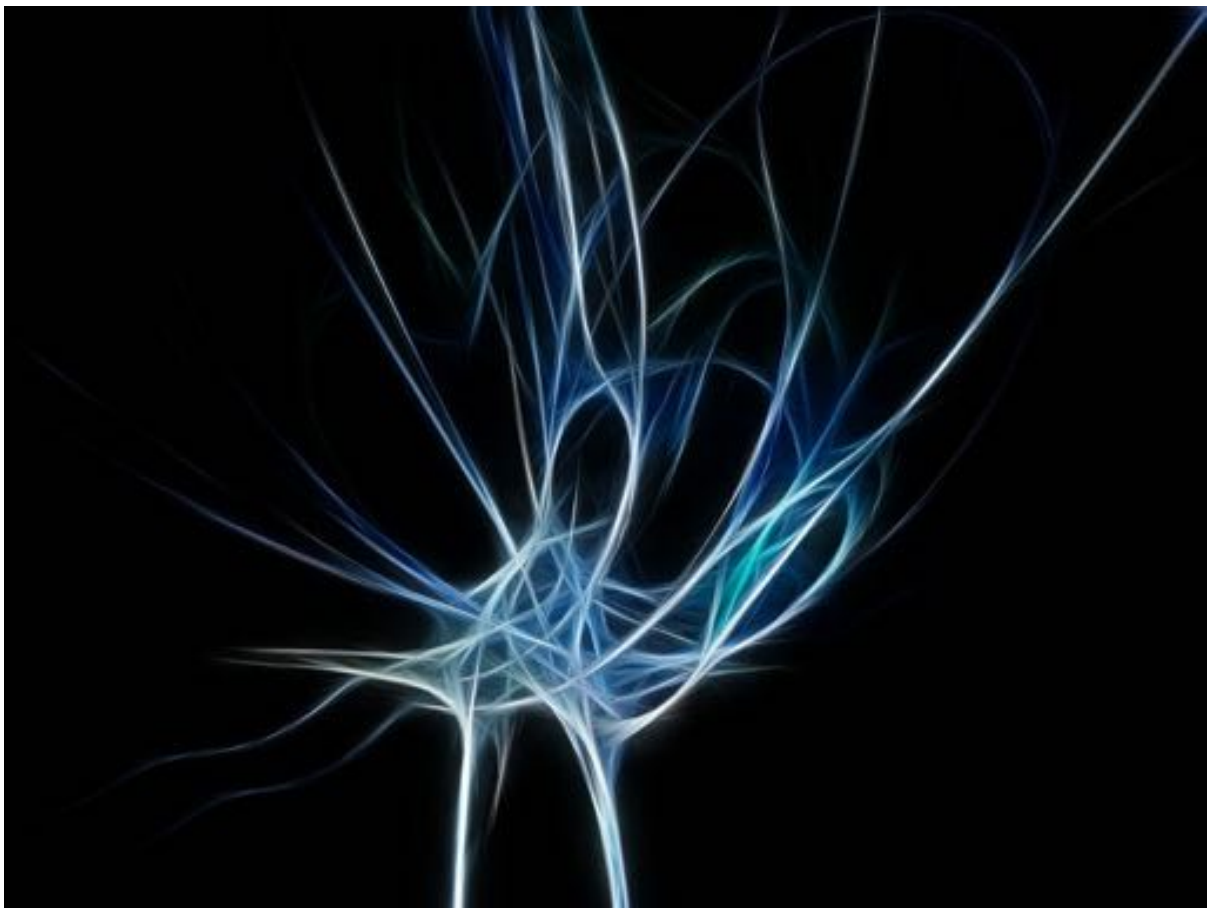
An Evolutionary Theory of Consciousness and Free Will



Developed from an Analysis of Tinbergen's Four Questions

by

ED GIBNEY

EXECUTIVE SUMMARY 3

SURVEY OF CURRENT THEORIES OF CONSCIOUSNESS 9

1 — Introduction to the Series 10
2 — The Illusory Self and a Fundamental Mystery 11
3 — The Hard Problem 13
4 — Panpsychist Problems with Consciousness 19
5 — Is Consciousness Just an Illusion? 22
6 — Introducing an Evolutionary Perspective 25
7 — More on Evolution 28
8 — Neurophilosophy 31
9 — Global Neuronal Workspace Theory 34
10 — Mind + Self 37
11 —Neurobiological Naturalism 40
12 — The Deep History of Ourselves 43
13 — (Rethinking) The Attention Schema 47
14 — Integrated Information Theory 54

AN EVOLUTIONARY THEORY OF CONSCIOUSNESS 61

15 — What is a Theory? 62
16 — A (Sorta) Brief History of the Definitions of Consciousness 65
17 — From Physics to Chemistry to Biology 76
18 — Tinbergen's Four Questions 92
19 — The Functions of Consciousness 96
20 — The Mechanisms of Consciousness 117
21 — Development Over a Lifetime (Ontogeny) 134
22 — Our Shared History (Phylogeny) 147
23 — Summary of My Evolutionary Theory 166

IMPLICATIONS FOR THE CONCEPT OF FREE WILL 178

My Review of "Just Deserts" by Daniel Dennett and Gregg Caruso 179
A Few Further Thoughts on Just Deserts 184
Another Free Will Debate — Kaufman v. Harris (Part 1/2) 186
Another Free Will Debate — Kaufman v. Harris (Part 2/2) 190
Some Thoughts on Sam Harris' Final Thoughts on Free Will 201
Summary of Freedom Evolves 205
Not My Final Thoughts on Free Will 212

ADDENDUM 216

The FAQs of Consciousness and Free Will 217

EXECUTIVE SUMMARY

Starting on the 15th of March 2020, I began writing a series of essays on consciousness and free will. After publishing 31 essays, the series was completed on the 3rd of October 2021. This exercise began as a simple survey of some current consciousness literature for my blog, which I thought would be a nice way to pass some time during the Covid-19 lockdowns. But, as the survey progressed, I discovered a hole missing from these studies. None of the major players had used **Tinbergen's four questions** (a standard tool of analysis in evolutionary studies) to look at all of the present and historical aspects of consciousness. I decided to work on this, which turned the blog series into a major research project. The results have been collated into this document.

The initial survey of theories of consciousness covered ideas from the following people:

- Authors Sam and Annaka Harris
- Philosophers David Chalmers, Phillip Goff, Keith Frankish, and Dan Dennett
- Neuroscientists Patricia Churchland, Stanislas Dehaene, Antonio Damasio, Todd Feinberg and Jon Mallatt, Joseph Ledoux, Michael Graziano, and Christof Koch

During this time, I was also invited to review a book about free will so I ended up writing essays examining the positions of the following people in that debate:

- Author Sam Harris
- Psychologist Scott Barry Kaufman
- Philosophers Gregg Caruso and Dan Dennett (who reviewed the entire field)

Any considerations of consciousness and free will are built on underlying beliefs about what we can hope to learn about the nature of the universe. My own starting point for these **epistemological** and **metaphysical** positions can be summarised as:

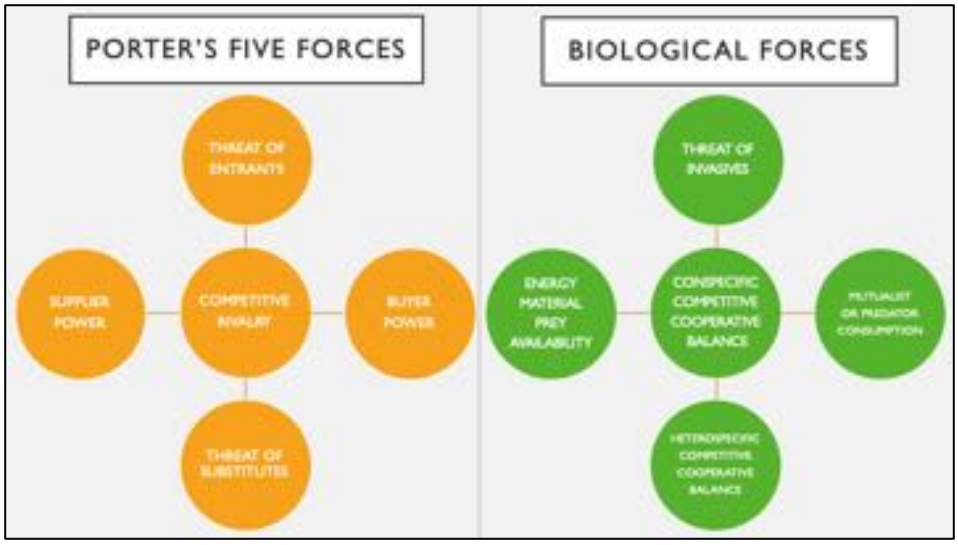
- **Epistemology:** It seems we will never be able to claim to know truth. Therefore, knowledge can only ever be a justified belief that is currently surviving our best tests.
- **Metaphysics:** The hypothesis that we live in a physical universe continues to survive. Although we do not have full explanations for all of nature, no non-natural phenomena have ever been reliably identified.

The "hard problem" for consciousness, as coined by David Chalmers, basically starts by noting that we humans all agree we have subjective experience. And evolutionary studies have shown us that there is an unbroken line in the history of life. But since water and rocks don't appear to have anything like consciousness, the question arises how inert matter could ever evolve into having the subjective experience that we humans undoubtedly feel?

Chalmers has suggested that subjectivity may be a fundamental property of the universe in the way that electromagnetism was discovered to be in the 19th century. I have come to agree with that conclusion. I hypothesise, however, that while the fundamental forces of physics are everywhere, and it is a fundamental property of the universe that these forces can be felt subjectively, this subjectivity can only emerge once subjects emerge. Until matter is organised into a living subject that is capable of responding to forces in such a way as to remain alive, it makes no sense to talk of non-living matter as experiencing or feeling those forces. Inert matter has no *structure* capable of receiving, registering, and responding to subjectivity. This

only occurs in actual subjects. Since the Greek word for force is *dynami*, I would therefore say that the universe has *pandynamism* rather than *panpsychism*. A psyche only originates and evolves along with life. The way that life emerged on this planet (abiogenesis) is still an open question, but the **RNA world** hypothesis has gathered enough plausible evidence to gain wide acceptance. Adding this to the theory of pandynamism get us from the inert landscape of the early universe to the rich and vibrant present-day world with biology and subjective experience. This bridges the gap between inert and living matter without imbuing properties to either class that violates our intuitive understanding.

If the theories of the RNA World and pandynamism hold up, this brings life into the world that experiences subjectivity. That would be a major discovery, but this view would also make it clear that something else emerges along with the emergence of such life. Given that living things are (to the best we know) merely structures of matter that have come to be organised so as to be self-sustaining and self-replicating, two new categories of things in the world appear which are related to that: 1) things that help life stay alive, and 2) things that harm life from staying alive. That division has no meaning in the worlds of physics or chemistry, but it is fundamental once biology emerges. Through the trials and errors of natural selection, living systems would become sensitive to these positive and negative aspects of the world and they would be preserved according to their success navigating them. In science, something exists to the extent that it exerts causal power over other things. Gravity exists because it exerts power over mass. Electricity exists because it exerts power over charged particles. Similarly, the power these categories exert over living things implies they exist too. I call them 'biological forces' and believe they can be understood in a comprehensive manner along the same lines that **Porter's Five Forces** are understood to impact the survival of organisations.



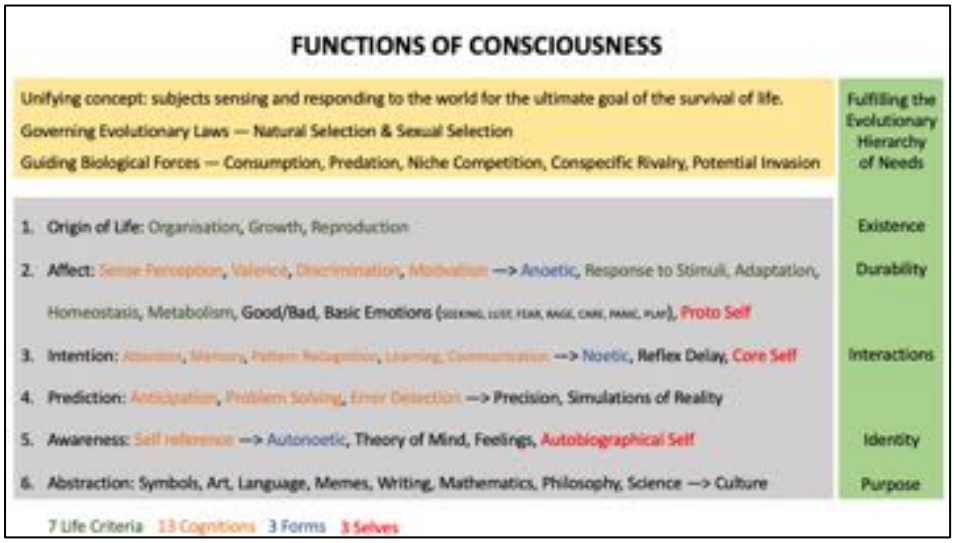
To date, **definitions of consciousness** stretch all the way from it being something as small as the private, ineffable, special feeling that only we rational humans have when we think about our thinking, right on down to it being a fundamental force of the universe that gives proto-feelings to an electron of what it's like to be that electron. As the **Wikipedia entry on consciousness** notes:

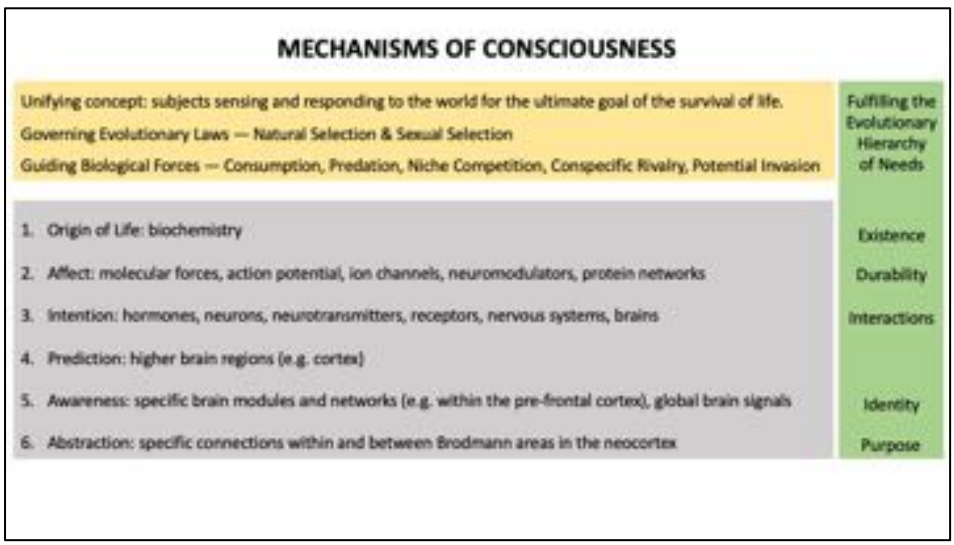
> The level of disagreement about the meaning of the word indicates that it either means different things to different people, or else it encompasses a variety of distinct meanings with no simple element in common.

