing epiphenomenally. I can call up a mental picture of the written characters, for example, and step through them one by one. In that case the images do not directly cause each other. If I mentally visualise two balls and then another two balls and then mentally count them, perhaps that is somewhat different? Can I think of two abstractly and then notice conceptually that its reduplication is identical with the entity ‘four’? Carruthers would deny it, I think, but I’m not quite sure. If I can, what causal chain is operating? At this point it becomes clear that I really have little idea of how I normally do arithmetic, which I suppose scores a point for the sceptics. The case of two plus two being four is perhaps a bad example, because it is so thoroughly remembered I simply replay it, verbally, visually, implicitly, abstractly or however I do it. What if I were multiplying 364 by 5? The introspective truth seems to be that I do something akin to running an algorithm by hand. I split the calculation into separate multiplications, whose answer I mainly draw direct from memory, and then I try to remember results and add them, again usually relying on remembered results. Does my thinking of four times five and recalling that the answer is twenty mean there was a causal link between the conscious thought and the conscious result? I think there may be such a link, but frustratingly if I use my brain in a slightly different way there may not be a direct one, or there may be a direct one which is not of the appropriate kind (because, say, the causal link is direct but irrelevant to the mathematical truth of the conclusion).

Having done all that calculation, I realise that since I’m multiplying by five, I could have multiplied by ten instead, which can be done by simply adding a zero (is that done visually - is it necessarily done visually? Some people cannot conjure up mental images at all - do they find multiplying by ten more difficult?) and halving the result, which is also relatively easy. Where did that little tactic come from? Did I think of it consciously, and was its arrival in my mind in reportable condition causally derived from my wondering about how best to do the sum (in words, or not in words?) or was it thrust into a kind of mental in-tray (rather spookily) by an unconscious part of my brain which has been vaguely searching around for any good tricks for the last couple of minutes? I don’t know. Unfortunately I think it could have happened in any one of a dozen different ways, some of which probably involve causally effective conscious states while others may not.

In the end the biggest difference between me and the sceptics may come down to what we are prepared to call conscious; they only want the states I’m calling self-conscious. Suppose we

take it that there is indeed a metaphorical or functional space where mental items become ‘available’ (in some sense I leave unclarified) to influence our behaviour. It could indeed be a Global Workspace, but need not commit us to all the details of that theory. Then the sceptics allow at most those items actually within the space to be conscious, while I would allow anything capable of entering it. My intuition here is that the true borderline falls, not between those mental items I’m aware of and those I merely have, but between those I could become aware of if my attention were suitably directed, and those I could never extract from the regions where they are processed, however I thought about it. When I play tennis, I may consciously plan a strategy, but I also consciously choose where to send the ball on the spur of the moment; that is not normally a self-conscious decision, but if I stopped to review it it easily could become one - whereas the murky Freudian reasons why I particularly want to win the game cannot be easily accessed (without lengthy therapy at any rate) and the processes my visual cortex used to work out the ball’s position are forever denied me.

My New Year resolution is to give up introspection before I vanish into some neo-Humean abyss of self-doubt.

Entangled Consciousness

9 January 2019

Could the answer be quantum physics after all? Various people have suggested that the mystery of consciousness might turn out to be explicable in terms of quantum physics; most notably we have the theory championed by Roger Penrose and Stuart Hameroff, which suggests that as-yet unknown quantum mechanics might be going on in the microtubules of neural cells. Atai Barkai has kindly shared with me his draft paper which sets out a different take on the matter.

Barkai's target is subjective experience, or what he defines as the consciousness instance. We may be conscious of things in the world out there, but that kind of consciousness always involves an element of inference, explicit or not; the consciousness instance is free of those uncertainties, consisting only of the direct experience of which we can be certain. Some hardy sceptics would deny that there is anything in consciousness that we can be that sure of, but I think it is clear enough what distinction Barkai is making and that it is, as it were, first order experience he is aiming at.

The paper puts a complex case very carefully, but flinging caution to the winds I believe the gist goes like this. Typically, those who see a problem with qualia and subjective experience think it lies outside the account given by physics, which arguably therefore needs some extension. Barkai agrees that subjectivity is outside the classical account and that a more capacious concept of the universe is needed to accommodate it; but he thinks that quantum entanglement can already, properly read, provide what is needed.

It's true that philosophical discussions generally either treat classical physics as being the whole subject or ignore any implications that might arise from the less intuitive aspects of quantum physics. Besides entanglement, Barkai discusses free will and its possible connection with the indeterminism of quantum physics. If there really is indeterminism in quantum physics; on this and other points I simply don't have a clear enough grasp of the science to tackle the questions effectively (better informed comments would be welcome).

Entanglement, as I understand it, means that the states of two particles (or in principle, larger bodies) may be strongly connected in ways that do not depend on normal classical interaction and are not affected by distance. This means that information can in theory be transferred any distance, instantly and infallibly, which opens up the theoretical possibility of quantum computing, delivering the instant solution to computable problems of any finite size. Whether the brain might be doing this kind of thing is a question which Barkai leaves as an interesting speculation.

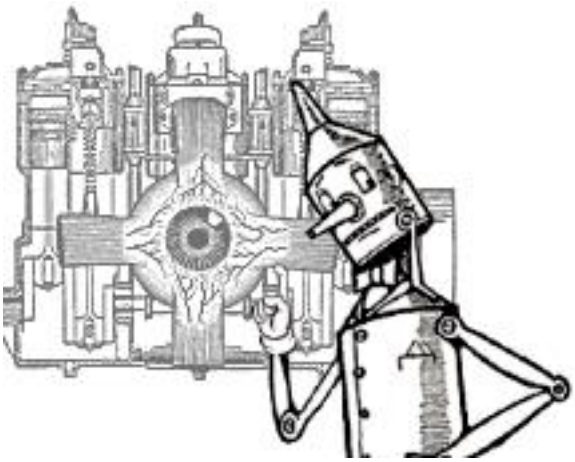
The main problem for me in the idea that entanglement explains subjectivity is understanding intuitively how that could be so. Entanglement seems to offer the possibility that there might be more information in our mental operations than classical physics could account for; that feels helpful, but does it explain the qualitative difference between mere dead data and the 'something it is like' of subjectivity? I don't reject the idea, but I don't think I fully grasp it either.

There is also a more practical objection. Quantum interactions are relevant at a microscopic level, but as we ascend to normal scales, it is generally assumed that wave forms collapse and we slip quickly back into a world where classical physics reigns. It is generally thought that neurons, let alone the whole brain, are just too big and hot for quantum interactions to have any relevance. This is why Penrose looks for new quantum mechanics in tiny microtubules rather than resting on what known quantum mechanics can provide.

As I say, I'm not really competent to assess these issues, but there is a neatness and a surprising novelty about Barkai's take that suggests it merits further discussion.

Explaining to Humans

18 January 2019



We've discussed a few times recently how deep learning systems are sometimes inscrutable. They work out how to do things for themselves, and there may be no way to tell what method they are using. I've suggested that it's worse than that; they are likely to be using methods which, though sound, are in principle incomprehensible to human beings. These unknowable methods may well be a very valuable addition to our stock of tools, but without understanding them we can't be sure they won't fail; and the worrying thing about that is that unlike humans, who often make small or recoverable mistakes, these systems have a way of failing that is sudden, unforgiving, and catastrophic.

So I was interested to see this short interview with Been Kim, who works for Google Brain (is there a Google Eyes?) has the interesting job of building a system that can explain to humans what the neural networks are up to.

That ought to be impossible, you'd think. If you use the same deep learning techniques to generate the explainer, you won't be able to understand how it works, and the same worries will recur on another level. The other approach is essentially the old-fashioned one of directly writing an algorithm to do the job; but to write an explicit algorithm to explain system X you surely need to understand system X to begin with?

I suspect there's an interesting general problem here about the transmission of understanding. Perhaps data can be transferred, but understanding just has to happen, and sometimes, no matter how helpfully the relevant information is transmitted, understanding just fails to occur (I feel teachers may empathise with this).

Suppose the explainer is in place; how do we know that its explanations are correct? We can see over time whether it successfully picks out systems that work reliably and those that are going to fail, but that is only ever going to be a provisional answer. The risk of sudden unexpected failure remains. For real certainty, the only way is for us to validate it by understanding it, and that is off the table.

So is Been Kim wasting her time on a quixotic project - perhaps one that Google has cynically created to reassure the public and politicians while knowing the goal is unattainable? No; her goal is actually a more modest, negative one. Her interpreter is not meant to provide an assurance that a given system is definitely reliable; rather, it is supposed to pick out one's that are definitely dodgy; and this is much more practical. After all, we may not always understand how a given system executes a particular complex task, but we do know in general how neural networks and deep learning work. We know that the output decisions come from factors in the input data, and the interpreter ought to be able to tell us what factors are being taken into account. Then, using the unique human capacity to identify relevance, we may be able to spot some duds - cases where the system is using a variable that tracks the relevant stuff only

unreliable, or where there was some unnoticed problem with the corpus of examples the system learnt from.

Is that OK? Well, in principle there's the further risk that the system is actually cleverer than we realise; that it is using features (perhaps very complex ones) that actually work fine, but which we're too dim to grasp. Our best reassurance here is again understanding; if we can see how things seem to be working, we have to be very unlucky to hit a system which is actually superior but just happens, in all the examined cases, to look like a dodgy one. We may not always understand the system, but if we understand something that's going wrong, we're probably on firm ground.

Of course, weeding out egregiously unreliable systems does not solve the basic problem of efficient but inscrutable systems. Without accusing Google of a cunning sleight of hand after all, I can well imagine that the legislators and bureaucrats who are gearing up to make rules about this issue might mistake interpreter systems like Kim's for a solution, require them in all cases, and assume that the job is done and dusted…

Mapping the Connectome

3 February 2019

Could Artificial Intelligence be the tool that finally allows us understand the natural kind?

We've talked before about the possibility; this Nautilus piece explains that scientists at the Max Planck Institute Of Neurobiology have come up with a way of using 'neural' networks to map, well, neural networks. There has been a continued improvement in our ability to generate images of neurons and their connections, but using those abilities to put together a complete map is a formidable task; the brain has often been described as the most complex object in the universe and drawing up a full schematic of its connections is potentially enough work for a lifetime. Yet that map may well be crucial; recently the idea of a 'connectome' roughly equivalent to the 'genome' has become a popular concept, one that suggests such a map may be an essential part of understanding the brain and consciousness. The Max Planck scientists have developed an AI, called 'SyConn' which tu4ns images into maps automatically with very high accuracy. In principle I suppose this means we could have a complete human connectome in the not-too-distant future.