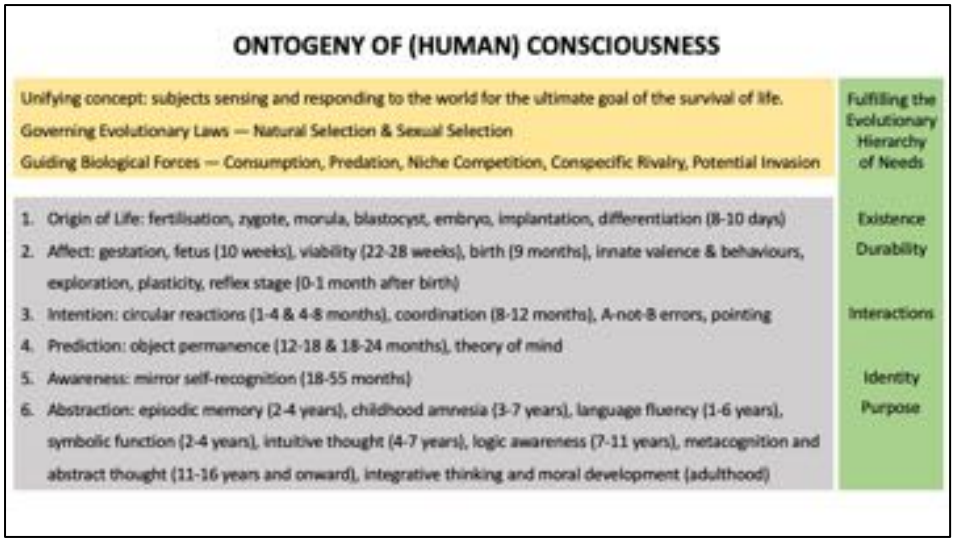
Considering the arguments above about pandynamism and biological forces, a new evolutionary theory of consciousness emerges that can capture all varieties of consciousness related to the history of living organisms. Namely:

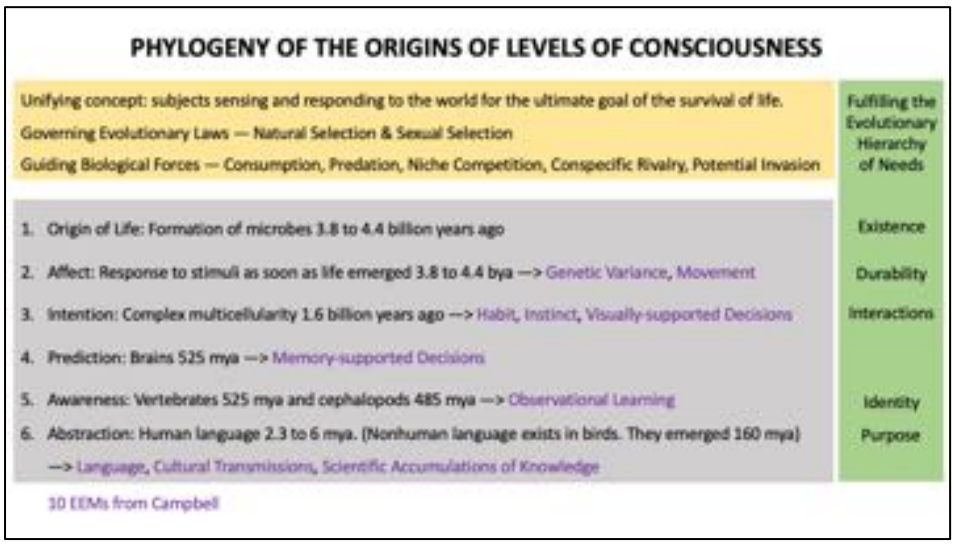
> **An Evolutionary Theory of Consciousness**
> Consciousness, according to this evolutionary theory, is an infinitesimally growing ability to sense and respond to any or all biological forces in order to meet the needs of survival. These forces and needs can vary from the immediate present to infinite timelines and affect anything from the smallest individual to the broadest concerns (both real and imagined) for all of life.

This is a very broad and inclusive definition that is intended to capture the entirety of the subject. Major figures in the field of consciousness studies have preferred to draw a line or circle around narrower conceptions and insist *that* is consciousness, but I find it much more helpful and informative to consider the broad spectrum of *all* aspects of consciousness and let the arguments over restrictive terminology melt away. In order to map the contours of such a broad definition, I spent several posts conducting a **Tinbergen analysis** of the **functions**, **mechanisms**, **ontogeny**, and **phylogeny** of consciousness. This is the standard procedure in evolutionary studies for coming to know all of the elements of any biological phenomenon. That massive review resulted in the following four charts:









For more details on each of these charts, see the individual essays where they were developed. Altogether, I believe these give a full picture of the various aspects of consciousness. The first tier in this hierarchy — **1) Origin of Life** — has already been discussed above. The remaining tiers are:

2) Affect: This is the valence, tone, or mood that is capable of distinguishing differences between good stimuli as opposed to bad ones, which results in responses of graduated arousal and intensity. **Mark Solms** calls this the primary experience and purpose of consciousness. He asks, rhetorically, how can affective arousal (i.e., the arousal of feeling) go on without any inner feel? It cannot. This accords with my theory of pandynamism, where such feelings are felt subjectively as soon as subjects appear and are affected by biological forces. At first, these affects will generate what we think of as instinctual unconscious reactions. These can involve any or all of **Jaak Panksepp's** seven basic emotions (in capital letters to denote a distinction between them and their common usage): SEEKING, FEAR, RAGE LUST, PANIC, CARE, and PLAY. Later, once many more structures have evolved, these affects can be registered, and eventually named, in conscious awareness.

3) Intention: This development in consciousness marks the ability of one reaction to interrupt or override others within an organism. From the perspective of an outside observer, choices appear to be made and there is a narrative sequence to life. Like affect, this can take

place unconsciously within humans, so presumably it can in other forms of life as well, but it does empirically exist in very simple life using cognitive abilities such as memory, pattern recognition, and learning. Much later in evolutionary history, this can also be accessed and rationally considered in order to create extremely complex and far-ranging intentions.

4) Prediction: Once intentions exist (either one's own or the intentions of others), the next development in consciousness is to take them into account by predicting how intentions will interact with the world. Organisms no longer just respond to the present by building up memories of the past; they begin to guess the future too. This appears to happen only in animals with brains that have neuroplasticity and can learn from experience. It also would seem that predictions about the intentions of others are particularly vital, which would explain why neurons and brains appear to have emerged during the Cambrian explosion due to the onset of predation. The success or failure of one's predictions about their predators or prey would have been a powerful driver for change in any arms race occurring in this new dimension of consciousness. Surprise and uncertainty would be a bad emotion for any prediction, which would eventually help to hone the development of feelings of precision to extremely high levels.

5) Awareness: The next level of consciousness comes in now that structures have evolved to trigger affective emotions in the present (level 2), evaluate the past to make complex choices (level 3), and predict further and further into the future what the actions of the self and others may result in (level 4). The interaction and comparison of these three phenomena allows for the dawning of awareness of a self that is different from others. The richness of this distinction grows with the number of sensations that are able to be evaluated against one another within more and more sophisticated models of elements of the world. Studies have shown that conscious awareness is indeed necessary for some types of learning that give organisms additional plasticity to respond to new and novel stimuli in their environment, thus cementing the evolutionary advantage of gaining and retaining this ability.

6) Abstraction: The final level of consciousness in this hierarchy comes when models of reality go beyond mere direct representation and begin to use symbolic representations to evoke, communicate, and manipulate thoughts and feelings about the world. While nonhuman animals have displayed rudimentary or latent abilities for abstraction, the emergence and development of this capability in humans has been of such enormous import that it is considered the latest of **the major transitions of evolution**. Symbols, art, and language have driven the cultural evolution of memes, writing, mathematics, philosophy, and science that make up all of the powerful products of human culture. The causes for the emergence of this type of consciousness are mysteriously shrouded in the history of one species at the moment, but there is no denying the power, for good or for ill, that this has enabled. May our fuller grasping of the biological forces that affect the consciousness of all of life motivate us to realise what good is and bring it into fruition.

These six levels have been developed and honed during the examination of all four of Tinbergen's questions. In each question, I found that the emergence of consciousness followed exactly along the same lines of these levels, which provides a great example of consilience where multiple streams of evidence are all pointing to the same thing. This will need to be verified and refined by researchers with more expertise in each of these areas, but for now it appears this new evolutionary theory is an exciting hypothesis.

Such a view of the evolution of consciousness has major implications for the understanding of free will. As **Dan Dennett** noted:

> "It is no mere coincidence that the philosophical problems of consciousness and free will are, together, the most intensely debated and (to some thinkers) ineluctably mysterious phenomena of all. As the author of five books on consciousness, two books on free will, and dozens of articles on both, I can attest to the generalization that you cannot explain consciousness without tackling free will, and vice versa."

In a nutshell, I agree with Dennett that we don't have the freest will imaginable, but we do have significant degrees of freedom, and that provides a kind of "free will worth wanting." As was the case with consciousness, an analysis using Tinbergen's four questions sheds much light on the emergence and expansion of these freedoms. And it turns out they are completely aligned with the emergence and expansion of consciousness as outlined above.

TINBERGEN ANALYSIS OF FREE WILL WITHIN HIERARCHIES OF CONSCIOUSNESS

Unifying concept: subjects sensing and responding to the world for the ultimate goal of the survival of life.
Governing Evolutionary Laws — Natural Selection & Sexual Selection
Guiding Biological Forces — Consumption, Predation, Niche Competition, Conspecific Rivalry, Potential Invasion

Hierarchies of Consciousness	Functions	Phylogeny (Appearance in History)	Ontogeny (Develop in Self)	Mechanisms	Fulfilling the Evolutionary Hierarchy of Needs
1. Origin of Life	No Free Will at First	***	***	***	Existence
2. Affect	In moment reflex choice of good/bad	Once life established	Innate valence	Molecular Forces ("Pandynamism")	Durability
3. Intention	Choice of present based on past	Complex multicellularity	Learned reactions	Chemical messages / nerve systems	Interactions
4. Prediction	Choosing between alternate futures	Brains	Improving guesses	Cortex	
5. Awareness	Decisions based on wider view of self	Vertebrates, cephalopods	Individual goals and preferences	Local modules / global signals	Identity
6. Abstraction	Expands freedom to abstract influences	Human language	Episodic and strategic planning	Specific neocortex regions	Purpose

I think it's easiest to grasp this table by focusing on the Functions column. Going from top to bottom, there is (1) no free will before the emergence of life. Once (2) life is established, the phenomenon of affect provides innate valences for making in the moment reflex choices between good or bad options for life. As (3) complex multicellularity develops mechanisms to learn and act on (unconscious at first) intentions, then life gains the freedom for choosing different actions in the present based on things it has learned in the past. Continuing on, the (4) development of brains enables modelling predictions of the world, which gives life freedom to choose between alternate futures. As all of these abilities lead to (5) the dawning of self-awareness where living organisms can begin to develop autobiographical narratives that inform choices over longer and longer time horizons depending on the quantity and quality of memories and predictions that have been developed. Finally, in the (6) realm of human language, we *Homo sapiens* have gained the freedom to be influenced by an infinite array of abstract representations. At this level, we can now see strategic planning of actions for decades of a life, which clearly drives the feelings of free will that exist in folk psychology.

For more on the implications that this view of free will has on our moral responsibility to deserve praise or blame for our actions, see **my review of *Just Deserts* by Gregg Caruso and Dan Dennett**.

SURVEY OF CURRENT THEORIES OF CONSCIOUSNESS

1 — Introduction to the Series

15 March 2020

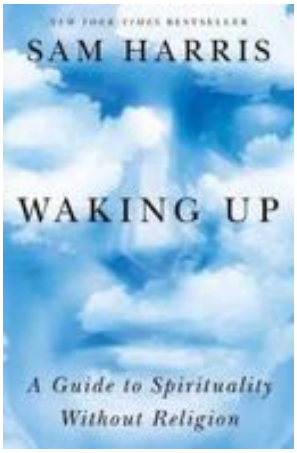
With the onset of the Covid-19 coronavirus, many of us are in the midst of a time for self-isolation. So, I figured this would be the perfect time to try to isolate the self.

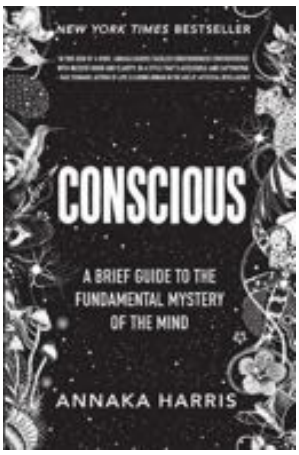
It's an exciting time for the study of consciousness right now with a wide range of ideas and scientific studies being discussed and released. For a while now, I have felt that I've not written enough about this subject to wrap my head around it all, but after going through a deluge of podcasts, talks, and articles about it, I think the time is finally right for me to dive in. I've also just agreed to take part in a one-on-one public discussion about this (and other things) for a Darwin Day lecture in 2021, so I really do need to get up to speed.

With that said, I'm going to do something different on the blog here. Rather than write one ginormous post about consciousness, I'm going to publish this as a long series of shorter posts. I'm going to try to do this every other day to give people a little time to read, listen, digest, and comment along the way, but I don't want this to spread out too long so the plot gets lost along the way. Right now, I've got 16 posts in mind, so this will take a little over a month. Hopefully that will get us through the worst of this pandemic.