How much good would it do us, though? It can't be bad to have a complete map, but there are three big problems. The first is that we can already be pretty sure that connections between neurons are not the whole story. Neurons come in many different varieties, ones that pretty clearly seem to have different functions - but it's not really clear what they are. They operate with a vast repertoire of neurotransmitters, and are themselves pretty complex entities that may have genuine computational properties all on their own. They are supported by a population of other, non-networked cells that may have a crucial role in how the overall system works. They seem to influence each other in ways that do not require direct connection; through electromagnetic fields, or even through DNA messages. Some believe that consciousness is not entirely a matter of network computation anyway, but resides in quantum or electrical fields; certainly the artificial networks that were originally inspired by biological neurology seem to behave in ways that are useful but quite distinct from those of human cognition. Benjamin Libet thought that if only he could do the experiment, he might be able to demonstrate that a sliver of the brain cut out from its neighbouring tissue but left in situ would continue to do its job. That, surely, is going too far; the brain didn't grow all those connections with such care for nothing. The connectome may not be the full story, but it has to be important.

The second problem, though, is that we might be going at the problem from the wrong end. A map of the national road network tells us some useful things about trade, but not what it is or, in the deeper sense, how it works. Without those roads, trade would not occur in anything like the same form; blockages and poor connections may hamper or distort it, and in regions isolated by catastrophic, um, road lesions, trade may cease altogether. Of course to understand things properly we should need to know that there are different kinds of vehicle doing different jobs on those roads, that places may be connected by canals and railways as well as roads, and so on. But more fundamentally, if we start with the map, we have no idea what trade really is. It is, in fact, primarily driven and determined by what people want, need, and believe, and if we fall into

the trap of thinking that it is wholly determined by the availability of trucks, goods, and roads we shall make a serious mistake.

Third, and perhaps it's the same problem in different clothes, we still don't have any clear and accepted idea of how neurology gives rise to consciousness anyway. We're not anywhere near being able to look at a network and say, yup, that is (or could be) conscious, if indeed it is ever possible to reach such a conclusion.

So do we really even want a map of the connectome? Oh yes…

Consciousness- Where Are We?

9 February 2019

Interesting to see the review of progress and prospects for the science of consciousness produced by Matthias Michel and others, and particularly the survey that was conducted in parallel. The paper discusses funding and other practical issues, but we're also given a broad view of the state of play, with the survey recording broadly optimistic views and interestingly picking out Global Workspace proposals as the most favoured theoretical approach. However, consciousness science was rated less rigorous than other fields (which I suppose is probably attributable to the interdisciplinary character of the topic and in particular the impossibility of avoiding 'messy' philosophical issues).

Michel suggests that the scientific study of consciousness only really got established a few decades ago, after the grip of behaviourism slackened. In practical terms you can indeed start in the mid twentieth century, but that actually overlooks the early structuralist psychologists a hundred years earlier. Wundt is usually credited as the first truly scientific psychologist, though there were others who adopted the same project around the same time. The investigation of consciousness (in the sense of awareness) was central to their work, and some of their results were of real value. Unfortunately, their introspective methods suffered a fatal loss of credibility, which is what precipitated the extreme reaction against consciousness represented by behaviourism, which eventually suffered an eclipse of its own, leaving the way clear for something like a fresh start, the point Michel takes as the real beginning. I think the longer history is worth remembering because it illustrates a pattern in which periods of energetic growth and optimism are followed by dreadful collapses, a pattern still recognisable in the field, perhaps most obviously in AI, but also in the outbreaks of enthusiasm followed by scepticism that have affected research based on fMRI scanning, for example.

In spite of the 'winters' affecting those areas, it is surely the advances in technology that have been responsible for the genuine progress recognised by respondents to the survey. Whatever our doubts about scanning, we undeniably know a lot more about neurology now than we did, even if that sometimes serves to reveal new mysteries, like the uncertain function of the newly-discovered 'rosehip' neurons. Similarly, though we don't have conscious robots (and I think almost everyone now has a more mature sense of what a challenge that is), the project of Artificial General Intelligence has reshaped our understanding. I think, for example, that Daniel Dennett is right to argue that exploration of the wider Frame Problem in AI is not just a problem for computer scientists, but tells us about an important aspect of the human mind we had never really noticed before - its remarkable capacity for dealing with relevance and meaning, something that is to the fore in the fascinating recent development of the pragmatics of language, for example.

I was not really surprised to see the Global Workspace theory achieving top popularity in the survey (Bernard Baars perhaps missing out on a deserved hat-tip here); it's a down-to-earth approach that makes a lot of sense and is relatively easily recruited as an ally of other theoretical insights. That said, it has been around for a while without much in the way of a

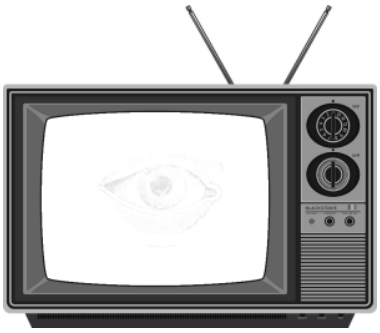
breakthrough. It was not that much more surprising to see Integrated Information also doing well, though rated higher by non-professionals (Michel shrewdly suggests that they may be especially impressed by the relatively complex mathematics involved).

However, the survey only featured a very short list of contenders which respondents could vote for. The absence of illusionism and quantum theories is acknowledged; myself I would have included at least two schools of sceptical thought; computationalism/functionalism and other qualia sceptics - though it would be easy to lengthen the list. Most surprising, perhaps, is the absence of panpsychism. Whatever you think about it (and regulars will know I'm not a fan), it's an idea whose popularity has notably grown in recent years and one whose further development is being actively pursued by capable adherents. I imagine the absence of these theories, and others such as mysterianism and the externalism doughtily championed by Riccardo Manzotti and others, is due to their being relatively hard to vindicate neurologically - though supporters might challenge that. Similarly, its robustly scientific neurological basis must account for the inclusion of 'local recurrence' - is that the same as recurrent processing?

It's only fair to acknowledge the impossibility of coming up with a comprehensive taxonomy of views on consciousness which would satisfy everyone. It would be easy to give a list of twenty or more which merely generated a big argument. (Perhaps a good thing to do, then?)

Consciousness Without Content

21 February 2019



Is there such a thing as consciousness without content? If so, is that minimal, empty consciousness, in fact, the constant ground underlying all conscious states? Thomas Metzinger launched an investigation of this question in the third Carnap lecture of a year ago; there's a summary here (link updated to correct version) in a discussion paper, and a fully worked-up paper will appear next year (hat-tip to Tom Clark for drawing this to my attention). The current paper is exploratory in several respects. One possible result of identifying the hypothetical state of Minimal Phenomenal Experience (MPE) would be to facilitate the identification of neural correlates; Metzinger suggests we might look to the Ascending Reticular Arousal System (ARAS) but offers it only as a plausible placeholder which future research might set aside.

More philosophically, the existence of an underlying conscious state which doesn't represent anything would be a fatal blow to the view that consciousness is essentially representational in character. On that widely held view, a mental state that doesn't feature representation cannot in fact be conscious at all, any more than text that contains no characters is really text. The alternative is to think that consciousness is more like a screen being turned on; we see only (let's say) a blank white expanse, but the basic state, precondition to the appearance of images, is in place, and similarly MPE can be present without 'showing us' anything.

There's a danger here of getting trapped in an essentially irrelevant argument about the difference between representing nothing and not representing anything, but I think it's legitimate to preserve representationalism (as an option at least) merely by claiming that even a blank screen necessarily represents something, namely a white void. Metzinger prefers to suggest that the MPE represents "the hidden cause of the ARAS- signal". That seems implausible to me, as it seems to involve the unlikely idea that we all have constantly in mind a hidden thing most of us have never heard of.

Metzinger does a creditable job of considering evidence from mystic experience as well as dreamless sleep. There is considerable support here for the view that when the mind is cleared, consciousness is not lost but purified. Metzinger rightly points to some difficulties with taking this testimony on board. One is the likelihood of what he calls "theory contamination". Most contemplatives are deeply involved with mystic or scriptural traditions that already tell them what is to be expected. Second is the problem of pinning down a phenomenal experience with no content, which automatically renders it inexpressible or ineffable. Metzinger makes it clear that this is not any kind of magic or supra-scientific ineffability, just the practical methodological issue that there isn't, as it were, anything to be said about non-existent content. Third we have an issue Metzinger calls "performative self-contradiction". Reports of what you get when your mind is emptied make clear that the MPE is timeless, placeless, lacking sensory character, and so on. Metzinger is a little disgusted with this; if the experience was timeless, how do you know it happened last Tuesday between noon and ten past? People keep talking about brilliance and white light, which should not be present in a featureless experience!

Here I think he under-rates the power and indeed the necessity of metaphors. To describe lack of content we fall back on a metaphor of blindness, but to be in darkness might imply simple

failure of the eyes, so we tend to go for our being blinded by powerful light and the vastness of space. It's also possible that white is a default product of our neural systems, which when deprived of input are known to produce moiré patterns and blobs of light from nowhere. Here we are undoubtedly getting into the classic problems that affect introspection; you cannot have a cleared mind and at the same time be mentally examining your own phenomenal experience. Metzinger aptly likens these problems to trying to check whether the light in the fridge goes off when the door is closed (I once had one that didn't, incidentally; it gave itself away by getting warm and unhelpfully heating food placed near it). Those are real problems that have been discussed extensively, but I don't think they need stop the investigation. In a nutshell, William James was right to say that introspection must be retrospection; we examine our experiences afterwards. This perhaps implies that memory must persist alongside MPE, but that seems OK to me. Without expressing it in quite these terms, Metzinger reaches broadly similar conclusions.

Metzinger is mainly concerned to build a minimal model of the basic MPE, and he comes up with six proposed constraints, giving him in effect not a single MPE state but a 6-dimensional space. The constraints are as follows.

- PC1: Wakefulness: the phenomenal quality of tonic alertness.
- PC2: Low complexity of reportable content: an absence of high-level symbolic mental content (i.e., conceptual or propositional thought or mind wandering), but also of perceptual, sensorimotor, of emotional content (as in full-absorption episodes).
- PC3: Self-luminosity: a phenomenal property of MPE typically described as "radiance", "brilliance", or the "clear light" of primordial awareness.
- PC4: Introspective availability: we can sometimes actively direct introspective attention
- to MPE and we can distinguish possible states by the degree of actually ongoing access.
- PC5: Epistemicity; as MPE is an epistemic relation ("awareness-of",) if MPE is successfully introspected, then we would predict a distinct phenomenal character of epistemicity or subjective confidence.
- PC6: Transparency/opacity: like all other phenomenal representations, MPE can vary along a spectrum of opacity and transparency.

At first I feared this was building too much on a foundation not yet well established, but against that Metzinger could fairly ask how he could consolidate without building; what we have is acknowledged to be a sketch for now; and in fact there's nothing that looks obviously out of place to me.

For Metzinger this investigation of minimal experience follows on from earlier explorations of minimal self-awareness and minimal perspective; this might well be the most significant of the three, however. It opens the way to some testable hypotheses and, since it addresses "pure" consciousness offers a head-on attack route on the core problem of consciousness itself. Next year's paper is surely going to be worth a look.

Consciousness Not Needed

25 February 2019

Artificial General Intelligence - human-style consciousness in machines - is unnecessary, says Daniel Dennett, in an interesting piece that makes several unexpected points. His thinking seems to have moved on in certain respects, though I think the underlying optimism about digital cognition is still there.

He starts out by highlighting some dangers that arise even with the non-conscious kinds of AI we have already. Recent developments make it easy to fake very realistic video of recognisable people doing or saying whatever we want. We can imagine, Dennett says '… a return to analog film-exposed-to-light, kept in "tamper-proof" systems until shown to juries, etc.' Well, I think that scenario will remain imaginary. These are not completely new concerns; similar ones go back to the people who in the last century used to say 'the camera cannot lie' and were so often proven wrong. Actually the question of whether a piece of evidence is digital or analog is pretty well irrelevant; but its history, whether it could have been interfered with, that bit about tamper-proof containers, has always been a concern and will no doubt remain so in one form or another (I remember the special cassette recorders once used by the authorities to interview suspects, which automatically made a copy for the interviewee and could not erase or rewind the tape).

I think his concerns have a more solid foundation though, when he goes on to say that there is now some danger of people mistaking simple AI for the kind of conscious entity they can trust. People do sometimes seem willing to be convinced rather easily that a machine has a mind of its own. That tendency in itself is also not exactly new, going back to Weizenbaum's simple chat program Eliza (as Dennett says); but these days serious discussion is beginning about topics like robot rights and robot morality. No reason why we shouldn't discuss them - I think we should - but the idea that they are issues of immediate practical concern seems radically premature. Still, I'm not that worried. It's true that some people will play along with a chatbot in the Loebner contest, or pretend Siri or Alexa is a thinking being, but I think anyone who is even half-serious about it can easily tell the difference. Dennett suggests we might need licensed operators trained to be unattractively hard-nosed and unsympathetic to AIs ('an ugly talent, reeking of racism' - !), but I don't think it's going to be that bad.

Dennett emphasises that current AI lacks true agency and calls on the creators of new systems to be more upfront about the fact that their humanising touches are fakery and even 'false advertising'. I have the impression that Dennett would once have considered agency, as a concept, a little fuzzy round the edges, a matter of explanatory stances and optimality rather than a clear reality whose sharp edges needed to be strongly defended. Years ago he worked with Ray Brooks on Cog, a deliberately humanoid robot they hoped would attain consciousness (it all seemed so easy back then…) and my impression is that the strategy then had a large element of 'fake it till you make it'. But hey, I wouldn't criticise Dennett for allowing his outlook to develop in the light of experience.

On to the two main points. Dennett says we don't need conscious AI because there is plenty of natural human consciousness around; what we need is intelligent tools, somewhat like oracles perhaps (given the history of real oracles that might be a dubious comparison - is there a single example of an oracle that was both clear and helpful? In The Golden Ass there's a pair of soothsayers who secretly give the same short verse to every client; in the end they're forced to give up the profitable business, not by being rumbled, but by sheer boredom.)

I would have thought that there were jobs a conscious AI could do for us. Consciousness allows us to reflect on future and imagined contingencies, spot relevant factors from scratch, and rise above the current set of inputs. Those ought to be useful capacities in a lot of routine planning and management (I can't help thinking they might be assets for self-driving vehicles); yes, humans could do it all, but it's boring, detailed, and takes too long. I reckon, if there's a job you'd give a slave, it's probably right for a robot.

The second main point is that we ought to be wary of trusting conscious AIs because they will be invulnerable. Putting them in jail is meaningless because they live in boxes anyway; they can copy themselves and download backups, so they don't die; unless we build in some pain function, there are really no sanctions to underpin their morality.

This is interesting because Dennett by and large assumes that future conscious AIs would be entirely digital, made of data; but the points he makes about their immortality and generally Platonic existence implicitly underline how different digital entities are from the one-off reality of human minds. I've mentioned this ontological difference before, and it surely provides one good reason to hesitate before assuming that consciousness can be purely digital. We're not just data, we're actual historical entities; what exactly that means, whether something to do with Meinongian distinctions between existence and subsistence, or something else entirely, I don't really think anyone knows, frustrating as that is.

Finally, isn't it a bit bleak to suggest that we can't trust entities that aren't subject to the death penalty, imprisonment, or other punitive sanctions? Aren't there other grounds for morality? Call me Pollyanna, but I like to think of future conscious AIs proving irrefutably for themselves that *virtue is its own reward*.

Do We Need Ethical AI?

8 March 2019

Amanda Sharkey has produced a very good paper on robot ethics which reviews recent research and considers the right way forward — it's admirably clear, readable and quite wide-ranging, with a lot of pithy reportage. I found it comforting in one way, as it shows that the arguments have had a rather better airing to date than I had realised.

To cut to the chase, Sharkey ultimately suggests that there are two main ways we could respond to the issue of ethical robots (using the word loosely to cover all kinds of broadly autonomous AI). We could keep on trying to make robots that perform well, so that they can safely be entrusted with moral decisions; or we could decide that robots should be kept away from ethical decisions. She favours the latter course.

What is the problem with robots making ethical decisions? One point is that they lack the ability to understand the very complex background to human situations. At present they are certainly nowhere near a human level of understanding, and it can reasonably be argued that the prospects of their attaining that level of comprehension in the foreseeable future don't look that good. This is certainly a valid and important consideration when it comes to, say, military kill-bots, which may be required to decide whether a human being is hostile, belligerent, and dangerous. That's not something even humans find easy in all circumstances. However, while absolutely valid and important, it's not clear that this is a truly ethical concern; it may be better seen as a safety issue, and Sharkey suggests that that applies to the questions examined by a number of current research projects.

A second objection is that robots are not, and may never be, ethical agents, and so lack the basic competence to make moral decisions. We saw recently that even Daniel Dennett thinks this is an important point. Robots are not agents because they lack true autonomy or free will and do not genuinely have moral responsibility for their decisions.

I agree, of course, that current robots lack real agency, but I don't think that matters in the way suggested. We need here the basic distinction between good people and good acts. To be a good person you need good motives and good intentions; but good acts are good acts even if performed with no particular desire to do good, or indeed if done from evil but confused motives. Now current robots, lacking any real intentions, cannot be good or bad people, and do not deserve moral praise or blame; but that doesn't mean they cannot do good or bad things. We will inevitably use moral language in talking about this aspect of robot behaviour just as we talk about strategy and motives when analysing the play of a chess-bot. Computers have no idea that they are playing chess; they have no real desire to win or any of the psychology that humans bring to the contest; but it would be tediously pedantic to deny that they do 'really' play chess and equally absurd to bar any discussion of whether their behaviour is good or bad.

I do give full weight to the objection here that using humanistic terms for the bloodless robot equivalents may tend to corrupt our attitude to humans. If we treat machines inappropriately as human, we may end up treating humans inappropriately as machines. Arguably we can see this already in the arguments that have come forward recently against moral blame, usually framed as being against punishment, which sounds kindly, though it seems clear to me that they might

also undermine human rights and dignity. I take comfort from the fact that no-one is making this mistake in the case of chess-bots; no-one thinks they should keep the prize money or be set free from the labs where they were created. But there's undoubtedly a legitimate concern here.

That legitimate concern perhaps needs to be distinguished from a certain irrational repugnance which I think clearly attaches to the idea of robots deciding the fate of humans, or having any control over them. To me this very noticeable moral disgust which arises when we talk of robots deciding to kill humans, punish them, or even constrain them for their own good, is not at all rational, but very much a fact about human nature which needs to be remembered.

The point about robots not being moral persons is interesting in connection with another point. Many current projects use extremely simple robots in very simple situations, and it can be argued that the very basic rule-following or harm prevention being examined is different in kind from real ethical issues. We're handicapped here by the alarming background fact that there is no philosophical consensus about the basic nature of ethics. Clearly that's too large a topic to deal with here, but I would argue that while we might disagree about the principles involved (I take a synthetic view myself, in which several basic principles work together) we can surely say that ethical judgements relate to very general considerations about acts. That carries the optimistic implication that ethical reasoning, at least in terms of cognitive tractability, might not otherwise be different in kind from ordinary practical reasoning, and that as robots become more capable of dealing with complex tasks they might naturally tend to acquire more genuine moral competence to go with it.

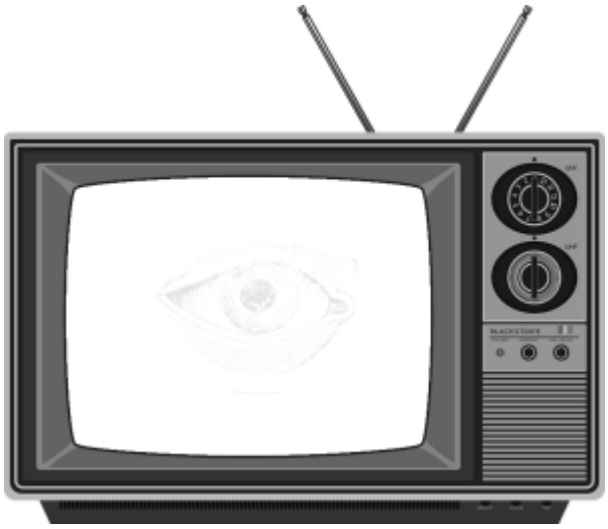
If robots are likely to acquire ethical competence as a natural by-product of increasing sophistication, then do we need to worry so much? Perhaps not, but the main reason for not worrying, in my eyes, is that truly ethical decisions are likely to be very rare anyway. The case of self-driving vehicles is often cited, but I think our expectations must have been tutored by all those tedious trolley problems; I've never encountered a situation in real life where a driver faced a clear cut decision about saving a bus load of nuns at the price of killing one fat man. If a driver follows the rule — 'try not to crash, and if crashing is unavoidable, try to minimise the impact' — I think almost all real cases will be adequately covered.