I'll be back soon for part 2. In the meantime, spread the word (non-contact please) to anyone else who might be interested in this. Maybe let me know in the comments what your current thinking on this subject is. It'd be interesting to see how that might change along the way. No matter what, I hope it will be a fun ride!

2 — The Illusory Self and a Fundamental Mystery





17 March 2020

As I begin this series on consciousness, a good place to start is by knocking down preconceived notions that may be common here. This is something the Buddhists have been doing for hundreds of years and a good guide to this way of thinking is Sam Harris. Sam has studied meditation for years, including on very long silent retreats in India, but he is now much better known as a hyper-rational neuroscientist who is one of the four horsemen of the New Atheists. How does he bring these experiences together? His books, meditation app, and podcast are full of discussions about this, but I'll choose two particular podcasts to focus on.

The first is **Episode #181- The Illusory Self**. Most of this episode is a discussion with the writer and meditation teacher Richard Lang, but the introductory comments from Sam from 5:27 to 13:30 are particularly useful for my purposes. Here are the most important lines:

- In today's podcast I want to give you sceptics one more shot understanding what I'm up to with meditation. There are specific insights here into the nature of mind that I consider to be the most important things I have ever learned.
- I've been slow to understand just how much intellectual work is being done for me by the fact that I've had certain experiences in meditation. And these experiences have made certain features of the mind obvious.
- The reality is that if you can pay sufficient attention to your mind, the illusion [of free will and the self] disappears. It becomes obvious that everything is just arising on its own, including one's thoughts and intentions and other mental precursors to action.
- Consider the analogy that I've sometimes used to the optic blind spot. You make two marks on a piece of paper. You stare at it. You close one eye, look at one of the marks, and bring the paper closer until the second mark disappears. This is a very simple procedure that allows you to see something right on the surface of consciousness that you would otherwise spend your entire lifetime overlooking.
- In seeing the blind spot, you're actually seeing something subjectively, as a matter of direct experience, that reveals a deeper truth about the eye. Well, I can also say that the non-existence of an unchanging self in the middle of experience, an ego, the feeling that we call I, is also predicted by the structure and function of the brain. ... There's no account of neuroanatomy or neurophysiology that would make sense of an unchanging self freely exercising its will. Meditation is ultimately a very simple procedure that allows one to discover the absence of this fake self directly.

Next, I'd draw your attention to **Episode #159 - Conscious**, which is a discussion with Sam's wife Annaka Harris about her book *Conscious: A Brief Guide to the Fundamental Mystery of the Mind*. Once again, there are a lot of interesting things said in this podcast, but here are the most important lines:

- The [hard] problem is, why is it that any configuration of non-conscious material can suddenly *have the experience* of being that matter? There's no explanation that we could think of that could make this less mysterious. It's always non-conscious matter getting arranged in a very specific way so that it suddenly lights up from the inside. It seems that no matter how much we know about the brain, there's nothing that will ever make this less mysterious.
- The most primary intuitions we have about consciousness live in two questions I like to keep asking myself. The first one is: is there any behaviour on the outside of a system that can tell us conclusively that consciousness is present in that system?
- The second question is: is consciousness doing anything? Is it serving a function?
- The idea that consciousness might not be doing anything is problematic from an evolutionary point of view because people wonder then why it would have evolved. Surely it must be doing something, because it must be expensive metabolically on some level.
- So the argument about the evolution of consciousness is one that sends many people down the path of wondering if it is possible that consciousness is a fundamental feature of matter, that it is there in some form all the way down.
- The name for that general family of views is panpsychism.
- In my book I cite the title of this great article by Philip Goff, which is: "Panpsychism may be crazy, but it's also most probably true." That got me to the point where I started to take panpsychism more seriously. ... Once you're able to break through the illusion of the self, these sorts of theories are easier to entertain or imagine.

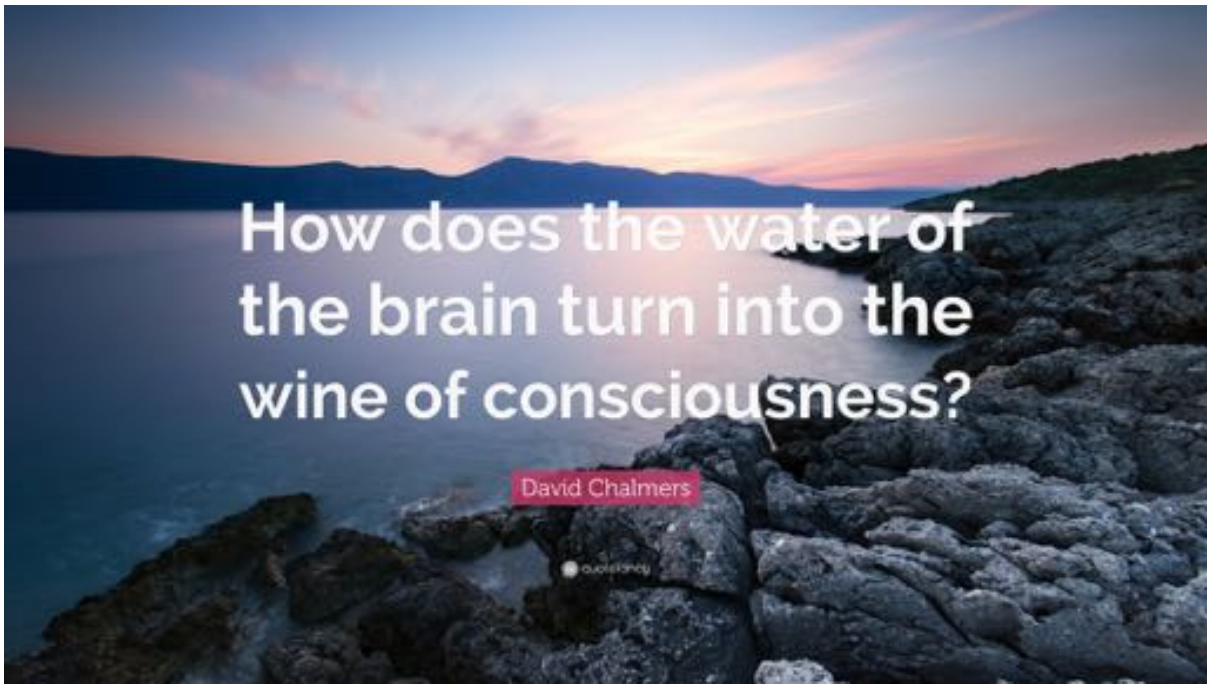
Brief Comments

Having done a bit of meditation over the last 15 years, I can see the value of paying close attention to where thoughts, feelings, and intentions arise from. I can easily agree with Sam that there is no "unchanging self in the middle of experience, an ego, the feeling that we call I." But whenever Sam goes on about there being no self, I like to remember Laurence Krauss telling him he was pretty sure he could find a self somewhere within the vicinity of his body. He and Sam could be using different definitions for what the self is, however, and that's something I'll explore more later.

As for Annaka's points, I first wanted to let her introduce the idea of "the hard problem of consciousness" here. There will be much more about that in the next article focusing on David Chalmers who coined that term. As for Annaka's primary questions about consciousness, I think the first one looking for conclusive evidence of consciousness is a common error of essentialist thinking in an evolving universe where lines are blurry and there are no on/off eternal essences. Dan Dennett will address that later but it's important to see right away that looking for "consciousness" doesn't reveal any obvious answers. As for what consciousness does, that depends a lot on how it is defined, which neuroscientists have been teasing out over the last several years. Whether they find panpsychism all the way down will be up for interpretation. I'll cover much more about that further down the line.

What do you think? Are you even a you? Is the hard problem of consciousness hard to you?

3 — The Hard Problem



19 March 2020

In the **last post** in this series, I shared a couple of podcasts that knocked down the common / religious / folk views of consciousness, which sees it as something separate from our bodies, unchanging, or immortal. Close observations of the world—whether scientific or meditative—just don't find any evidence for that kind of consciousness. And yet, we seem confident that we, ourselves, have it. So where does consciousness come from? That has been the subject of **the mind-body problem** in philosophy for centuries. Most modern people (especially of non-religious persuasions) now see the mind as embedded in the body. But since bodies are made up of the same physical stuff as the rest of the observable universe, it's unclear how minds could possibly ever arise from such stuff. In 1996, in his book ***The Conscious Mind***, the philosopher David Chalmers called this "the hard problem of consciousness" and it remains a deep sticking point for philosophers and scientists today.

I've heard Chalmers talk about this to loads of different people (e.g. Tom Stoppard **discussed his play about it** with him), but the best conversation I've come across was with the physicist Sean Carroll on his podcast **Mindscape - Episode 25**. The first 50 minutes of the podcast are particularly relevant, so here are the most important lines from that:

- [Sean Carroll] David describes himself as a naturalist, someone who believes in just the natural world, not a supernatural one. Not a dualist who thinks there's a disembodied mind or anything like that. But he's not a physicalist. He thinks that the natural world not only has physical properties, but mental properties as well. He's convinced of the problem, but he's not wedded to any solutions yet.
- [David Chalmers] The hard problem of consciousness is the problem of explaining how physical processes in the brain somehow give rise to subjective experience. ... When it comes to explaining behaviour, we have a pretty good bead on how to explain that. In principle, you find a circuit in the brain, maybe a complex neural system, which maybe

performs some computations, produces some outputs, generates the behaviour. Then, in principle, you've got an explanation. It may take a century or two to work out the details, but that's roughly the standard model in cognitive science. This is what, 20 odd years ago, I called the easy problem. Nobody thinks they are easy in the ordinary sense. The sense in which they are easy is that we've got a paradigm for explaining them.

- [DC] The really distinctive problem of consciousness is posed not by the behavioural parts but by the subjective experience. By how it feels from the inside to be a conscious being. I'm seeing you right now. I have a visual image of colours and shapes that are sort of present to me as an element of the inner movie of the mind. I'm hearing my voice, I'm feeling my body, I've got a stream of thoughts running through my head. This is what philosophers call consciousness or subjective experience. I take it to be one of the fundamental facts about ourselves, that we have this kind of subjective experience.
- [SC] Sometimes I hear it glossed as "what it is like" to be a subjective agent.
- [DC] That's a good definition of consciousness actually put forward by my colleague Thomas Nagel in an article back in 1974 called "What is it like to be a bat?" His thought was that we don't know what it is like. We don't know what a bat's subjective experience is like. It's got this weird sonar capacity that doesn't correspond directly to anything we humans have. But presumably there is something it is like to be a bat. A bat is conscious. On the other hand, people would say there is nothing it is like to be a glass of water. If that's right, the glass of water is not conscious. So, this "what it's like" way of speaking is a good way of serving as an initial intuition pump for the difference we're getting at between systems that are conscious and systems which are not.
- [SC] The other word that is sometimes invoked in this context is the "qualia" of the experiences we have. There is one thing that it is to see the colour red, and a separate thing to have the experience of the redness of red.
- [DC] This word qualia may have gone a little out of favour over the last 20 years, but you used to have a lot of people speaking of qualia as a word for the sensory qualities that you come across in experience. The paradigmatic one would be the experience of red vs. the experience of green. There are many familiar questions about this. How do I know that my experience of the thing we call red is the same as the experience you have? Maybe our internal experiences are swapped. That would be inverted qualia, if my red were your green. ... We know that some people are colour blind. They can't make a distinction between red and green. ... I have friends that have this and I'm often asking them, what is it like to be you? Is it all just shades of blue and yellow? We know that what it is like to be them can't be what it is like to be us.
- [DC] When it comes to consciousness, we're dealing with something subjective. I know I'm conscious not because I've measured my behaviour or anybody else's behaviour, but because it's something I've experienced directly from the first-person point of view. You're probably conscious, but it's not like I can give a straight up operational definition of it. We could come up with an AI that says it's conscious. That would be very interesting. But would that settle the question of whether it's having subjective experience? Probably not.
- [SC] Alan Turing noted a "consciousness objection" [to his Turing test], but said he can't possibly test for that so it's not meaningful.
- [DC] Yes. But it turns out consciousness is one of the central things that we value. A) It's one of the central properties of our minds. B) Many people think it's what actually gives lives meaning and value. If we weren't conscious, if we didn't have subjective experience, then we'd basically just be automata for whom nothing has any meaning or value. So I think when it comes to the question of whether sophisticated AI's are conscious or not, its going to be absolutely central to how we treat them, to whether they have moral status, whether we should care if they continue to live or die, whether they get rights, and so on.

- [SC] To get our cards fully on the table, neither of us are coming at this from a strictly dualist position. Neither of us are resorting to a Cartesian disembodied mind that is a separate substance. Right? As a first hypothesis, we both want to say that we are composed of atoms and obeying the laws of physics. Consciousness is somehow related to that but not an entirely separate category interacting with us. Is that fair to say?
- [DC] Yes, although there are different kinds and degrees of dualism. My background is in mathematics, computer science, and physics, so my first instincts are materialist. To try to explain everything in terms of the processes of physics: e.g. biology in terms of chemistry and chemistry in terms of physics. This is a wonderful great chain of explanation, but when it comes to consciousness, this is the one place where that great chain of explanation seems to break down. That doesn't mean these are the properties of a soul or some religious thing which has existed since the beginning of time and will go on after our death. People call that substance dualism. Maybe there's a whole separate substance that's the mental substance and somehow that interacts and connects up with our physical bodies. That view, however, is much harder to connect to a scientific view of the world.
- [DC] The version I end up with is sometimes called property dualism. This is the idea that there are some extra properties of things in the universe. This is something we already have in physics. During Maxwell's era, space and time and mass were seen as fundamental. Then Maxwell wanted to explain electromagnetism and there was a project that tried to explain it in terms of mass and space and time. That didn't work. Eventually, we ended up positing charge as a fundamental property with some new laws of physics governing these electromagnetic phenomena and that became just an extra property in our scientific picture of the world. I'm inclined to think that something slightly analogous to this is what we have to do with consciousness.
- [SC] You think that even if neuroscientists got to the point where, for every time a person was doing something we would all recognise as having a conscious experience, even if it was silent—for example, experiencing the redness of red—they could point to exactly the same neural activity going on in the brain, you would say this still doesn't explain my subjective experience?
- [DC] Yes. That's in fact a very important research program going on right now. People call it the program of finding the neural correlates of consciousness (the NCC). We're trying to find the NCC or neural systems that act precisely when you are conscious. This is a very important research program, but it's one for correlation, not explanation. We could know when a special kind of neuron fires in a certain pattern that that always goes along with consciousness. But the next question is why. Why is that? As it stands, nothing we get out of the neural correlates of consciousness comes close to explaining that matter.
- [DC] We need another fundamental principle that connects the neural correlates of consciousness with consciousness itself. Giulio Tononi, for example has developed his Integrated Information Theory where he says consciousness goes along with a mathematical measure of the integration of information, which he calls phi. The more phi you have, the more consciousness you have. Phi is a mathematically and physically respectable quantity that is very hard to measure, but in principle you could find it and measure it. There are questions of whether this is actually well defined in terms of the details of physics and physical systems, but it's at least halfway to something definable. But even if he's right that phi—this informational property—correlates perfectly with consciousness, there's still this question of why.
- [DC] *Prima facie*, it looks like you could have had a universe where the integration of information is going on, but no consciousness at all. And yet, in our universe, there's consciousness. How do we explain that fact? What I regard as the scientific thing to do at this point is to say that in science, we boil everything down into fundamental principles and laws, and we need to postulate a fundamental law that connects, say phi, with

consciousness. Then that would be great, maybe that's going to be the best we can do. In physics, there's a fundamental law of gravitation, or a grand unified theory that unifies all these different forces. You end up with some fundamental principles and you don't take them further. Something has to be taken as basic. Of course, you want to minimise our fundamental principles and properties as far as we can. Occam's razor says don't multiply entities without necessity. Every now and then, however, we have necessity. Maxwell was right about this with electromagnetism. Maybe I'm right about the necessity in the case of consciousness too.

- [SC] You've hinted at one of your most famous thought experiments there by saying you can imagine a system with whatever phi you want, but we wouldn't call it conscious. You take that idea to the extreme and say there could be something that looks and acts just like a person but doesn't have consciousness.
- [DC] Yes. This is the philosopher's thought experiment of the zombie. ... The philosopher's zombie is a creature that is exactly like us functionally, behaviourally, and maybe physically, but it's not conscious. It's very important to say that nobody, certainly not me, is arguing that such zombies actually exist. ... I'm very confident there isn't such a case now, but the point is that it at least seems logically possible. There's no contradiction in the idea of there being an entity just like you without consciousness. That's just one way of getting at the idea that somehow consciousness is something extra and special that is going on. You could put the hard problem of consciousness as, why aren't we zombies?
- [SC] How can I be sure that I'm not a zombie?
- [DC] There's a very good argument that I can't be sure *you're* not a zombie. All I have is access to your behaviour. But the first-person case is different. In the first-person case, I'm conscious, I know that more directly than I know anything else. Descartes said in the 1640's this is the one thing I can be certain of. I can doubt everything about the external world, but I can't doubt that I'm thinking. I think therefore I am. I think it's natural to take consciousness as our primary epistemic datum. Whatever you say about zombies I know that I'm not one of them because I know I'm conscious.
- [SC] What makes me worried is that the zombie would give itself all those same reasons. So, how can I be sure I'm not that zombie?
- [DC] To be fair, you've put your finger on the weakest spot of the zombie hypothesis and for the ideas that come from it. In my first book, *The Conscious Mind*, I had a whole chapter about this called this "The Paradox of Phenomenal Judgment" that stems from the fact that my zombie twin would say, and do, and write all of the things I was. We shouldn't take possible worlds too seriously, but what is going on in the zombie world is what philosophers call eliminativism, where there is no such thing as consciousness and the zombie is making a mistake. There is a respectable program in philosophy that says we're basically in that situation in our world, and lately there has been an upsurge in people taking this seriously. It's called illusionism.
- [DC] Illusionism is the idea that consciousness is some kind of internal introspective illusion. Think about what's going on with the zombie. The zombie thinks it has special properties of consciousness, but it doesn't. All is dark inside. Illusionists say, actually, that's our situation. It seems to us we have all these special properties—those qualia, those sensory experiences—but in a way, all is dark inside for us as well. There is just a very strong introspective mechanism that makes us think we have those special properties. That's illusionism.
- [DC] I've been thinking about this a lot and wrote an article called "The Meta Problem of Consciousness" that just came out. The hard problem of consciousness is why are we conscious, why do these physical processes give rise to consciousness. The meta problem of consciousness is: why do we think we're conscious? Why do we think there's a problem of consciousness? Remember, the hard problem says the easy problems are about

behaviour, and the hard problem is about experience. Well, the meta problem is ultimately about behaviour. It's about the things we do and the things we say. Why do people go around writing books about this? Why do they say, "I'm conscious", "I'm feeling pain"? Why do they say, I have these properties that are hard to explain in functional terms? That's a behavioural problem. That's an easy problem.

- [SC] Aside from eliminativism and illusionism, which are fairly hard core on one side, or forms of dualism on the other side, there is this kind of "emergent" position one can take that is physicalist and materialist at the bottom, but doesn't say that therefore things like consciousness and subjective experiences don't exist or are illusions. They are higher order phenomena like tables or chairs. They are categories that we invent to help us organise our experience of the world.
- [DC] My view is that emergence is sometimes used as a magic word to make us feel good about things we don't understand. How do you get from this to this? It's emergent! But what do you really mean by emergent? I wrote an article about this once where I distinguished weak emergence from strong emergence. Weak emergence is basically the kind you get from lower level structural dynamics explaining higher level structural dynamics: the behaviour of a complex system, the way traffic flows in a city, the dynamics of a hurricane etc. You get all sorts strange and surprising and cool phenomena emerging at the higher level. But still, once you understand the lower level mechanisms well enough, the higher-level ones just follow transparently. It's just lower level structure giving you higher level structure according to the following simple rules. When it comes to consciousness, it looks like the easy problems may be emergent in this way. Those may turn out to be low-level structural and functional mechanisms that produce these reports and these behaviours that lead us to being awake, and no one would be surprised if these were weakly emergent in that way. But none of that seems to add up to an explanation of subjective experience, which just looks like something new. Philosophers sometimes talk about emergence in a different way. Strong emergence involves something fundamentally new emerging via new fundamental laws. Maybe there's a fundamental law that says when you get this information being integrated then you get consciousness. I think consciousness may be emergent in that sense, but that's not a sense that helps the materialist. If you want consciousness to be emergent in a sense that helps the materialist, you have to go for weak emergence and that is ultimately going to require reducing the hard problem to an easy problem.
- [DC] Everyone has to make hard choices here and I don't want to let you off the hook by just saying, "Ah it's all ultimately going to be the brain and a bunch of emergence." There's a respectable materialist research program here, but that involves ultimately turning the hard problem into an easy one. All you are going to get from physics is more and more structure and dynamics and functioning and so on. For that to turn into an explanation of consciousness, you need to find some way to deflate what needs explaining in the case of consciousness to a matter of behaviour and functioning. And maybe say the extra thing that needs explaining, that's an illusion. People like Dan Dennett, who I respect greatly, has tried to do this for years, for decades. At the end of the day, most people look at what Dennett's come up with and they say "Nope, not good enough. You haven't explained consciousness." If you can do better, then great.
- [DC] I've explored a number of different positive views on consciousness. What I haven't done is commit to any of them. I see various different interesting possibilities, each of which has big problems. Big attractions, but also big problems to overcome.

Brief Comments

I've never given much weight to Chalmers' zombie problem. Relying on "conceivable worlds" strikes me as a reformulated **ontological argument** for the existence of God—i.e. if you

can imagine it, it must be so. But our imaginations can be wrong in all sorts of ways; possibly even in ways we can't imagine. That's why Descartes was wrong too. *Cogito ergo sum* should have been **I think, therefore I think I think**.

In this interview, however, Chalmers has convinced me there is a "hard" problem, but I think it is misnamed. Hard implies that it could be cracked. But what Chalmers keeps retreating to is ultimately an unanswerable question. After every new explanation of consciousness that could ever come along—from believing that consciousness is in our bodies, all the way to defining a theoretically perfect neural correlates of consciousness—Chalmers continually just asks, "Why?" Why is there consciousness rather than none? I think this is perfectly analogous to asking "why is there something rather than nothing?" But As Arne Naess pointed out, all worldviews have to start with some hypotheses. You can never get outside of everything in order to see everything. To claim that you can, is like trying to blow a balloon up from the inside. And Chalmers' infinite regression of "why" sure seems a balloon we can never get outside of.

So, I'd like to make a distinction for Chalmers' hard problem between the how and the why. How do physical processes lead to subjective experience? Why do physical processes lead to subjective experience? The ultimate why is ultimately an impossible problem. The how's along the way to that ultimate why may be difficult, but we can make progress with them. And they can tell us important things about life. Maybe it will turn out that consciousness—whatever we mean by that—will be fundamental to the universe in the way that electromagnetism is right now. Or maybe we'll find something else. But let's spend our time studying those hows, rather than getting caught up debating impossible whys.

Of course, there are other problems with objectively studying these "easy" problems of subjective consciousness. And that's what we'll look at next time.

What do you think? Is the hard problem of consciousness hard? Impossible? Easy? Or something else?

4 — Panpsychist Problems with Consciousness

21 March 2020

In the **last post**, I acknowledged that there may indeed be an impossible problem for studies of consciousness. David Chalmers makes his distinction here between the "easy problems" and the "hard problem", but by renaming them the "hows" and "the ultimate why", it becomes easier to see that Chalmers is really just playing the infinite regression game of why, why, why, why, why..... That is always an impossible loop to get out of, but we can still make efforts towards each new why whenever we find it useful and possible.

Before continuing down that path, however, we have to deal with an objection being raised that it is not in fact possible to study consciousness. This objection is currently being made by a philosopher of consciousness from Durham University in the UK named Philip Goff. If you'll remember from **post 2** in this series, Goff is a prominent proponent of panpsychism, which is the idea that *psyche* (mind) is *pan* (everywhere). Panpsychism is one of the concepts that physicalists / materialists like Sam and Annaka Harris are increasingly considering as a solution to the problem of how conscious entities arise from seemingly non-conscious materials. They think that maybe consciousness is just a fundamental attribute of the universe. I think we have a lot of investigating to do into our definitions and understanding of consciousness before we can make much sense of that claim, but Philip Goff doesn't think we can even do that. To best understand his point of view, I recommend reading an open exchange of letters ("**On the Problem of Consciousness, Panpsychism, and More**") which Goff had with the philosopher Massimo Pigliucci. Pigliucci has also been a professor of science in the fields of ecology and evolution, and he has written a book about how to distinguish between science and pseudoscience, so he is more than up for the task of debating Goff. Here are the most important points they made:

Philip Goff:

- 1st core issue: the problem of consciousness is radically unlike any other scientific problem. Perhaps the most obvious reason is that consciousness is unobservable. ... What we want a theory of consciousness to explain are the qualities of experience, e.g. the [*redness*] of red experiences. These qualities can only be known about by attending to experience from the 1st-person perspective; they are invisible to 3rd-person observation. This makes the problem of consciousness utterly unique: in every other scientific problem, we are trying to explain the data of 3rd-person observation.
- 2nd core issue: the case against materialism. There is something that needs explaining that can only be known about from the first-person perspective. We know that consciousness exists not from observation and experiment, but from our immediate

awareness of our own feelings and experiences. ... If the predicates of neuroscience could convey what it's like to see red, then a colour-blind neuroscientist would be able to know what it's like to see red by reading relevant neuroscience.

- 3rd core issue: is panpsychism coherent? Overall, I can't see any reason to doubt the coherence of the claim that experiential properties are the categorical properties underlying those dispositions.
- 4th core issue: why should we believe panpsychism? Panpsychism, I believe, is the simplest theory able to accommodate both 3rd-person observation and experiment, and the subjective qualities of experience. ... I think physical science alone cannot explain consciousness and hence we must turn to alternative ways of accounting for it.

Responses from Massimo Pigliucci:

- [1st core issue] Are you then discarding a lot of what psychology and cognitive science has done since the demise of behaviourism? Because part of the business of those sciences is to systematically study first-person phenomena, including people's intentions, motivations, emotions, and so forth. All of which are not directly observable and become data only via self-reporting. That has not been an obstacle to the scientific investigation of those phenomena, which we can even study experimentally, for instance, by inserting electrodes in the brain, or using localized magnetic stimulation and asking the subjects what they feel. Why you think this is an issue at all is beyond my comprehension, frankly. ... A scientific theory of consciousness—if we will have one—will provide a detailed mechanistic understanding of how the human brain generates first-person experience, using people's self-reports as data. Once we have that, there is nothing above and beyond it that requires further explanation. We would be done.
- [2nd core issue] What you call "knowledge of qualitative experience," and allege to be beyond scientific reach, I call experience. You are using "knowledge" in a very loose fashion. ... That would be a category mistake: we are talking about explaining the experience, not having it. ... Experiential knowledge is a different beast from theoretical knowledge. Science isn't going to give you the experience. ... It used to be that people would make the kind of argument you are putting forth to the effect that there was something special, irreducible to materialism, about life. They called it élan vital, vital essence. You are postulating the consciousness equivalent of an élan vital, for which there is no need.
- [3rd core issue] If by coherent you mean logically so, then sure, we agree. But literally an infinite number of models of the world are logically coherent. That doesn't help at all. ... You seem convinced that analytical metaphysics, the kind of approach developed in ancient Greece and that I would have thought died with Descartes, is still a valuable project. You are not the only one, of course; David Chalmers is another prominent advocate. But this is simply a rabbit hole that leads to an absurd proliferation of "coherent" or—worse yet—simply "conceivable" scenarios that tell us absolutely nothing about how the world actually works.
- [4th core issue] The issue is whether there is empirical reason to consider panpsychism. ... If you think that your theory does not, and cannot, make contact with empirical reality, then you simply don't have a theory. You have a speculation that can never be tested. ... There is absolutely nothing in modern physics or biology that hints at panpsychism, and you have acknowledged that no empirical evidence could possibly bear on the issue. That acknowledgement, for me, is the endpoint of our discussion. Once data are ruled out as arbiters among theories, those theories become pointless, just another clever intellectual game. ... The path you, Chalmers, and others are attempting to chart has already been tried, centuries ago, and has brought us—as David Hume put it—nothing but sophistry and illusion.