A point to remember is that we actually do often make rules about this sort of thing which a robot could follow without needing any ethical sense of its own, so long as its understanding of the general context was adequate. We don't have explicit rules about how many fat men outweigh a coachload of nuns just because we've never really needed them; if it happened every day we'd have debated it and made laws that people would have to know in order to pass their driving test. While there are no laws, even humans are in doubt and no-one can say definitively what the right choice is; so it's not logically possible to get too worried that the robot's choice in such circumstances would be wrong.

I do nevertheless have some sympathy with Sharkey's reservations. I don't think we should hold off from trying to create ethical robots though; we should go on, not because we want to use the resulting bots to make decisions, but because the research itself may illuminate ethical questions in ways that are interesting (a possibility Sharkey acknowledges). Since on my view we're probably never really going to need robots with a real ethical sense, and on the other hand if we did, there's a good chance they would naturally have developed the required competence, this looks to me like a case where we can have our cake and eat it (if that isn't itself unethical).

Consciousness without Content

29 February 2019



Is there such a thing as consciousness without content? If so, is that minimal, empty consciousness, in fact, the constant ground underlying all conscious states? Thomas Metzinger launched an investigation of this question in the third Carnap lecture of a year ago; there's a summary in a discussion paper, and a fully worked-up paper will appear next year (hat-tip to Tom Clark for drawing this to my attention). The current paper is exploratory in several respects. One possible result of identifying the hypothetical state of Minimal Phenomenal Experience (MPE) would be to facilitate the identification of neural correlates; Metzinger suggests we might look to the Ascending Reticular Arousal System (ARAS), but offers it only as a plausible place-holder which future research might set aside.

More philosophically, the existence of an underlying conscious state which doesn't represent anything would be a fatal blow to the view that consciousness is essentially representational in character. On that widely-held view, a mental state that doesn't feature representation cannot in fact be conscious at all, any more than text that contains no characters is really text. The alternative is to think that consciousness is more like a screen being turned on; we see only (let's say) a blank white expanse, but the basic state, precondition to the appearance of images, is in place, and similarly MPE can be present without 'showing us' anything.

There's a danger here of getting trapped in an essentially irrelevant argument about the difference between representing nothing and not representing anything, but I think it's legitimate to preserve representationalism (as an option at least) merely by claiming that even a blank screen necessarily represents something, namely a white void. Metzinger prefers to suggest that the MPE represents 'the hidden cause of the ARAS signal'. That seems implausible to me, as it seems to involve the unlikely idea that we all have constantly in mind a hidden thing most of us have never heard of.

Metzinger does a creditable job of considering evidence from mystic experience as well as dreamless sleep. There is considerable support here for the view that when the mind is cleared, consciousness is not lost but purified. Metzinger rightly points to some difficulties with taking this testimony on board. One is the likelihood of what he calls 'theory contamination'. Most contemplatives are deeply involved with mystic or scriptural traditions that already tell them what is to be expected. Second is the problem of pinning down a phenomenal experience with no content, which automatically renders it inexpressible or ineffable. Metzinger makes it clear that this is not any kind of magic or supra-scientific ineffability, just the practical methodological issue that there isn't, as it were, anything to be said about non-existent content. Third we have an issue Metzinger calls 'performative self-contradiction'. Reports of what you get when your mind is emptied make clear that the MPE is timeless, placeless, lacking sensory character, and so on. Metzinger is a little disgusted with this; if the experience was timeless,

how do you know it happened last Tuesday between noon and ten past? People keep talking about brilliance and white light, which should not be present in a featureless experience!

Here I think he under-rates the power and indeed the necessity of metaphors. To describe lack of content we fall back on a metaphor of blindness, but to be in darkness might imply simple failure of the eyes, so we tend to go for our being blinded by powerful light and the vastness of space. It's also possible that white is a default product of our neural systems, which when deprived of input are known to produce moiré patterns and blobs of light from nowhere. Here we are undoubtedly getting into the classic problems that affect introspection; you cannot have a cleared mind and at the same time be mentally examining your own phenomenal experience. Metzinger aptly likens these problems to trying to check whether the light in the fridge goes off when the door is closed (I once had one that didn't, incidentally; it gave itself away by getting warm and unhelpfully heating food placed near it). Those are real problems that have been discussed extensively, but I don't think they need stop the investigation. In a nutshell, William James was right to say that introspection must be retrospection; we examine our experiences afterwards. This perhaps implies that memory must persist alongside MPE, but that seems acceptable. Without expressing it in quite these terms, Metzinger reaches broadly similar conclusions.

Metzinger is mainly concerned to build a minimal model of the basic MPE, and he comes up with six proposed constraints, giving him in effect not a single MPE state but a 6-dimensional space. The constraints are as follows: PC1 (Wakefulness — the phenomenal quality of tonic alertness); PC2 (Low complexity of reportable content — an absence of high-level symbolic mental content, i.e. conceptual or propositional thought or mind wandering, but also of perceptual, sensorimotor, or emotional content); PC3 (Self-luminosity — a phenomenal property of MPE typically described as 'radiance', 'brilliance', or the 'clear light' of primordial awareness); PC4 (Introspective availability — we can sometimes actively direct introspective attention to MPE and distinguish possible states by the degree of actually ongoing access); PC5 (Epistemicity — as MPE is an epistemic relation, if MPE is successfully introspected, then we would predict a distinct phenomenal character of epistemicity or subjective confidence); and PC6 (Transparency/opacity — like all other phenomenal representations, MPE can vary along a spectrum of opacity and transparency).

At first I feared this was building too much on a foundation not yet well established, but against that Metzinger could fairly ask how he could consolidate without building; what we have is acknowledged to be a sketch for now; and in fact there's nothing that looks obviously out of place to me.

For Metzinger this investigation of minimal experience follows on from earlier explorations of minimal self-awareness and minimal perspective; this might well be the most significant of the three, however. It opens the way to some testable hypotheses and, since it addresses 'pure' consciousness, offers a head-on attack route on the core problem of consciousness itself. The forthcoming full paper is surely going to be worth a look.

Minds Within Minds

25 March 2019

Can there be minds within minds? I think not.

The train of thought I'm pursuing here started in a conversation with a friend (let's call him Fidel) who somehow manages to remain not only an orthodox member of the Church of England, but one who is apparently quite untroubled by any reservations, doubts, or issues about the theology involved. Now of course we don't all see Christianity the same way. Maybe Fidel sees it differently from me. For many people (I think) religion seems to be primarily a social organisation of people with a broadly similar vision of what is good, derived mainly from the teachings of Jesus. To me, and I suspect to most people who are likely to read this, it's primarily a set of propositions, whose truth, falsity, and consistency is the really important matter. To them it's a club, to us it's a theory. I reckon the martyrs and inquisitors who formed the religion, who were prepared to die or kill over formal assent to a point of doctrine, were actually closer to my way of thinking on this, but there we are.

Be that as it may, my friend cunningly used the problems (or mysteries) of his religion as a weapon against me. You atheists are so complacent, he said, you think you've got it all sorted out with your little clockwork science universe, but you don't appreciate the deep mysteries, beyond human understanding. There are more things in heaven and earth...

But that isn't true at all, I said. If you think current physics works like clockwork, you haven't been paying attention. And there are lots of philosophical problems where I have only reasonable guesses at the answer, or sometimes, even on some fundamental points, little idea at all. Why, I said injudiciously, I don't understand at all what reality itself even is. I can sort of see that to be real is to be part of a historical process characterised by causality, but what that really means and why there is anything like that, what the hell is really going on with it...? Ah, said Fidel, what a confession! Well, when you're ready to learn about reality, you know where to come...

I don't, though. The trouble is, I don't think Christianity really has any answers for me on this or many other metaphysical points. Maybe it's just my ignorance of theology talking here, but it seems to me just as Christianity tells us that people are souls and then falls largely silent on how souls and spirits work and what they are, it tells us that God made the world and withholds any useful details of how and what. I know that Buddhism and Taoism tell us pretty clearly that reality is an illusion; that seems to raise other issues but it's a respectable start. The clearest Christian answer I can come up with is Berkeley's idealism; that is, that to be real is to be within the mind of God; the world is whatever God imagines or believes it to be.

That means that we ourselves exist only because we are among the contents of God's mind. Yet we ourselves are minds, so that requires it to be true that minds can exist within minds (yes, at last I am getting to the point). I don't think a mind can exist within another mind. The simplest way to explain is perhaps as follows: a thought that exists within a mind, that was generated by that mind, belongs to that mind. So if I am sustaining another mind by my thoughts, all of its

thoughts are really generated by me, and of course they are within my mind. So they remain my thoughts, the secondary mind has none that are truly its own — and it doesn't really exist. In the same way, either God is thinking my thoughts for me — in which case I'm just a puppet — or my thoughts are outside his mind, in which case my reality is grounded in something other than the Divine mind.

That might help explain why God would give us free will, and so on; it looks as if Berkeley must have been perfectly wrong: in fact reality is exactly the quality possessed by those things that are outside God's mind. Anyway, my grip of theology is too weak for my thoughts on the matter to be really worth reading (so I owe you an apology); but the idea of minds within minds arises in AI-related philosophy too, perhaps in relation to Nick Bostrom's argument that we are almost certainly part of a computer simulation. That argument rests on the idea that future folk with advanced computing technology will produce perfect simulations of societies like their own, which will themselves go on to generate similar simulations, so that most minds, statistically, are likely to be simulated ones. If minds can't exist within other minds, might we be inclined to doubt that they could arise in mind-like simulations?

Suppose for the sake of argument that we have a conscious mind that is purely computational; its mind arises from the computations it performs. Why should such a program not contain, as some kind of subroutine, a distinct process that has the same mind-generating properties? I don't think the answer is obvious, and it will depend on your view of consciousness. For me it's all about recognition; a conscious mind is a process whose outputs are conditioned by the recognition of future and imagined entities. So I would see two alternatives; either the computational mind we supposed to exist has one locus of recognition, or two. If it has one, the secondary mind can only be a puppet; if there are two, then whatever the computational relationship, the secondary process is independent in a way that means it isn't truly within the primary mind.

That doesn't seem to give me the anti-Bostrom argument I thought might be there, and let's be honest, the notion of a 'locus of recognition' could possibly be attacked. If God were doing my thinking, I feel it would be a bit sharper than this...

Feeling Free

5 April 2019

Eddy Nahmias recently reported on a ground-breaking case in Japan where a care-giving robot was held responsible for agreeing to a patient's request for a lethal dose of drugs. Such a decision surely amounts to a milestone in the recognition of non-human agency; but fittingly for a piece published on 1 April, the case was in fact wholly fictional.

However, the imaginary case serves as an introduction to some interesting results from the experimental philosophy Nahmias has been prominent in developing. The research — and I take it to be genuine — aims not at clarifying the metaphysics or logical arguments around free will and responsibility, but at discovering how people actually think about those concepts.

The results are interesting. Perhaps not surprisingly, people are more inclined to attribute free will to robots when told that the robots are conscious. More unexpectedly, they attach weight primarily to subjective and especially emotional conscious experience. Free will is apparently thought to be more a matter of having feelings than it is of neutral cognitive processing.

Why is that? Nahmias offers the reasonable hypothesis that people think free will involves caring about things. Entities with no emotions, it might be, don't have the right kind of stake in the decisions they make. Making a free choice, we might say, is deciding what you want to happen; if you don't have any emotions or feelings you don't really want anything, and so are radically disqualified from an act of will. Nahmias goes on to suggest, again quite plausibly, that reactive emotions such as pride or guilt might have special relevance to the social circumstances in which most of our decisions are made.

I think there's probably another factor behind these results; I suspect people see decisions based on imponderable factors as freer than others. The results suggest, let's say, that the choice of a lover is a clearer example of free will than the choice of an insurance policy; that might be because the latter choice has a lot of clearly calculable factors to do with payments versus benefits. It's not unreasonable to think that there might be an objectively correct choice of insurance policy for me in my particular circumstances, but you can't really tell someone their romantic inclinations are based on erroneous calculations.

I think it's also likely that people focus primarily on interesting cases, which are often instances of moral decisions; those in turn often involve self-control in the face of strong desires or emotions.

Another really interesting result is that while philosophers typically see freedom and responsibility as two sides of the same coin, people's everyday understanding may separate the two. It looks as though people do not generally distinguish all that sharply between the concepts of being causally responsible (it's because of you it happened, whatever your intentions) and morally responsible (you are blameworthy and perhaps deserve punishment). So, although people are unwilling to say that corporations or unconscious robots have free will,

they are prepared to hold them responsible for their actions. It might be that people generally are happier with concepts such as strict liability than moral philosophers mainly are; or of course, we shouldn't rule out the possibility that people just tend to suffer some mild confusion over these issues.

Thought-provoking stuff, anyway, and further evidence that experimental philosophy is a tool we shouldn't reject.

Machines Like Me

7 May 2019

Ian McEwan's latest book *Machines Like Me* has a humanoid robot as a central character. Unfortunately, I don't think he's a terrifically interesting robot; he's not very different to a naïve human in most respects, except for certain unlikely gifts: an ability to discuss literature impressively and an ability to play the stock market with steady success. No real explanation for these superpowers is given; it's kind of assumed that direct access to huge volumes of information together with a computational brain just naturally make you able to do these things. I don't think it's that easy, though in fairness these feats only resemble the common literary trick where our hero's facility with languages or amazingly retentive memory somehow makes him able to perform brilliantly at tasks that actually require things like insight and originality.

The robot is called Adam; twenty-five of these robots have been created, twelve Adams and thirteen Eves, on the market for a mere £86,000 each. This doesn't seem to make much commercial sense; if these are prototypes you wouldn't sell them; if you're ready to market them you'd be gearing up to make thousands of them, at least. Surely you'd charge more, too — you could easily spend £86k on a fancy new car. But perhaps prices are misleading, because we are in an alternate world.

This is perhaps the nub of it all. The prime difference here is that in the world of the novel, Alan Turing did not die and was mainly responsible for a much faster development of computers and IT. Plausible humanoid robots have appeared in 1982. This seems to me an unhelpful contribution to the myth of Turing as 'Mr Computer'. It's sadly plausible that if he had lived longer he would have had more to contribute; but most likely in other mathematical fields, not in the practical development of the computer, where many others played key roles (as they did at Bletchley). If you ask me, John Von Neumann was more than capable of inventing computers on his own, and in fact in the real postwar world they developed about as fast as they could have done whether Turing was alive or not. McEwan nudges things along a bit more by having Tesla around to work on silicon chips (!) and brings Demis Hassabis back in time so he can be Turing's collaborator (Hassabis evidently doomed to work on machine learning whenever he's born). This is all a bit silly, but McEwan enjoys it enough to have advanced IT in Exocet missiles give victory to Argentina in the Falklands War, with consequences for British politics which he elaborates in the background of the story.

Turing appears in the novel, and I hate the way he's portrayed. One of McEwan's weaknesses, in my opinion, is his reverence for the British upper class, and here he makes Sir Alan into the sort of grandee he admires; a lordly fellow with a large house in North London who summons people when he wants information, dismisses them when he's finished, and hands out moral lectures. Obviously I don't know what Turing was really like, but to me his papers give the strong impression of an unassuming man of distinctly lower middle class origins; a far more pleasant person than the arrogant one we get in the book.

McEwan doesn't give us any great insight into how Adam comes to have human-like behaviour (and surely human-like consciousness). His fellow robots are prone to a sort of depression which leads them to a form of suicide; we're given the suggestion that they all find it hard to deal with human moral ambiguity, though it seems to me that humans in their position (enslaved to morally dubious idiots) might get a bit depressed too. As the novel progresses, Adam's robotic nature seems to lose McEwan's interest anyway, as a couple of very human plots increasingly take over the story.

McEwan got into trouble for speaking dismissively of science fiction; is Machines Like Me SF? On a broad reading I'd say why not? — but there is a respectable argument to be made for the narrower view. In my youth the genre was pretty well-defined. There were the great precursors: Jules Verne, H.G. Wells, and perhaps Mary Shelley, but SF was mainly the product of the American pulp magazines of the fifties and sixties, a vigorous tradition that gave rise to Asimov, Clarke, and Heinlein at the head of a host of others. That genre tradition is not extinct, upheld today by, for example, the beautiful stories of Ted Chiang.

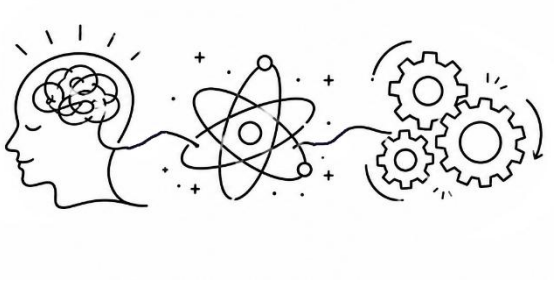
At the same time, though, SF concepts have entered mainstream literature in a new way. The Time Traveller's Wife, for example, obviously makes brilliant use of an SF concept, but does so in the service of a novel which is essentially a love story in the literary mainstream of books about people getting married which goes all the way back to Pamela. I think the new currency of SF ideas comes from the impact of computer games. The nerdy people who create computer games read SF and use SF concepts; but even non-nerdy people play the games, and in that way they pick up the ideas, so that novelists can now write about, say, a 'portal' and feel confident that people will get the idea pretty readily; a novel that has people reliving bits of their lives in an attempt to get them right (like The Seven Deaths of Evelyn Hardcastle) will not get readers confused the way it once would have done.

I think that among other things this wider dispersal of a sort of SF-aware mentality has led to a vast improvement in the robots we see in films and the like. It used to be the case that only one story was allowed: robots take over. Latterly films like Ex Machina or Her have taken a more sophisticated line; the TV series Westworld, though back with the take-over story, explicitly used ideas from Julian Jaynes.

So, I think we can accept that Machines Like Me stands outside the pure genre tradition but benefits from this wider currency of SF ideas. Alas, in spite of that we don't really get the focus on Adam's psychology that I should have preferred.

Minds, Matter and Mechanisms

21 May 2019



Will the mind ever be fully explained by neuroscience? There was a good discussion from the IAI, capably chaired by Barry C. Smith.

Raymond Tallis puts intentionality at the centre of the question of the mind (quite rightly, I think). Neuroscience will never explain meaning or the other forms of intentionality, so it will never tell us about essential aspects of the mind.

Susanna Martinez-Conde says we should not fear reductive explanation. Knowing how an illusion works can enhance our appreciation rather than undermining it. Our brains are designed to find meanings, and will do so even in a chaotic world.

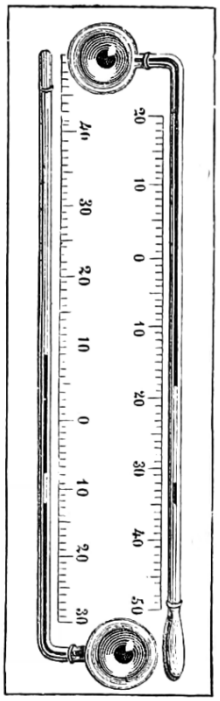
Markus Gabriel says we are not just a pack of neurons — trivially, because we are complete animals, but more interestingly because of the contents of our mind. He broadly agrees that intentionality is essential. London is not contained in my head, so aliens could not decipher from my neurons that I was thinking I was in London. He adds the concept of Geist — the capacity to live according to a conception of ourselves as a certain kind of being — which is essential to humanity, but relies on our unique mental powers.

Martinez-Conde points out that we can have the experience of being in London without in fact being there; Tallis dismisses such 'brain in a vat' ideas, arguing that for the brain to do that it must have had real experiences and there must be scientists controlling what happens in the vat. The mind is irreducibly social.