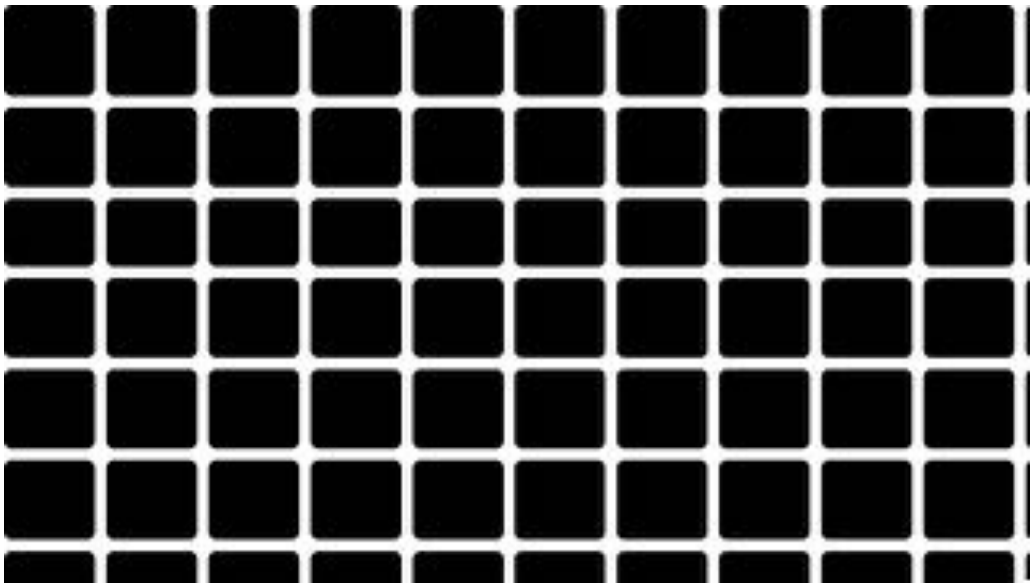
Brief Comments

I had the good fortune to meet Philip Goff recently when he gave a talk about these ideas to a small, local, Humanist group. He's a nice guy who is impressively well-versed on the literature of materialism and consciousness, but I have to say that his arguments strike me as deeply confused. His latest book for the general reader is titled *Galileo's Error: Foundations for a New Science of Consciousness*. But when I pressed him for how it could possibly be scientific if he also thinks the study of consciousness is empirically impossible, he admitted that was a question his editors asked, and he didn't have a good answer for them other than that we needed to rethink what we mean by science. I'm sorry, but that sounds exactly like pseudoscience, and Massimo did an excellent job of dismissing it.

To me, materialism / physicalism is still **a viable primary hypothesis**, and scientific investigations may yet find deeply sufficient explanations for consciousness in such a material universe. Goff worries that we can't get 3rd-person reports on consciousness for science, but that's literally true for everything. As a recent article in Scientific American pointed out ("**How to Make the Study of Consciousness Scientifically Tractable**"), there is no 3rd-person, objective, view from nowhere. "There is always a somewhere, a perspective, a subject." The key is realising that all progress in knowledge comes from "intersubjective confirmation". Naomi Oreskes called this "scientific consensus" in her latest book, *Why Trust Science?*, which I recently **reviewed**.

What do you think? Before we go on, are there other fundamental questions you have about studying consciousness? Or have we reached intersubjective confirmation that scientific consensus is possible? Let me know in the comments.

5 — Is Consciousness Just an Illusion?



Those dots aren't actually there. What else could be an illusion?

23 March 2020

In the **third post** in this series, David Chalmers mentioned that there has been an upsurge within consciousness philosophy towards a position called *illusionism*. In today's post, I want to begin to explore that position by listening to Keith Frankish, a leading proponent of illusionism. In an October 2019 episode of the Rationally Speaking podcast, Frankish discussed **Why Consciousness is an Illusion**. Here are the most important points from that discussion:

- The simplest way to put it is that [illusionism is] offering a different model of what consciousness is. This model rejects a central theory that dominates most people's thinking about consciousness. Consciousness in *that sense* is illusory and doesn't exist.
- Our common-sense view of what our inner experience is like is not as solid and reliable as we think. We tend to assume we encounter a presentation of the visual world that is full and complete in every detail right down to the periphery, but it turns out that is wrong. That is an illusion. That's an introspective illusion. Even in the matrix we would be having this illusion that we have a complete visual field. Once you allow that, you're opening a wedge here to the idea that introspection itself might be a construction.
- Let me say a bit more about the realist picture and just how odd this picture is. Dan Dennett calls this a sort of Cartesian theatre. The idea that there is this inner display of experience for conscious awareness. The outer world effectively creates this private cinema screen that we (and who are we?) witness. This kind of view of introspection does presuppose an introspect-or. That's one thing that needs to be hashed out.
- For Descartes this was easy because he envisioned an immaterial soul doing the witnessing, and it has special access. I suppose if there is an immaterial soul then all bets are off as to what it can do. But most philosophers now think it is just a brain. We aren't two things, but just an embodied brain.
- We are complicated creatures by any account, and we have some sort of self-awareness of our own mental processes, but it wouldn't be surprising if that picture weren't totally accurate. Why would nature have equipped us to be super-neuroscientists or to have a

super understanding of our own mental processes? We don't need that. Maybe we have something that's much more sketchy and caricatured.

- Here's a way I put this in a paper. These properties are anomalous [i.e. deviating from what is standard, normal, or expected]. They're not like other properties of the body like digestion, respiration, reproduction, etc. They're also not like other mental properties like emotion or other things that don't involve consciousness—we don't have good cognitive accounts of what's going on there. ... There are three approaches we could take to that. One is to say that yes they are anomalous and we've got to do some radical theorising to account for them. We have to do some heavy-duty metaphysics to say they're not a part of the physical world. Or perhaps (and this is gaining in popularity) they are a fundamental feature of reality, like the intrinsic nature of all matter is conscious in this way, or that all matter has this intrinsic phenomenal aspect to it. [That's panpsychism.]
- [*Digression from illusionism to consciousness in general:]*
- If you really want to be realist about consciousness, you've got to put it into the natural world somewhere, and it doesn't fit in easily. So, maybe, [panpsychism is] one way of getting consciousness into the natural world. Or maybe it just pops into existence when you get complex enough brains. That's a sort of emergence. You start where nature is building brains and the original ones don't have this phenomenal aspect to them, they just process information and get bigger and bigger and bigger. At some point between the first organisms and us, the lights came on inside. All that information processing, which was doing all the work, led to the lights coming on in a phenomenal aspect. Then the question is when did this happen? We can't be sure, because we can't strictly tell if other creatures, or indeed anyone else has this. There is a sort of arbitrariness here where things click on.
- Is this any more arbitrary than the fuzziness surrounding the definition of life? I think consciousness is worse than this in two ways.
- One is that there doesn't seem to be an in-between condition where there is a little bit of an interior world. Either there is something it is like to be something that has this first-person experience or there isn't. It may be very impoverished or boring for what it is like to be an electron or an amoeba or whatever, but it is still either or. It either does have this first-person experience or it doesn't. It's hard to see how it could have half a perspective. The inner lights are either on or off.
- Second, with life it's just difficult to specify what you count as life and what you don't. There is no hidden fact here; it's just what you say. It's a terminological issue. With consciousness, there are radically hidden facts. No matter how clearly we define this thing, we can't tell where it is and where it isn't. If someone says my cup has it, there is no test you can do to prove it.
- [*Returning to illusionism:*]
- Let me get back to the three broad positions you can take on this. ... [The third position] is a more conservative response that says we can kind of explain all this in terms of standard resources of cognitive science by talking about representations in the brain and maybe some sort of self-awareness. Maybe when we start to represent our own awareness to ourselves, that's when this apparent subjective experience comes in. ... That's been the standard view. Illusionism just goes a bit further. Yes, there are some sort of introspective mechanisms here, but what they are doing is misrepresenting their targets. It's not that these brain states, these targets, really *are* that. We have these simple, private, qualitative states. But actual brain states are much more complicated than that. Brain states merely present like that. And that is the illusion.
- Here is an analogy. In the Middle Ages, people thought other people were possessed by demons. Modern psychology gives a different explanation of what is happening. Now, do we say that's what demons really are? Schizophrenia is what demons are? Or do you say, "Stop thinking about it that way. Stop using the word demons altogether. This isn't an

explanation of what demons really are." That's what I'm asking us to do with consciousness.

- Some people start with the presumption that qualia is presented to us in a way that is immediate and transparent. They are revealed to us. There is nothing hidden about them. Just by having the experience, and attending to the experience, we can know the character of that property. I think it's pretty obvious that if that is your conception of the problem that needs to be explained, then science isn't going to help you with that. This presupposes a relationship between the subject and the object that you couldn't have in any physical conception of the world. To these people, to suggest that science has something to say here is to miss the point of the target for the whole debate.
- But we can reconfigure that. We can reconceptualise that; i.e. we are not hard wired to think that way. People who are into Buddhist philosophy tell me that this is what Buddhist thinkers have been doing for a long time. So, I think it's an open question about how able we are to shake off this picture.

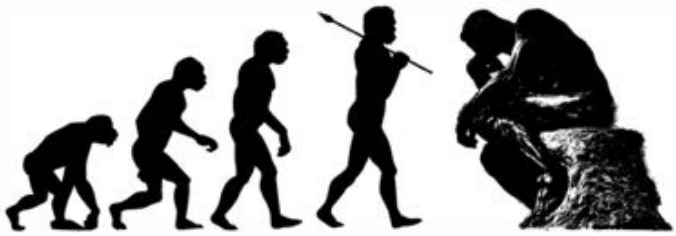
Brief Comments

As I noted in the **second post** of this series, Sam Harris does indeed use observations from his meditation practice to puncture the idea that consciousness is "presented to us in a way that is immediate and transparent." So, illusionism, while sounding pretty dire on the face of it, seems to be nothing more than the resting place for people who have dropped the supernatural attributes of consciousness, but haven't made the leap to panpsychism to explain it either. Illusionism doesn't say that conscious experience doesn't exist; just that it isn't what people generally think it is. This is important to note because there are still a lot of philosophers who ridicule illusionism by misunderstanding the position.

The problem I see with Frankish's view is that he's still talking about consciousness like an essentialist, talking as if it were one essence that either exists or does not. His claim that consciousness is either on or off seems deeply problematic in an evolving universe. But not all illusionists feel that way. What might they think is behind the illusion then? That will be the subject of the next two posts from perhaps the most famous developer of this idea—Dan Dennett.

What do you think? Are you more comfortable with dualism, panpsychism, or illusionism? Or do you have another name for your position here?

6 — Introducing an Evolutionary Perspective



25 March 2020

In the **last post**, I introduced illusionism using an interview with Keith Frankish. (Which he himself **retweeted**!) I mentioned that illusionists don't think our conscious experience *is* an illusion, just that our experience of it is papering over what's really going on behind the scenes. It's a little like pointing out that old projection movies give us the illusion of fluid motion on the screen when in reality there is just a series of still images flying by too fast for us to perceive. But what, then, is the reality behind our illusion of consciousness?

That's the question we'll be diving into for pretty much the rest of the series. Most of the research into that question has been done by neuroscientists, but before we get to them, there's one more pure philosopher we ought to consider to help set the stage, and that is Dan Dennett. Dennett has been prominently working on consciousness for decades. I'll be honest that I've never gone back and read his 1991 doorstopper ***Consciousness Explained***, but I figure his 2017 book ***From Bacteria to Bach and Back: The Evolution of Minds*** would supplant that now. Plus, the neuroscience has exploded since 1991, so why go back? The wikipedia entries that I've linked to for those books offer very quick summaries, but I'm going to focus in this post on **a Google Talk that Dennett gave about FBtBaB**. Here are the most important points from his hour-plus talk:

- The history of life is an R&D project, a design process that exploits information in the environment to create, maintain, and improve the design of things.
- R&D takes time and energy. There are two main types: evolution by natural selection, and human intelligent design. There have been intelligent designers for only about 100,000 years. You should not read back our intelligent design efforts into nature.
- Evolutionary design is purposeless, foresightless, extremely costly (99% of everything that ever lived died childless), and very slow. Intelligent design is purposeful, goal directed, somewhat foresighted, governed by cost considerations, and relatively fast.
- A termite mound might be 70 million clueless termites. A brain might be 86 billion clueless neurons. There are no captains, lieutenants, or generals in the brain. How [then] do you get a mind capable of intelligent design out of such a brain?
- Short answer: You can't do much carpentry with your bare hands, and you can't do much thinking with your bare brain. A termite colony is a bare brain. Intelligent designers have well-equipped brains. They have thinking tools.
- Long answer: Cultural evolution designed thinking tools that impose novel structures on our brains: virtual machines that could travel and be installed on different brains to give them powers they otherwise didn't have ("apps we download into our necktops").
- Darwin's strange inversion of reason: in order to make a perfect and beautiful machine, it is not necessary to know how to make it.