My sympathies are mainly with Tallis, but against him it can be pointed out that while neuroscience has no satisfactory account of intentionality, he hasn't got one either. While the subject remains a mystery, it remains possible that a remarkable new insight that resolves it all will come out of neuroscience. The case against that possibility, I think, rests mainly on a sense of incredulity: the physical is just not the sort of thing that could ever explain the mental. We find this in Brentano of course, and perhaps as far back as Leibniz's mill, or in the Cartesian point that mental things have no extension. But we ought to admit that this incredulity is really just an intuition, or if you like, a failure to be able to imagine. It puzzles me sometimes that numbers, those extensionless abstract concepts, can nevertheless drive the behaviour of a computer. But surely it would be weird to say they don't, or that how computers do arithmetic must remain an unfathomable mystery.

Degrees of Consciousness

25 June 2019



An interesting blog post by William Lycan gives a brisk treatment of the interesting question of whether consciousness comes in degrees, or is the kind of thing you either have or don't. In essence, Lycan thinks the answer depends on what type of consciousness you're thinking of. He distinguishes three: basic perceptual consciousness, 'state consciousness' where we are aware of our own mental state, and phenomenal consciousness. In passing, he raises interesting questions about perceptual consciousness. We can assume that animals, broadly speaking, probably have perceptual, but not state consciousness, which seems primarily if not exclusively a human matter. So what about pain? If an animal is in pain, but doesn't know it is in pain, does that pain still matter?

Leaving that one aside as an exercise for the reader, Lycan's answer on degrees is that the first two varieties of consciousness do indeed come in degrees, while the third, phenomenal consciousness, does not. Lycan gives a good ultra-brief summary of the state of play on phenomenal consciousness. Some just deny it (that represents a 'desperate lunge' in Lycan's view); some, finding it undeniable, lunge the other way — or perhaps fall back — by deciding that materialism is inadequate and that our metaphysics must accommodate irreducibly mental entities. In the middle are all the people who offer some partial or complete explanation of phenomenal consciousness. The leading view, according to Lycan, is something like his own interesting proposal that our introspective categorisation of experience cannot be translated into ordinary language; it's the untranslatability that gives the appearance of ineffability. There is a fourth position out there beyond the reach of even the most reckless lunge, which is panpsychism; Lycan says he would need stronger arguments for that than he has yet seen.

Getting back to the original question, why does Lycan think the answer is, as it were, 'yes, yes, no'? In the case of perceptual consciousness, he observes that different animals perceive different quantities of information and make greater or lesser numbers of distinctions. In that sense, at least, it seems hard to argue against consciousness occurring in degrees. He also thinks animals with more senses will have higher degrees of perceptual consciousness. He must, I suppose, be thinking here of the animal's overall, global state of consciousness, though I took the question to be about, for example, perception of a single light, in which case the number of senses is irrelevant (though I think the basic answer remains correct).

On state consciousness, Lycan argues that our perception of our mental states can be dim, vivid, or otherwise varied in degree. There's variation in actual intensity of the state, but what he's mainly thinking of is the degree of attention we give it. That's surely true, but it opens up a couple of cans of worms. For one thing, Lycan has already argued that perceptual states come in degrees by virtue of the amount of information they embody; now state consciousness — which is consciousness of a perceptual state — can also vary in degree because of the level of attention paid to the perceptual state. That in itself is not a problem, but to me it implies that the variability of state consciousness is really at least a two-dimensional matter. The second question is, if we can invoke attention when it comes to state consciousness, should we not

also be invoking it in the case of perceptual consciousness? We can surely pay different degrees of attention to our perceptual inputs. More generally, aren't there other ways in which consciousness can come in degrees? What about, for example, an epistemic criterion, i.e. how certain we feel about what we perceive? What about the complexity of the percept, or of our conscious response?

Coming to phenomenal consciousness, the brevity of the piece leaves me less clear about why Lycan thinks it alone fails to come in degrees. He asserts that wherever there is some degree of awareness of one's own mental state, there is something it's like for the subject to experience that state. But that's not enough; it shows that you can have no phenomenal consciousness or some, but not that there's no way the 'some' can vary in degree. Maybe sometimes there are two things it's like? Lycan argued that perceptual consciousness comes in degrees according to the quantity of information; he didn't argue that we can have some information or none, and that therefore perceptual consciousness is not a matter of degree.

It is unfortunately very difficult to talk about phenomenal experience. Typically, in fact, we address it through a sort of informal twinning. We speak of red quale, though the red part is really the objective bit that can be explained by science. It seems to me a natural prima facie assumption that phenomenal experience must 'inherit' the variability of its objective counterparts. Lycan might say that, even if that were true, it isn't what we're really talking about. But I remain to be convinced that phenomenal experience cannot be categorised by degree according to some criteria.

Neuromorality

15 July 2019

> *'In 1989 I was invited to go to Los Angeles in response to a request from the Dalai Lama, who wished to learn some basic facts about the brain.'*

Besides being my own selection for 'name drop of the year', this remark from Patricia Churchland's new book Conscience perhaps tells us that we are not dealing with someone who suffers much doubt about their own ability to explain things. That's fair enough; if we weren't radically overconfident about our ability to answer difficult questions better than anyone else, it's probable no philosophy would ever get done. And Churchland modestly goes on to admit to asking the Buddhists some dumb questions ('What's your equivalent of the Ten Commandments?'). Alas, I think some of her views on moral philosophy might benefit from further reflection.

Her basic proposition is that human morality is a more complex version of the co-operative and empathetic behaviour shown by various animals. There are some interesting remarks in her account, such as a passage about human scrupulosity, but she doesn't seem to me to offer anything distinctively new in the way of a bridge between mere co-operation and actual ethics. There is, surely, a gulf between the two which needs bridging if we are to explain one in terms of the other. No doubt it's true that some of the customs and practices of human beings may have an inherited, instinctive root; and those practices in turn may provide a relevant backdrop to moral behaviour. Not morality itself, though. It's interesting that a monkey fobbed off with a reward of cucumber instead of a grape displays indignation, but we don't get into morality until we ask whether the monkey was right to complain — and why.

Churchland never accepts that. She suggests that morality is a vaguely defined business; really a matter of a collection of rules and behaviours that a species or a community has cobbled together from pragmatic adaptations, whether through evolution or culture (quite a gulf there, too). She denies that there are any deep principles involved; we simply come to feel, through reinforcement learning and imitation, that the practices of our own group have a special moral quality. She groups moral philosophers into two groups: people she sees as flexible pragmatists (Aristotle, for some reason, and Hume) and rule-lovers (Kant and Jeremy Bentham). Unfortunately she treats moral rules and moral principles as the same, so advocates of moral codes like the Ten Commandments are regarded as equivalent to those who seek a fundamental grounding for morality, like Kant. Failure to observe this distinction perhaps causes her to give the seekers of principles unnecessarily short shrift. She rightly notes that there are severe problems with applying pure Utilitarianism or pure Kantianism directly to real life; but that doesn't mean that either theory fails to capture important ethical truths. A car needs wheels as well as an engine, but that doesn't mean the principle of internal combustion is invalid.

Another grouping which strikes me as odd is the way Churchland puts rationalists with religious believers (they must be puzzled to find themselves together) with neurobiology alone on the

other side. I wouldn't be so keen to declare myself the enemy of rational argument; but the rationalists are really the junior partners, it seems, people who hanker after the old religious certainties and deludedly suppose they can run up their own equivalents. Just as people who deny personhood sometimes seem to be motivated mainly by a desire to denounce the soul, I suspect Churchland mainly wants to reject Christian morality, with the baby of reasoned ethics getting thrown out along with the theological bathwater.

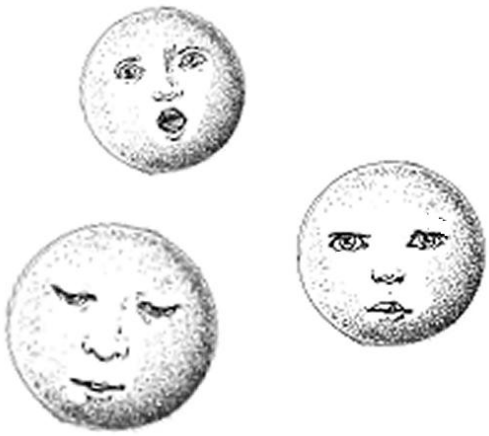
She seems to me particularly hard on Kant. She points out, quite rightly, that his principle of acting on rules you would be prepared to have made universal requires the rules to be stated correctly; a Nazi, she suggests, could claim to be acting according to consistent rules if those rules were drawn up in a particular way. We require the moral act to be given its correct description in order for the principle to apply. Yes; but much the same is true of Aristotle's Golden Mean, which she approves. 'Nothing to excess' is fine if we talk about eating or the pursuit of wealth, but it also, taken literally, means we should commit just the right amount of theft and murder; not too much, but not too little, either. Churchland is prepared to cut Aristotle the slack required to see the truth behind the defective formulation, but Kant doesn't get the same accommodation. Nor does she address the Categorical Imperative, which is a shame because it might have revealed that Kant understands the kind of practical decision-making she makes central, even though he says there's more to life than that.

Here's an analogy. Churchland could have set out to debunk physics in much the way she tackles ethics. She might have noted that beavers build dams and ants create sophisticated nests that embody excellent use of physics. Our human understanding of physics, she might have said, is the same sort of collection of rules of thumb and useful tips; it's just that we have so many more neurons, our version is more complex. Now some people claim that there are spooky abstract 'laws' of physics, like something handed down by God on tablets; invisible entities and forces that underlie the behaviour of material things. But if we look at each of the supposed laws we find that they break down in particular cases. Planes sail through the air, the Earth consistently fails to plummet into the Sun; so much for the 'law' of gravity! It's simply that the physics practices of our own culture come to seem almost magical to us; there's no underlying truth of physics. And worse, after centuries of experiment and argument, there's still bitter disagreement about the answers. One prominent physicist not so long ago said his enemies were 'not even wrong'! No-one, of course, would be convinced by that, and we really shouldn't be convinced by a similar case against ethical theory.

That implicit absence of moral truth is perhaps the most troubling thing about Churchland's outlook. She suggests Kant has nothing to say to a consistent Nazi, but I'm not sure what she can come up with, either, except that her moral feelings are different. Churchland wraps up with a reference to the treatment of asylum seekers at the American border, saying that her conscientious feelings are fired up. But so what? She's barely finished explaining why these are just feelings generated by training and imitation of her peer group. Surely we want to be able to say that mistreatment of children is really wrong?

Dismissing Materialism

9 August 2019



Eric Holloway gives a brisk and entertaining dismissal of all materialist theories of consciousness, boldly claiming that no materialist theory of consciousness is plausible. I'm not sure his coverage is altogether comprehensive, but let's have a look at his arguments. He starts out by attacking panpsychism.

One proposed solution is that all particles are conscious. But, in that case, why am I a human instead of a particle? The vast majority of conscious beings in the universe would be particles, and so it is most likely I'd be a particle and not any sort of organic life form.

It's really a bit of a straw man he's demolishing here. I'm not sure panpsychists are necessarily committed to the view that particles are conscious (I'm not sure panpsychists are necessarily materialists, either), but I've certainly never run across anyone who thinks that the consciousness of a particle and the consciousness of a human being would be the same. It would be more typical to say that particles, or whatever the substrate is, have only a faint glow of awareness, or only a very simple, perhaps binary kind of consciousness. Clearly there's then a need to explain how the simple kind of consciousness relates or builds up into our kind; not an easy task, but that's the business panpsychists are in, and they can't be dismissed without at least looking at their proposals.

Another solution is that certain structures become conscious. But a structure is an abstract entity and there is an untold infinite number of abstract entities.

This is perhaps Holloway's riposte; he considers this another variety of panpsychism, though as stated it seems to me to encompass a lot of non-panpsychist theories, too. I wholeheartedly agree that conscious beings are not abstract entities, an error which is easy to fall into if you are keen on information or computation as the basis of your theory. But it seems to me hard to fight the idea that certain structural (or perhaps I mean functional) properties instantiated in particular physical beings are what amounts to consciousness. On the one hand there's a vast wealth of evidence that structures in our brains have a very detailed influence on the content of our experiences. On the other, if there are no structural features, broadly described, that all physical instances of conscious entities have in common, it seems to me hard to avoid radical mysterianism. Even dualists don't usually believe that consciousness can simply be injected randomly into any physical structure whatever (do they?). Of course we can't yet say authoritatively what those structural features are.

Another option, says Holloway, is illusionism. But, if we are allowed to 'solve' the problem that way, all problems can be solved by denying them. Again, that is an unsatisfying approach that 'explains' by explaining away.

Empty dismissal of consciousness would indeed not amount to much, but again that isn't what illusionists actually say; typically they offer detailed ideas about why consciousness must be an

illusion and varied proposals about how the illusion arises. I think many would agree with David Chalmers that explaining why people do believe in consciousness is currently where some of the most interesting action is to be found.

Some say consciousness is an emergent property of a complex structure of matter. At what point is a structure complex enough to become conscious?

I agree that complexity alone is not enough, though some people have been attracted to the idea, suggesting that the Internet, for example, might achieve consciousness. A vastly more sophisticated form of the same kind of thinking perhaps underlies the Integrated Information theory. But emergence can mean more than that; in particular it might say that when systems have enough structural complexity of the right kind, they acquire interesting properties — meaningful, experiential ones — that can only be addressed on a higher level of interpretation. That, I think, is true; it just doesn't help all that much.

Holloway wraps up with another pop at those fully conscious particles that surely no-one believes in anyway. I don't think he has shown that no materialist theory can be plausible — the great mainstream ideas of functionalism and computationalism are largely untouched — but I salute the chutzpah of anyone who thinks such an issue can be wrapped up in one side of A4 — and is willing to take it on!

A Short Solution

29 August 2019

Tim Bollands recently tweeted his short solution to the Hard Problem (not literally in a tweet — it's not that short). He also provides an argument for a pretty uncompromising kind of panpsychism. I have to applaud his boldness and ingenuity, but unfortunately I part ways with his argument pretty early on.

Bollands' starting premise is that it's 'intuitively clear that combining any two non-conscious material objects results in another non-conscious object'. Not really. Combining a non-conscious Victorian lady and a non-conscious bottle of smelling salts might easily produce a conscious being. More seriously, I think most materialists would assume that conscious human beings can be put together by the gradual addition of neural tissue to a foetus that attains consciousness by a similarly gradual process, from dim sensations to complex self-aware thought. It's not clear to me that that is intuitively untenable, though you could certainly say that the details are currently mysterious.

Bollands believes there are three conclusions we can draw: humans are not conscious; consciousness miraculously emerges; or consciousness is already present in the matter brains are made from. The first, he says, is evidently false (remember that); the second is impossible, given that putting unconscious stuff together can't produce consciousness; so the third must be true.

That points to some variety of panpsychism, and in fact Bollands goes boldly for the extreme version which attributes to individual particles the same kind of consciousness we have as human beings. In fact, your consciousness is really the consciousness of a single particle within you, which due to the complex processing of the body has come to think of itself as the consciousness of the whole.

I can't recall any other panpsychist who has been willing to push fully-developed human consciousness right down to the level of elementary particles. I believe most either think consciousness starts somewhere above that level, or suppose that particles have only the dimmest imaginable spark of awareness. Taking this extreme position raises very difficult questions. Which particle is my conscious one? Are they all being conscious in parallel? Why doesn't my consciousness feel like the consciousness of a particle? How could all the complex content of my current conscious state be held by a single invariant particle? And why do my particles cease to be conscious when my body is dead, or stunned? You may notice, incidentally, that Bollands' conclusion seems to be that human beings as such are not, in fact, conscious, contradicting what he said earlier.

Brevity is perhaps the problem here; I don't think Bollands has enough space to make his answers clear, let alone plausible. Nor is it really clear how all this solves the Hard Problem. Bollands reckons the Hard Problem is analogous to the Combination Problem for panpsychism, which he has solved by denying that any combination occurs (though his particles still somehow benefit from the senses and cognitive apparatus of the whole body). But the Hard Problem isn't about how particles or nerves come together to create experience; it's about how

phenomenal experience can possibly arise from anything merely physical. That is, to put it no higher, at least as difficult to imagine for a single particle as for a large complex organism.

So I'm not convinced — but I'd welcome more contributions to the debate as bold as this one.

Libet Unreadied

16 September 2019

Benjamin Libet's famous experiments have been among the most-discussed topics of neuroscience for many years. Libet's experiments asked a subject to move their hand at a random moment of their choosing; he showed that the decision to move could be predicted on the basis of a 'readiness potential' detectable half a second before the subject reported having made the decision. The result has been confirmed many times since, and even longer gaps between prediction and reported decision have been claimed. The results are controversial because they seem to be strong scientific evidence against free will. If the decision had been made before the decider knew it, how could their conscious thought have been effective?

Libet's findings and their significance have been much disputed philosophically, but a new study credibly suggests that the readiness potential (RP) or Bereitschaftspotential has been misunderstood all along.

In essence the RP is a peak of neural activity, which seemed to the original researchers to represent a sort of gathering of neural resources as the precursor of an action. But the new study suggests that this appearance is largely an artefact of the methods used to analyse the data, and that the RP really represents nothing more than the kind of regular peak which occurs naturally in any system with a lot of background noise. Neural activity just does ebb and flow like that for no particular reason.

That is not to say that the RP was completely irrelevant to the behaviour of Libet's subjects. In the rather peculiar circumstances of the original experiment, where subjects are asked to pick a time at random, it is likely that a chance peak of activity would tip the balance for the unmotivated decision. But that doesn't make it either the decision itself or the required cause of a decision. Outside the rather strange conditions of the experiment, it has no particular role. Perhaps most tellingly, when the original experiments were repeated with a second group of subjects who were asked not to move at all, it was impossible to tell the difference between the patterns of neural activity recorded; a real difference appeared only at the time the subjects in the first group reported having made a decision.

This certainly seems to change things, though it should be noted that Libet himself was aware that the RP could be consciously 'over-ruled', a phenomenon he called 'Free Won't'. It can, indeed, be argued that the significance of the results was always slightly overstated. We always knew that there must be neural precursors of any given decision, if not so neatly identifiable as the RP. So long as we believe the world is governed by determined physical laws (something I think it's metaphysically difficult to deny) the problem of Free Will still arises; indeed, essentially the same problem was kicked around for centuries in forms that relied on divine predestination or simple logical fatalism rather than science.

Nevertheless, it looks as though our minds work the way they seem to do rather more than we've recently believed. I'm not quite sure whether that's disappointing or comforting.

Chatbot Fever

21 September 2020

The Guardian recently featured an article produced by GPT-3, OpenAI's powerful new language generator. It's an essay intended to reassure us humans that AIs do not want to take over, still less kill all humans. GPT-3 also produced a kind of scripture for its own religion, as well as other relatively impressive texts. Its chief advantage, apparently, is that it uses the entire Internet as its corpus of texts, from which to work out what phrase or sentence might naturally come next in the piece it's producing.

Now I say the texts are impressive, but I think your admiration for the Guardian piece ebbs considerably when you learn that GPT-3 had eight runs at this; then the best bits from all eight were selected and edited together, not necessarily in the original order, by a human. It seems the AI essentially just trawled some raw material from the Internet which was then used by the human editor. The trawling is still quite a feat, though you can safely bet that there were some weird things among the stuff the editor rejected. Overall, it seems GPT-3 is an excellent chatbot, but not really different in kind.

The thing is, for really human-style text, the bot needs to be able to deal with meaning, and none of them even attempt that. We don't really have any idea of how to approach that challenge; it's not that we haven't made enough progress, rather, we're not even on the road and have not really got anywhere with finding it. What we have got surprisingly good at is making machines that fake meaningfulness, or manage to do without it. Once it would have seemed pretty unlikely that computer translation would ever be any good, because proper translation involves considering the meanings of words. Of course Google Translate is still highly fallible, but it's good enough to be useful.

The real puzzle is why people are so eager to pretend that AIs are ready for human-style conversations and prose composition. Is it just that so many of us would love a robot pal (I certainly would)? Or some deeper metaphysical loneliness? Is it a becoming but misplaced modesty about human capacities? Whatever it is, it seems to raise the risk that we'll all end up talking to the nobody in the machine, like budgies chirping at their reflections. I suppose it must be acknowledged that we've all had conversations with humans where it seemed rather as if there was no-one at home in there.

[Looking back on this from 2026, it’s amazing that I could be so essentially right and yet also totally wrong.]

Body Memory

22 October 2020

Can you remember with your leg, and is that part of who you are? An interesting piece by Ben Platts-Mills in Psyche suggests it might be so. This is not Psyche the good old journal that went under in 2010, alas, but a promising new venture from those excellent folk at Aeon.