- Turing's strange inversion of reason: in order to be a perfect and beautiful computing machine, it is not necessary to know what arithmetic is.
- These yield Dennett's bumper sticker: Competence Without Comprehension
- The upshot of this is that the mind (consciousness, understanding, etc.) is the effect, not the cause. It's not a mind-first universe; it's a matter-first universe. Minds came recently.
- There's a difference between how come ("Why are planets spheres?") and what for ("Why are ball bearings spheres?"). The teleology of "what for" enters the world gradually. Darwin showed that it didn't always have to be there.
- Panpsychism is the view that everything is conscious. And I almost agree with it, but I just have to change the view a little bit. I call my view "pan-niftyism" — every atom is nifty, every electron is nifty. The question then is, is there any difference between panpsychism and pan-niftyism? They both explain the same thing—nothing. To say conscious things are made out of conscious things doesn't necessarily follow. Coloured things aren't made of coloured things.
- The (Paul) MacCready Explosion: 10,000 years ago, human population plus livestock and pets were approximately 0.1% of terrestrial vertebrate biomass. Today, it is 98%. This is probably the biggest, fastest, biological change on the planet ever. Genes don't explain it. Technology does.
- [The creation of] eukaryotic cells was one of the first great transfers of technology. A recent one is the invasion of human brains by symbiotic thinking tools called memes.
- These memes are "free floating reasons" as opposed to the reasons that saturate the biotic world. Trees, fungi, bacteria, non-human animals, etc. all do things for reasons. But they aren't aware of them. We can be.
- Bach was a top example of experimenting with purpose. He deeply understood his instruments and the history and theory of music in order to prolifically produce genius compositions. How [then] to get from blind genetic evolution to Bach?
- First step is synanthropic words. Synanthropic means things that thrive along with humans (e.g. seagulls, cockroaches, etc.). Nobody owned the first words; they were just habits that developed. [E.g. screeching for certain predators or specific dangers.]
- Next are domesticated words. Domesticated means the reproduction is controlled. For words, this means conscious choosing of one over the other. This leads to differential replication. Meanings or pronunciations can change over time, but the best ones survive, usually without even noticing why.
- The next step are coined words, deliberately designed, although their survival is still down to selection. Then there are technical terms, which are very carefully designed, and curated under strong group pressure. E.g. phenotype vs. genotype. These are hyper-domesticated words.
- This describes the age of intelligent design—ever-controlled more and more in a top-down method. Now, however, we are entering the age of post-intelligent design, where we have learned that the power of evolution is smarter than we are so we can create without comprehension. [Thus going from Bach back to bacteria.]

Brief Comments

I really don't have much to add to this other than that it's a good introduction to the ideas that complexity can arise very gradually without foresight, and the cultural evolution of language is a good hypothesis for providing an instrumental difference maker in the kind of minds that we humans have. If you are reading this post on a website called evolutionary philosophy, you probably already agree with this. But I wanted to stop and make this point specifically before l go on a deeper dive into Dennett's thoughts in the next post.

Oh and I had to share this talk because I loved Dennett's quip about pan-niftyism. That

surrender to explaining nothing is essentially my view of the panpsychists' project (if you can even call it a project). So, this post puts a nice bow on the end of that discussion too.

What do you think? Do you have any hesitations or questions about the role evolution can play in the history of that thing we currently call consciousness?

7 — More on Evolution

27 March 2020

In the **last post**, I introduced Dan Dennett's evolutionary perspective on consciousness. I mentioned that he's been working on this for decades, and during that time he has been a ...productive... philosopher to say the least. That sometimes makes him challenging to keep up with, but I personally think his quality is very high, so I wanted to spend one more post with him before making the transition to hearing from neuroscientists.

In this post, I'll be relying on another podcast with Sean Carroll — **Episode 78: Dan Dennett on Minds, Patterns, and the Scientific Image**. In a recent **January 2020 tweet**, Dan Dennett himself said that this was, "Another excellent interview, this time with Sean Carroll. If you haven't overdosed on Dennett in the last few days, this will clarify key points." Here, then, are some of those key clarifying points:

- [Do you have a simple definition of consciousness?] No. But that's okay. That's the way science works too. There's no perfect definition of time or energy, but scientists get on with it.
- Consciousness emerges (in the innocent sense, not the woo one), and the idea that consciousness is one thing, that everything in the universe is either conscious or not, that the light is either on or off—that is a fundamental error. But it is very widespread.
- The search for the simplest form of consciousness, therefore, is a snipe hunt. Starfish have some elements of consciousness, so do trees, and bacteria. (But not electrons.) We can argue about motor proteins. The question of "where do you draw the line?" is an ill-motivated question. Where do you draw the line between night and day?
- Electrons can't accrue memories. They do not change over billions of years. They do not participate in the arrow of time, so there is no way for them to be said to have intentions, feelings, purposes, or goals.
- Human consciousness is much different from the consciousness of other species. This is an embattled view, but I'm pretty sure of it. It's hard to see this because consciousness has a moral dimension and we want to be kind to animals. But don't worry. The conscious properties we share with mammals and birds, and to some degree with reptiles and fish, are significant. Moral significance itself is also a graded notion.
- UK law says it is now illegal to throw a live octopus onto a hot grill. This one species is an honorary vertebrate. It's not all cephalopods, although maybe it should be. Lobsters can

be boiled. Squid can be grilled live. Vertebrates must all be treated humanely. The law has to draw a line and these need to be reasonable to a vast majority of the people.

- Human minds are profoundly different from other minds, because they are obliged to articulate reasons. This is why I'm interested in the history and evolution of language.
- If I ask you to picture a rope and climbing up it, you can do it. I specifically chose those objects and actions because it is exactly what a chimp in a zoo is familiar with. If I asked a chimp to do the same thing, could it? We don't know, but I suspect not, because you can't do it wordlessly. You need to be able to interact using language. Without language, I don't think you have the cognitive systems for self-simulation and self-probing that we have. ... Language allows us to be conscious of things we otherwise wouldn't be able to be conscious of. If you believe that recursion and self-representation are crucial to consciousness, then language is a huge part of that as a useful tool.
- Degrees of freedom is something I'm using more lately. It is an opportunity for control. Degrees of freedom can be clamped or locked down to be removed. How many degrees of freedom do humans have? Millions and millions of things we can think of. We have orders of magnitude more that we can think of than a bear does, even with roughly the same number of cells. So, our complexity is higher. The options a bear has are a vanishing subset of the options that we have. Learning to control these options is not now a science. It is an art.
- Many theories of consciousness only have half of the theory. The upward stream. But what then? What does consciousness enable or take away from? The answer is that almost anything can happen [with consciousness]. But we need a neuroscientific theory as to how that happens.

Brief Comments

I can't say that Dennett puts a foot wrong here. His commitment to evolutionary thinking and following evidence leads to some conclusions that are out of step with much of society, but I find myself pretty much right there with him. I would question his point about electrons not having *any* elements of consciousness, but that's probably just based on terminology, and speculation that we may someday get from physics to chemistry to biology (where Dennett finds conscious elements). Without a good theory of **abiogenesis** (i.e. the origin of life), Dennett seems happy to pragmatically confine himself to studying consciousness *as if* it were a material phenomenon. I agree that's a useful hypothesis to hold until something better comes along.

I also really liked Dennett's use of the engineering terminology "degrees of freedom". This reminds me of "the parable of the immune system" that the evolutionary scientist David Sloan Wilson often uses to make a point. For example, on The Psychology Podcast (**Episode 167: Evolution and Contextual Behaviour Science**), Wilson said:

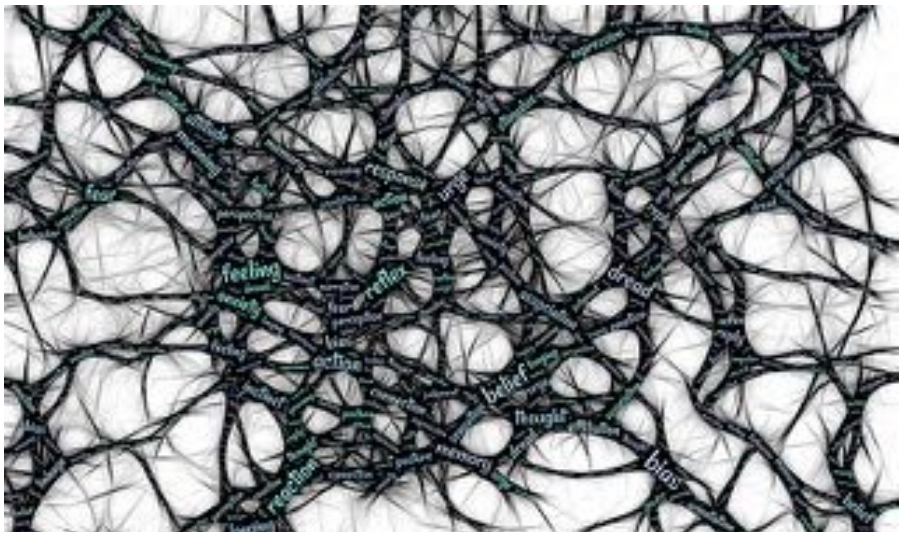
"The human immune system is immensely modular. We inherit it, and it does not change during our lifetime. It is something that evolved by genetic evolution, but it is triggered by environmental circumstances just as the evolutionary psychologists like to point out. The adaptive component of the immune system is highly evolutionary. That's the ability of antibodies to vary and for the successful antigens to be ramped up. So that's an evolutionary process that takes place during the lifetime of the organism. The whole thing is densely modular but also amazingly open-ended. Why can't we say the same thing about the human behavioural system?"

It seems obvious (to me anyway) that we *can* say the same thing about our behaviour—that it adapts during our lifetimes to successful and unsuccessful interactions with the environment. And it seems that more and more consciousness might give life more and more degrees of freedom as it helps an organism make more and more sense of its environment. But to really

consider that, we'll need to consider Dennett's questions, “But what then? What does consciousness enable or take away from?” And do to that, it's time to turn to the neuroscientific theories of consciousness being developed and explored by scientists.

What do you think? Does Dennett's evolutionary perspective continue to make sense? Are there any gaps in the story that need more explanation?

8 — Neurophilosophy



29 March 2020

As I make the transition in this series from philosophy to neuroscience, a natural step between these two disciplines (some might even call it an evolutionary step) would clearly be with the work of **Patricia Churchland**. She's a philosopher and neuroscientist who thinks that "philosophers are increasingly realising that to understand the mind one must understand the brain."

I'll start with a few snippets from the podcast Nous, and its episode: **Patricia Churchland on How We Evolved a Conscience**. Churchland has a book out now called ***Conscience: The Origins of Moral Intuition***, which I know is not the same thing as consciousness, but her discussion still has some relevant information for us.

- There were some philosophers who thought that if we went off and really studied the language of what we MEAN by the word consciousness, we'd be able to understand it. But other philosophers said, wait a minute, we might be mistaken about what we mean.
- Philosophy is a proto-science that must remain in touch with empirical discoveries. Science cannot tell us why something is right or wrong. However, science gives us all sorts of information that we take into account.
- Why did we become social? It started when we became warm blooded. Warm blooded creatures need about 10 times more nutrition though. One way to compensate for this requirement was for mammals to develop a new structure in the brain—a cortex—which allowed them to store a tremendous amount of information in the brain and to integrate it. The cortex relied on the subcortical parts of the brain for motivations, sleep/wake patterns, etc., but the cortex allowed for a kind of predictive prowess that had not been seen on the planet before.
- This all comes with a cost though. You can't have memory unless you can build structure on the neuron. To tune the brain up to an environment requires that you are super immature when you are born. Snakes just are born and go off into the world. Mammals can't. It was like evolution took a step backwards. This immaturity then led to the need for caregiving, which led to parents who care. Once caring for offspring turns on, family units, sociality, norms, and morality all take off.

That's a nice short, sharp, prod to get us philosophers studying the evolution of brains. A much more rigorous argument can be found in Churchland's essay "**Neurophilosophy**", which was a chapter in the fantastic edited collection ***How Biology Shapes Philosophy: New Foundations for Naturalism***. Here are some useful points from that essay:

- The words "mind" and "brain" are distinct. Even so, that linguistic fact leaves it open whether mental processes are in fact processes of the physical brain. … [For physicalists] the important problem concerns how the brain learns and remembers, how the brain enables us to see and hear and think, and how it enables us to move our eyes, legs, and whole body. Their problem concerns the nature of the brain mechanisms that support mental phenomena. Interestingly, dualists also have a closely related set of problems: how does soul stuff work such that we learn and remember, see and hear and think, and so forth. Whereas in neuroscience, physicalists have a vibrant research program to address such questions, dualists have no comparable program. No one has the slightest idea how soul stuff does anything.
- Studies of a few patients who had suffered bilateral damage to the hippocampus showed them to be severely impaired in learning new things. … Memory losses associated with dementing diseases also linked memory with neural loss and further suggested the tight link between the mental and the neural. Important also are studies of attention using brain imaging along with single neuron physiology. These varied studies suggest that at least three anatomic networks, connected but somewhat independent of the other, are involved in different aspects of attention: alerting, orienting, and executive control.
- Developments in psychology, especially visual psychology, also implicated neural networks in mental functions, and this work tended to dovetail well with neuroscientific findings on the visual system. Explanations of color vision, for example, depended on the retina's three cone types and on opponent processing by neurons in the cortical areas. … Visual hallucinations were known to be caused by physical substances such as LSD or ketamine, and consciousness could be obliterated by drugs such as ether, as well by other substances employed by anesthesiologists, such as propofol. No evidence linked these drugs to soul stuff.
- Short-term memory can be transiently blocked by a blow to the head or by a drug such as scopolamine; emotions and moods can be affected by Prozac and by alcohol; decision making can be affected by hunger, fear, sleeplessness, and cocaine; elevated levels of cortisol cause anxiety. Very specific changes in whole-brain activity corresponding to periods of sleep versus dreaming versus being awake have been documented, and explanations for the neuronal signature typifying these three states have made considerable progress. In aggregate, these findings weighed in favor of the physical brain, not of some spooky "soul stuff."
- A methodological point may be pertinent in regard to the dualist's argument: however large and systematic the mass of empirical evidence supporting the empirical hypothesis that consciousness is a brain function, it is always a logically consistent option to be stubborn and to insist otherwise, as do Chalmers and Nagel. Here is the way to think about this: identities—such as that temperature really is mean molecular kinetic energy, for example—are not directly observable. They are underwritten by inferences that best account for the mass of data and the appreciation that no explanatory competitor is as successful. One could, if determined, dig one's heels in and say, "temperature is not mean molecular kinetic energy, but rather an occult phenomenon that merely runs parallel to KE." It is a logically consistent position, even if it is not a reasonable position.
- With the benefit of contemporary physics, we can see that the causal interaction between nonphysical stuff such as a soul with physical stuff such as electrons would be an anomaly relative to the current and rather well-established laws of physics. More exactly, it would

affect the law of conservation of energy. If brains can cause changes external to the physical domain, there should be an anomaly with respect to conservation of energy. No such anomaly has ever been seen or measured.

Brief Comments

In previous posts, we saw how argument alone could make the case that thinking of consciousness as a non-material or panpsychic phenomenon is not helpful. Now, we see a glut of empirical evidence supporting the idea that consciousness is a physical phenomenon. Does that prove the case? Of course not. Knowledge is never proved in this way. Churchland's point, however, is exquisite, and right on the nose, that one can always dig their heels in about this and remain *consistent*, while also being unreasonable. This is something all philosophers should keep in mind.