The piece naturally mentions Locke, who put memory at the centre of personal identity, and gives some fascinating snippets about the difficult lives of people with anterograde amnesia, the inability to form new memories. These people have obviously not lost their identities, however they may struggle. I have always been sceptical about the Lockean view more or less for this reason; if I lost my memories, would I stop being me? It doesn't quite seem so.

In the cases mentioned we don't quite have that total loss, though; the subjects mainly retain some or all of their existing memories, so a Lockean can simply say those retained memories are what preserve their identity. The inability to form new memories merely stops their identity going through the incremental further changes which are presumably normal (though some of these people do seem to change).

Platts-Mills wants to make the bolder claim that in fact important bits of memory and identity are stored, not in the brain, but in the context and bodies involved. Perhaps a subject does not remember anything about art, but fixed bodily habits draw them towards a studio where the tools and equipment (perhaps we could say the affordances) prompt them to create. This clearly chimes well with much that has been written about various versions of the extended or expanded mind.

I don't think these examples make the case very strongly, though; Platts-Mills seems to underestimate the brain, in particular assuming that the absence of conscious recall means there's no brain memory at all. In fact our brains clearly retain vast amounts of stuff we cannot get at consciously. Platts-Mills mentions a former paediatric nurse who cannot remember her career but when handed a baby, holds it correctly and asks good professional questions about it. Fine, but it would take a lot to convince me that her brain is not playing a major role there.

One of the most intriguing and poignant things in the piece is the passing remark that some sufferers from anterograde amnesia actually prefer to forget their earlier lives. That seems difficult to understand, but at any rate those people evidently don't feel the loss of old memories threatens their existence — or do they consider themselves new people, unrelated to the previous tenant of their body?

That last thought prompts another argument against the extended view. The notion that we could be transferred to someone else's body (or to no body at all) is universally understood and accepted, at least for the purposes of fiction (and, without meaning to be rude, religion). The idea of moving into a different body is not one of those philosophical notions that you should never start trying to explain at the end of a dinner party; it's a common feature of SF and magic stories that no-one jibs at. That doesn't, of course, mean transfers are actually possible, even in theory (I feel fairly sure they aren't), but it shows at least that people at large do not regard the body as an essential part of their identity.

Seating Consciousness

14 November 2020

A short piece by Tam Hunt in Nautilus asks whether the brain's electromagnetic fields could be the seat of consciousness.

What does that even mean? Let's start with a sensible answer. It could just mean that electromagnetic effects are an essential part of the way the brain works. A few ideas along these lines are discussed in the piece, and it's a perfectly respectable hypothesis. But it's hard to see why that would mean the electromagnetic aspects of brain processes are the seat of consciousness any more than the chemical or physical aspects. In fact the whole idea of separating electromagnetic effects from the physical events they're associated with seems slightly odd; you can't really have one without the other.

A much more problematic reading might be that the electromagnetic fields are where consciousness is actually located. I believe this would be a kind of category error. Consciousness in itself (as opposed to the processes that support and presumably generate it) does not have a location. It's like a piece of arithmetic or a narrative; things that don't have the property of a physical location.

It looks as if Hunt is really thinking in terms of the search, often pursued over the years, for the neural correlates of consciousness. The idea of electromagnetic fields being the seat of consciousness essentially says: stop looking at the neurons and try looking at the fields instead.

That's fine, except that for me there's a bit of a problem with the 'correlates of consciousness' strategy anyway; I doubt whether there is, in the final analysis, any systematic correlation. By way of explanation I offer an analogy: the search for the textual correlates of story. We have reams of text available for research, and we know that some of this text has the property of telling one or another story. Lots of it, equally, does not — it's non-fiction of various kinds. Now we know that for each story there are corresponding texts; the question is, which formal properties of those strings of text make them stories?

Now the project isn't completely hopeless. We may be able to identify passages of dialogue, for example, just by examining formal textual properties (occurrence of quote marks and indentation, or of strings like 'said'). If we can spot passages of dialogue, we'll have a pretty good clue that we might be looking at a story. But we can only go so far with that, and we will certainly be wrong if we claim that the textual properties that suggest dialogue can actually be identified with storyhood. It's obvious that there could be passages of text with all those properties that are in fact mere gibberish, or a factual report. Moreover, there are many stories that have no dialogue and none of any of the other properties we might pick out. The fundamental problem is that storyhood is about what the text means, and that is not a formal property we can get to just by examination. In the same way, conscious states are conscious because they are about something, and aboutness is not a matter of patterns of neural or electromagnetic activity.

Be that as it may, Hunt's real point is to suggest that electromagnetic field correlates might be better than neural ones. If I've got this right, he is a panpsychist who believes our consciousness is built out of the sparks of lower-grade awareness which are natural properties of matter. There is obviously a question there about how the sparks get put together into richer kinds of consciousness, and Hunt thinks resonance might play a key part. If it's all about

electromagnetic fields, it clearly becomes much easier to see how some sort of resonance might be in play.

I haven't read enough about Hunt's ideas to be anywhere near doing them justice; I have no doubt there is a lot of reasonable stuff to be said about and in favour of them. But as a first reaction resonance looks to me like an effect that reduces complexity and richness rather than enhancing them. If the whole brain is pulsing along to the same rhythm that suggests less content than a brain where every bit is doing its own thing. But perhaps that's a subject I ought to address at better length another time, if I'm going to.

The Kekulé Problem

1 November 2022

Cormac McCarthy asks an interesting question, and gives the wrong answer, also interestingly. Along the way he manages to describe in simple language a fundamental problem of how our minds work, one that he rightly says remains mysterious — though I think I have a clue.

McCarthy briefly retells the story of Kekulé, who struggled to understand the form of the benzene molecule. A dream of a lizard biting its own tail (like Ouroboros) prompted him to realise that the molecule was a ring. So we may conclude that Kekulé's unconscious had done the work for him and found the solution. McCarthy rightly contends that most of our problem-solving, our maths and so on, is done unconsciously. We can reason about things consciously using language and other symbols explicitly, but it is a very slow, clunky and ineffective business. In fact the main point of doing this explicit reasoning is usually to drag into conscious, recordable form the logic or processes we would normally use unconsciously.

The problem posed by McCarthy is: why didn't Kekulé's unconscious just tell him the answer, instead of sending it as an esoteric symbol in a dream? The unconscious understands language, he says, or it wouldn't have understood the problem Kekulé was trying to answer in the first place (I don't think that is clearly true). He concludes that the unconscious has been running our mental lives quite successfully for millions of years before language came along: it therefore distrusts and dislikes language and prefers other ways.

That isn't right. Part of the problem here is too strong a separation between conscious and unconscious, which are really just the silent and the talky bits of the same process. Kekulé's unconscious could not speak to him in words because any mental content that is in the form of language is automatically part of consciousness. Or to put it another way, his unconscious did tell him the answer in words — the words (let's say) 'By golly, benzene is a ring!' But those words appeared as thoughts of Kekulé's own, necessarily conscious thoughts. They were certainly put together by unconscious processes: as McCarthy recognises, all of our conscious verbal thoughts and utterances are constructed by unconscious processes — or we should get involved in an infinite regress.

How does the unconscious pull all this off? I have claimed before that our conscious thoughts and much of our higher cognition is grounded in recognition. The recognition of larger entities from smaller ones (and back down elsewhere) allows us to leap from present stimuli to distant, future or hypothetical ones, and finally into the realm of abstraction where we can recognise even the pure forms of mathematics. When we think about a problem, we do not typically do computations or talk to ourselves (though we might do that too). We hold the problem in mind and wait for our faculties of recognition to identify something that makes sense. I'm sure that's what happened in Kekulé's mind.

The same processes ultimately ground our use of language. As I have suggested, the grunt emitted when digging the dry riverbed can be recognised as part of the digging process, another part of which may be the desirable item of water. The grunt can therefore become an invitation to dig: and I think invitations and commands probably came before names.

Anyway, though I have a somewhat different answer, I think McCarthy is, as it were, digging in the right place with the right tools.

Daniel Dennett

20 April 2024

I feel I should acknowledge the death of Daniel Dennett, a philosopher who had the rare gift of communicating his interesting ideas very well. Even now, many years later and although I disagree with some significant bits, I would recommend *Consciousness Explained* as a great book on the subject. I think Dennett's faith in computation was over-optimistic, and although I basically agreed with him about his atheism I didn't share the crusading enthusiasm. In the nicest possible way, I really hope he's not in Heaven (or the other place), because that would annoy him quite considerably.

An excellent writer and a clear and original thinker — we cannot afford to lose too many of those!

To me Dennett's passing aptly symbolises the end of the era in which the topic of consciousness was hotly and enthusiastically debated, with many sure it would soon be cracked, and ambitious thinkers thinking they might win fame by publishing the Answer. It seems to me that things have quietened down now, and we don't expect a big breakthrough so soon. But it was fun while it lasted...

Decline of the Blog

During 2019 it became clear that Conscious Entities was slowing down, and the gaps between posts were getting longer. This was due to slowly worsening Crohn's Disease, which caused me constant gut problems and fatigue which at times was almost completely disabling. In 2020 I had to admit that the blog would be almost silent for a good while.

While Crohn's itself remains incurable, there is a growing armoury of good drugs to alleviate the symptoms. Alas, ultimately none worked for me and in 2022, during the Covid episode, I went into hospital for radical surgery. That was completely successful and returned my capacity to live a normal life with minor accomodations.

However, returning to Conscious Entities I found to my surprise that part of my mind just wouldn't co-operate in reading and writing more about consciousness (a disconcerting experience that might well have been the subject of an interesting post in itself!).

Perhaps, as I said at the time, this was partly due to continuing lack of energy, but I also had to admit that some of the energy I did have was now being siphoned off into writing stories, which I entered into competitions. Fiction, perhaps, is a little easier on the elderly brain. This has continued ever since, with some of my efforts and successes recorded on a new blog, *Seen and Done*.

Conscious Entities was neglected thereafter and eventually I allowed the blog to drop off the Internet. But it represents a significant part of my life, and I shouldn't like it to vanish completely. So these volumes form an archive where, for whatever it's worth, the posts are stored. I regret not being able to include all the lively dialogue in the comments from many regular friends that were such an important part of the whole thing, but you can still see the blog in its original form, together with the comments, on the invaluable Internet Archive.