What do you think? Any other important points jump out at you from these quotes?

9 — Global Neuronal Workspace Theory

Stanislas Dehaene — if I'm implicitly biased towards him, I now know why.

31 March 2020

For the rest of the research in this series, I'm going to be going over the work of neuroscientists. This is because, as Patricia Churchland stated in the **last post**, "Philosophy is a proto-science that must remain in touch with empirical discoveries." As a philosopher, however, my goal here is not to gain or present a detailed lesson of all the most complicated inner workings of the brain (read neuroscientists for that). Nor is it to get into a deep debate about the methodologies, assumptions, and conclusions of the people working in this field (read philosophers of science for that). What I'm looking for in this series are findings or hypotheses which have implications for the rest of my philosophical worldview. Is that going to require *some* knowledge of brain anatomy and mechanisms? Yes. But it's not that scary or difficult.

One of the best guides for this world is Dr. Ginger Campbell, whose podcast **Brain Science** is up to 170 episodes now as of this post. Recently, Campbell posted an incredible four-part series on consciousness that was really a key inspiration for me to finally tackle this subject as well. In the first of these podcasts (called **What is Consciousness?**), Campbell gave her own summaries of some of the latest and best books on consciousness. Before she dives into them, Campbell notes that while they do have their differences, there are still three concepts they all share:

1. Consciousness requires a brain
2. Consciousness is a product of evolution
3. Consciousness is embodied

While I'm always happy to hear from people with an evolutionary perspective, previous posts in this series make it clear that there are enough quibbles about the term "consciousness" to remain wary of saying it is a coherent enough concept to deserve a label. That throws into question whether a brain is required for it or not. But, if you grant each neuroscientist their hypothetical definition of consciousness, then we can understand what they are talking about and the rest of their claims remain valid within that perspective.

Okay. Time for the first summary. Campbell kicked off her series by discussing Stanislas Dehaene's book ***Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts***. Here are the most important points from that:

- Three key ingredients were required to move the study of consciousness into the lab: 1) a better definition of consciousness; 2) methods to manipulate consciousness experimentally; and 3) a new respect for the study of subjective phenomena (compared to behaviourism).
- The definition Dehaene uses is called Global Neuronal Workspace Theory (an offshoot of Bernard Baars' **Global Workspace Theory**)
- GNWT states: consciousness is global information broadcasting within the cortex.
- Consciousness adds functionality, ability to hold information in mind, and flexible behaviour.
- Wakefulness, vigilance, and attention enable conscious access, but they are separate things.
- Some of the main methods used to study this are: binocular rivalry, attentional blink, and masking.
- No amount of introspection can tell us how our brain works.
- Most of what our brain does is outside of our conscious access. Many phenomena do not require consciousness to occur. We drastically underestimate this.
- If our brains can do so much without consciousness, then what is it for?
- Brains make unconscious predictions as if they were using Beysian logic, but seem to need consciousness to interpret ambiguous images. Also, consciousness plays a very important role in learning (e.g. subliminal learning doesn't work).
- Only consciousness allows us to entertain lasting thoughts. It also allows us to create algorithms, a step-by-step way of solving a problem. It allows for flexible routing of information, and appears to be necessary for making a final decision.
- Consciousness is an important element of social information sharing. It condenses information, [making it easier to transfer].
- Our self is just a database that is filled through social experience. Consciousness is the mind's reality simulator.
- When conscious access occurs: brain activity is strongly activated when a threshold of awareness is crossed. At that point the signal spreads to many brain areas. There are four highly reproducible signals associated with this. Signature 1: activation in parietal and prefrontal circuits. Signature 2: a slow web called P3 that pairs late, approximately 1/3 sec after stimulus (i.e. consciousness lags behind the world). Signature 3: deep brain electrodes detect late and sudden bursts of high frequency oscillations. Signature 4: information exchange across distant brain areas.
- Virtually every circuit in the brain, cortical and subcortical, can participate in conscious and unconscious processes.
- In Global Neuronal Workspace Theory, conscious access occurs when perception, or any other signal, crosses a critical threshold and is broadcast across the brain.
- 50 milliseconds seems to be a limit for the shortest exposure to a signal that we can detect.
- We can only perceive one signal at a time. And there is a 1/3 second time lag. Error prediction makes up for this.
- Consequences of consciousness include: the ability to respond, the ability to hold ideas in our mind, and the ability to act flexibly.
- Dehaene does not show mere "correlates of consciousness" because correlation does not show causation. Correlation just finds things that are present when consciousness is perceived, and absent when it is not perceived. Dehaene's four signatures fit this.

Causation would require recreation of conscious states using artificial means and this is now being done using deep brain stimulation.
- Higher brain regions do appear to be essential.
- Putting together all the evidence inescapably leads to a reductionist conclusion. The electrical activities of neurones can create a state of mind, or equally destroy an existing one.
- Dehaene thinks Chalmers swapped the labels. It is the easy problem that is hard, while the hard problem seems hard because it engages ill-defined intuitions. Once our intuition is educated by cognitive neuroscience and computer simulations, he thinks Chalmers' hard problem will evaporate.

Brief Comments

Dehaene offers lots of persuasive evidence for the brain activities that occur during events that we humans can report (i.e. conscious vs. unconscious activities). It is fascinating to see the list of functions this enables as that presumably provides some guides about what is likely to have evolved later as the long evolutionary history of consciousness has unfolded. For example, it seems plain to me that there would be a massive evolutionary advantage for a brain to be able to predict reality rather than wait 1/3 of a second for the processing of inputs. So far, that seems like a good candidate to help answer the question of what consciousness is for. I'll wait to look at more evidence from other scientists, though, before proclaiming too much. Stick around for that in the next few posts.

What do you think? Is Chalmers' hard problem fading away as our understanding of the correlates of consciousness grows? Or as we even begin to dabble in the causation of our conscious experience itself? If this is all too new or confusing to give an answer to that, I recommend trying a short video on **Global vs. Local Theories** that is part of a recently released introductory course on the brain and consciousness. Let me know in the comments if that helps or if anything else would.

10 — Mind + Self

Photo by Alberto Gamazo (https://is.gd/KVwanB)

2 April 2020

In the **last post**, I noted that I was going to be relying on Dr. Ginger Campbell's **Brain Science** podcast for summaries of the latest work on consciousness by neuroscientists. She kicked off her recent four-part series on consciousness with an episode called **What is Consciousness?** where she gave summaries of some of the latest and best books on this subject. Three of the five books she covered were written by neuroscientists. (The other two were by Sean Carroll and Dan Dennett who I've already covered.) The first of those was by Stanislas Dehaene, which I discussed in the last post. Next up, is Antonio Damasio's book ***The Strange Order of Things: Life, Feeling, and the Making of Cultures***. Here are the most important points from that:

- Damasio defines consciousness as: mind + self.
- A mind emerges from the brain when an animal is able to create images and to map the world and its body.
- Consciousness requires the addition of self-awareness. This begins at the level of the brain stem, with "primordial feelings." The self is built up in stages starting with the proto self made up of primordial feelings, affect alone, and feeling alive. Then the core self is developed when the proto self is interacting with objects and images such that they are modified and there is a narrative sequence. Finally comes the autobiographical self, which is built from the lived past and the anticipated future.
- Mind precedes consciousness.
- Consciousness includes wakefulness, mind, and self.
- Consciousness is the feeling that my body exists independent of other objects.
- Affect or feelings came first. Long before consciousness. (*A la* Panksepp.) Feelings evolve from homeostatic signals and so affect evolved very early. Damasio called this "the strange order of things" because it's the opposite of what many scientists assume.
- Damasio stresses the importance of embodiment because homeostasis is the primary mechanism driving life. Feelings are mental experiences that are conscious by

definition. The emotive response triggered by sensory stimuli are the qualia of philosophical tradition. This subjectivity is the critical enabler of consciousness.
- Emotions are chemical reactions. Feelings are the conscious experience of emotions. (This can be slightly confusing as it is not always used consistently in Damasio's work.)
- Early life was regulated without feelings and there was no mind or consciousness. Then, during the Cambrian explosion, vertebrates appeared and all vertebrates have feelings.
- Valence / value evolved much earlier. Even bacteria can go toward food and away from danger.
- Feelings are not neural events alone. They are interpretations of body signals (such as a fast heartbeat). Feelings are, through-and-through, simultaneously, and interestingly, phenomena of both bodies and nervous systems.

For just a bit more on this, Antonio Damasio gave a TED talk in 2011 called, **The quest to understand consciousness**. Here are a few extra details he used during this talk:

- Three levels of self to consider: proto self, core self, and autobiographical self.
- Autobiographical self has prompted: extended memory, reasoning, imagination, creativity, and language.
- Out of these came the instruments of culture: religions, justice, trade, the arts, science, and technology.

Brief Comments

I may be jumping the gun here, but Damasio's distinction between the mind and the self appear to me to map neatly onto **the two brain networks scientists just proved are key to consciousness**. The DAT (dorsal attention network) sounds like it produces the streaming images of the outside world, which Damasio calls mind. And the DMN (default mode network) monitors the internal states of our bodies, generating the sense of a relatively stable but historically changing identity, which Damasio calls the self. As the article I linked to says, consciousness is reported when the DAT and DMN are both activated. In other words, when both mind and self are active. This is something to consider as we go forward. (And, by the way, **default mode networks have been detected in macaques, chimpanzees, and even rats**.)

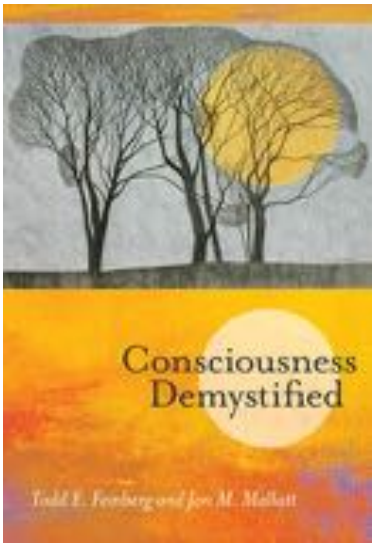
I also like Damasio's distinctions between emotions, feelings, and valences. This fits very well with **my own system** for mapping cognitive appraisals (i.e. judging if something is good, bad, or unknown, aka valenced) onto different events in the past, present, or future, in order to generate the things we typically call emotions (but which Damasio would distinguish as feelings). I can certainly get behind his distinction here. I could also adopt his labelling. And I think he's got "the strange order of things" right by saying the chemical emotional responses would have come first before the feelings in our self became able to identify them. This would clearly be the order of things in a material universe where physics led to chemistry, biology, and then psychology. This is another thing to consider as we put together the evolutionary story of consciousness.

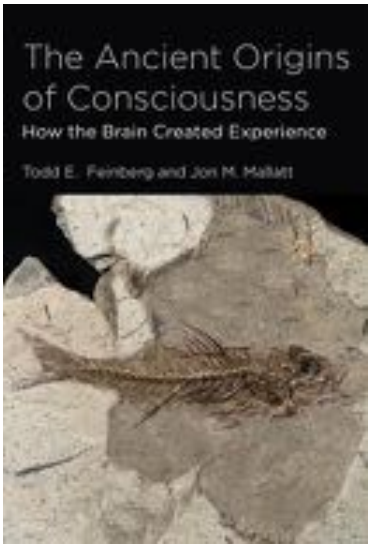
Finally, I'll just explain the brief reference Damasio made to Panksepp. In my first peer-reviewed philosophy paper about **Bridging the Is-Ought Divide**, I mentioned Panksepp's work when I said: "Evolutionary neuroscientist Jaak Panksepp of Bowling Green State University has identified seven emotional systems in humans that originated deeper in our evolutionary past than the Pleistocene era. The emotional systems that Panskepp terms Care (tenderness for others), Panic (from loneliness), and Play (social joy) date back to early primate evolutionary history, whereas the systems of Fear, Rage, Seeking, and Lust, which govern

survival instincts for the individual, have even earlier, premammalian origins." I cited this work as potential evidence for the evolution of morality from care of the self to care for others, but of course it is also evidence of the development of the concept of the self too.

What do you think? Do Damasio's distinctions make sense to you? Do they map onto concepts you find helpful or not? Let me know what you think of this in the comments.

11 —Neurobiological Naturalism





Two books that look pretty applicable to this series...

4 April 2020

In the last post, I mentioned that Dr. Ginger Campbell reviewed three books about consciousness in her magnificent **Brain Science** podcast that were written by neuroscientists. The first two were written by Stanislas Dehaene and Antonio Damasio, which I covered in the last two posts. Now, we get to a book written by Todd Feinberg and Jon Mallatt called ***Consciousness Demystified***. This is their most recent book, published in 2018, so that's the one Campbell covered in depth. However, since this is a refined and perhaps popularised version of the book they published in 2016 called ***The Ancient Origins of Consciousness*** (which sure sounds appropriate for this series), I thought I should pull a couple of summary points from that book too. Here, then, are the most important items I found:

- Feinberg and Mallatt use a much broader view of consciousness than Dehaene or Damasio.
- They use the term "neurobiological naturalism" to address the hard problem, which is an elaboration of John Searle's **biological naturalism**.
- F&M's goal is to bridge the gap between what the brain does and subjective experience.
- Neurobiological naturalism rests on three principles: 1) Life. F&M say consciousness is grounded in the unique features of life. 2) Neural features. This consciousness correlates with neural activity. 3) Naturalistic manner. Nothing supernatural is needed.
- Primary consciousness is broken down into three elements: 1) Exteroceptive—Damasio's mapping of the outer world. 2) Interoceptive—signals from inside the body. 3) Affective—the experience of feeling, emotion, or mood.
- The intercommunicating axons of affective pathways branch a lot more than in the exteroceptive pathways, sending signals to many different parts of the system. Another difference is that affective circuits communicate less through short-distance neurotransmitter chemicals and more through far-diffusing neuromodulator chemicals than do exteroceptive circuits.
- Four problems arise then: 1) Referral—we don't experience anything inside our brain. It's all referred to from the outside world or from our bodies. 2) Mental unity—how is it all

put together into a single experience. 3) Mental causation—how do thoughts cause action. 4) The perceived qualia of objects.

- Breaking the hard problem into four smaller problems makes things more manageable.
- E.g. mental unity is a process, not locatable to a single brain region. It requires synchronised oscillations to unify multiple networks.
- There is evidence that all vertebrates and some invertebrates enjoy consciousness. This is from a combination of anatomical and behavioural evidence, including operant learning.
- F&M see qualia (subjective experience) as having two unique features: 1) a unique neurobiology; and 2) the fact that it is exclusively first-person. So, therefore, we need two answers. They argue that the first person subjectivity comes from 1) the life process, combined with 2) the neurobiological pathways.
- Responding to Chalmers' famous question "Why is experience one way rather than another?" they write: "Our theory of neurobiological naturalism argues that animal experience is fundamentally and inextricably built on the foundation of life. Therefore, we must distinguish purely computational mechanisms, for example computers and any other known non-living computational device, as well as cognitive theories of consciousness that likewise centre on information processing, from the theories that invoke the biological and neural properties of a living brain. We hypothesise that experience and qualia are living processes that cannot be explained solely by non-biological computation. Our view of the hard problem begins and rests on the essential role that biology plays in making animal experience and qualia possible."
- There are several keys to the mystery of consciousness and subjective experience. One is that consciousness is incredibly diverse, coming from a multi-factorial combination of life and various unique neurobiological structures and processes. They also argue that qualia should not be treated as a single thing and that subjective experiences emerge when a sufficient level of neural complexity evolves. They argue repeatedly that the neurobiological problems should NOT be conflated with the philosophical problem.

- In *The Ancient Origins of Consciousness*, Feinberg and Mallatt conted that consciousness is about creating image maps of the environment and oneself. But systems that do it with orders of magnitude less sophistication than humans can still trigger our intuition of a fellow conscious being.
- After assembling a list of the biological and neurobiological features that seem responsible for consciousness, and considering the fossil record of evolution, Feinberg and Mallatt argue that consciousness appeared much earlier in evolutionary history than is commonly assumed. About 520 to 560 million years ago, they explain, the great "Cambrian explosion" of animal diversity produced the first complex brains, which were accompanied by the first appearance of consciousness. Simple reflexive behaviours evolved into a unified inner world of subjective experiences. From this they deduce that all vertebrates are and have always been conscious—not just humans and other mammals, but also every fish, reptile, amphibian, and bird. Considering invertebrates, they find that arthropods (including insects and probably crustaceans) and cephalopods (including the octopus) meet many of the criteria for consciousness. The obvious and conventional wisdom–shattering implication is that consciousness evolved simultaneously but independently in the first vertebrates and possibly arthropods more than half a billion years ago.
- To Feinberg and Mallatt, real consciousness is indicated by the optic tectum making a multi-sensory map of the world, attending to the most important object in this map, and then signalling behaviours based on the map.
- Isomorphic maps are the cornerstone of image-based sensory consciousness. These maps evolved in early vertebrates more than 520 million years ago, and this process was the

natural result of the extraordinary innovations of the camera eye, neural crest, and placodes. These events led to the mental images that mark the creation of the mysterious explanatory gaps and the subjective features of consciousness.

- The Defining Features of Consciousness are: Level 1) General Biological Features: life, embodiment, processes, self-organising systems, emergence, teleonomy, and adaption. Level 2) Reflexes of animals with nervous systems. Level 3) Special Neurobiological Features: complex hierarchy (of networks); nested and non-nested processes, aka recursive; isomorphic representations and mental images; affective states; attention; and memory.
- *The Ancient Origins of Consciousness* does not address higher levels of consciousness: full-blown self-awareness, meta-awareness, recognition of the self in mirrors, theory of mind, access to verbal self-reporting.

Brief Comments

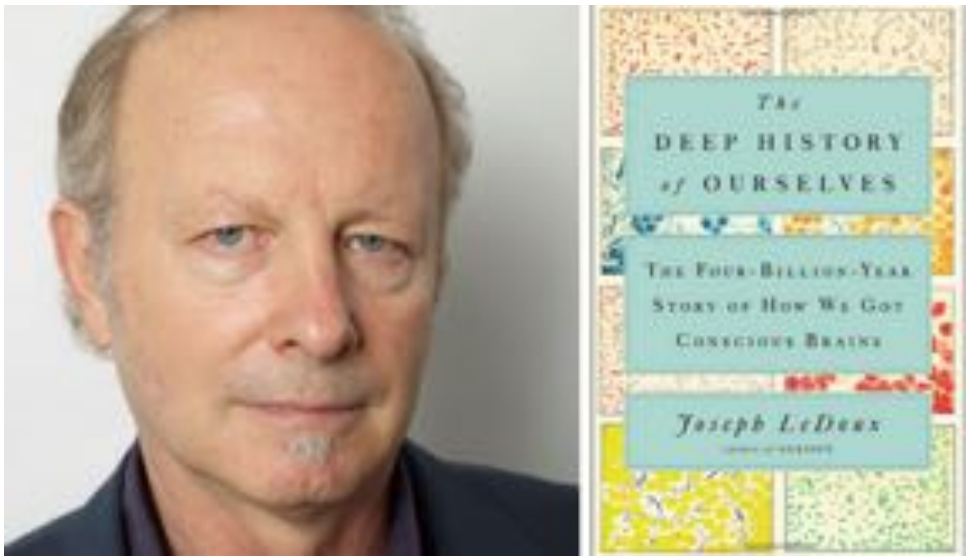
These books are apparently rammed full of good details about the internal brain structures involved with lots of discretely-named aspects of consciousness, and the evolutionary history of these anatomical features. That's certainly helpful for my project. However, the philosopher in me can't also help agreeing with the top Amazon review for *Consciousness Demystified*, which called it a disappointing bait and switch. The reviewer said, "In other words, in spite of their stated 'main goal' to address the explanatory gap between a third-person, objective description of how the brain works and the mystery of why that gives rise to (or amounts to) subjective, conscious experience, in fact they finally conclude that this explanatory gap is only a 'philosophical problem' instead of a 'neurobiological problem' and thus not really what their book was ever intended to explain anyway."

I have already gone over how the "philosophical problem" raised by Chalmers is actually an impossible problem so it doesn't bother me that Feinberg and Mallatt didn't tackle it. But by naming their books as they have, and promising early on to clear up the so-called hard problem, Feinberg and Mallat have disappointed more than a few readers. Then, by merely asserting that consciousness only arises from natural living processes, they lose credibility by failing to acknowledge (**as Searle did**) the possibility that alternate arrangements of matter, other than biological brains, could bring forth consciousness. While I'd still put money on the uniqueness of biology leading to the uniqueness of the consciousness that we recognise (think about how that consciousness changes for tiny changes in the biology), I don't pretend that this is a sure bet.

Feinberg and Mallat's addition of "affect" to the mix of "exteroception" (what Damasio calls *mind*) and "interoception" (what Damasio calls *self*) is interesting, but probably due to their expanded conception of consciousness. I agree with them it is certainly something that is a part of this full range of experiences that can get lumped into "consciousness", but the note about how the affective circuits communicate "through far-diffusing neuromodulator chemicals" reminds me of the brain being *awash in an emotion*, which presumably Damasio would say can occur in a non-conscious fashion, which is why it is not a part of his more limited definition of consciousness.

What do you think? Did anything else in Feinberg and Mallatt's research or hypotheses add to your thinking about consciousness? As always, let me know in the comments.

12 — The Deep History of Ourselves



Good thing that soul patch is only used in one of his specialties.

6 April 2020

We're in the home stretch now for this series on consciousness. In the last three posts, I went over the summaries of books that Dr. Ginger Campbell provided in one of her **Brain Science** podcast episodes. That one episode was particularly useful, but it was just the first of a four-part series on consciousness. The next three episodes were one-on-one interviews with three more neuroscientists about their own studies of consciousness. Those interviews will provide the last three pieces of external research for my series.

The **first interview** was with Joseph LeDoux about his book ***The Deep History of Ourselves: The Four-Billion-Year Story of How We Got Conscious Brains***. What a great evolutionary title! LeDoux is a Professor of Neural Science and Psychology at NYU who has spent the last thirty years studying the brain mechanisms of fear and emotional memory. He's also the guitarist and songwriter for a funky band called The Amygdaloids who gave us the hep-cat, jazzy, yet informative little number **Fearing**. (Pretty awesome.) For a more straightforward lesson about consciousness, however, here are the highlights from LeDoux's interview with Dr. Campbell:

- **Higher-order representations** is the category LeDoux prefers from among the 20 different theories of consciousness.
- How far back in evolution does the ability to detect and respond to danger go? Other nonhuman animals do this. Even bees. But it's much older still. Protozoa like paramecia or amoeba do it. Even bacteria do. In fact, it goes all the way back to the beginning of life.
- It's not just detecting danger either — incorporating nutrients, balancing fluids and ions, thermoregulation, reproduction for the species to survive — all of these behaviours exist in animals, but also in single-cell microbes. Value / valence / affect has also been present since the beginning of life (e.g. bacteria swim toward or away from things).
- So, behaviour and even learning and memory do not require nervous systems.
- When we do those things, we have subjective experiences about them, but those subjective experiences are not essential to the actions.

- What is the relationship between behaviour and consciousness? We see behaviour in others so we attribute the same thoughts and feelings that we do. This makes sense for other human brains, but it is more and more dissimilar for other brains.
- When we detect danger, we feel fear. But that may not always be the case. Split brain cases show one side getting a signal, the body acts, but then the other side can't say why.
- I hypothesised that emotional systems could generate non-conscious behaviours. I was able to trace the pathways through the amygdala to do this. Other research showed the amygdala is involved in implicit / non-conscious memories as opposed to conscious memories about detecting and responding to danger. I used this model for memories and applied it to emotions—i.e. implicit vs. explicit emotions. I thought of conscious explicit emotions as the product of cortical areas. Non-conscious emotions come out of the amygdala. The amygdala doesn't experience fear; it just produces responses.
- When stimuli are presented to patients, but masked so they can't detect it consciously, the visual cortex and amygdala are activated and that's it. When the stimulus is not masked, you get activation in the visual cortex, the amygdala, and the prefrontal cortex as well. ... In order to be conscious of an apple, it not only needs to be represented in your visual cortex, it needs to be re-represented, which involves the prefrontal cortex. ... So, the prefrontal cortex is emerging as an important area in the consolidation of our conscious experiences into what they are.
- In other words, the ability to respond to and detect danger may be as old as life, but the feeling of fear may be a much more recent addition.
- [Here's my 1st crazy idea.] What came first was cognition not emotion. I'm defining cognition as the ability to form internal representations of stimuli and to perform behaviours based on those representations. Cues are enough to stimulate the behaviour independent of the presence of the stimuli themselves. The representation alone is enough to guide the behaviour. That capacity exists in invertebrates, and on into all vertebrates, e.g. fish and reptiles. When you get to mammals, you have a much more complex form of cognitive representation, where it begins to look deliberative, i.e. the ability to form mental models that can be predictive of things not existing. It's a much more complicated thing than having a static memory of what was there.
- We assume that because mammals behave in much the same way that we do, they must be experiencing the same things. But the amygdala example of fear gives us some reason to be cautious about that. The short summary is that you should actually assume behaviour is unconscious unless proven otherwise.
- In humans, we all know that we have these conscious experiences. In an experiment, we ask, "Can the response in this experiment be explained by a conscious state?" We have to rule out that the response is not coming from a non-conscious state. But we have a vast cognitive unconscious repository of information that allows us to get through the day without having to consciously evaluate everything we do (e.g. speaking grammatically, anticipating what we are looking at before we see it, completing patterns on the basis of limited information). To separate these conscious and non-conscious responses you can do experiments, and these have indeed happened.
- The gold standard for whether a response is conscious or not is whether you can talk about it. This doesn't mean language and consciousness are identical, just that you have access to the experience to think about it (and we use language to discuss that access with one another). In non-human animal research, that doesn't exist. It would be good for animals to treat them *as if* they had conscious experiences, but it's not a scientific demonstration to watch behaviour and say that they do.
- Darwin, when faced with resistance about humans evolving from animals, responded not by saying that people have bestial qualities, but by saying that animals have human qualities. This set the debate on a track that has been difficult to get past. There was

tremendous anthropomorphism in the late 19th century. That led to the radical behaviourist movement in psychology where all cognitive experience was eliminated from research. The cognitive revolution brought back the mind, but as an information processing system with inputs being conscious and unconscious. This gave us the "cognitive unconscious", which is a middle ground between the choice the behaviourists gave us between conscious vs. reflex machines.

- Anthropomorphism may be an important innate human quality, but that doesn't mean it's an accurate concept. And maybe we just can't know either.
- As a brief aside, usages of the limbic system, triune brain, and serial evolution of additive brain functions are all outdated now.
- [Here's my 2nd crazy idea.] Emotions are not initially a product of natural selection. Emotions are conscious experiences constructed by cognitive processes. The possibility then exists that the cognitive abilities that are unique in the human brain might be responsible for those emotions. Maybe emotions came in with the early humans. Maybe they came in as byproducts, or what Stephen J. Gould called exaptations. If this cognitive model is correct, then emotions are based on mental schema (bodies of memories about certain categories of experiences), for example, a fear schema. When in danger, a template is activated. This has implications for medicine to treat emotions. For example, people taking medicine for social anxiety find it easy to go to parties (they are less timid), but they still feel anxious when there. ... Drugs alone won't be enough to treat problems. Cognitive Behavioural Therapy is required in the end.
- A particular human experience is where you know the experience is happening to you. We can't rule that out in other animals, but neurological evidence suggests that it's not happening. This "autonoetic consciousness" represents the view of the self as the subject. It enables mental time-travel (i.e. you can review past experiences and possible future states). Other animals can learn from the past, but in a simple way. They can also have shifts in perspectives to those of others, but they don't have this notion of the self that is part of these experiences. Non-conscious alternatives can always account for the behaviour in animals.
- Every person has the same human brain. There are things in our prefrontal cortex, structures ("frontal pole"), and connections that are unique to humans. But mice have their own unique brain area. Other animals may also have their own unique ways of experience. We have to be subtle and not simply say conscious or nonconscious. Consciousness isn't one thing. There's autonoetic consciousness. There's noetic consciousness (an awareness of facts and the world). Working memory, for example, is very similar in other primates but not other mammals. There's anoetic consciousness, which is a body awareness (i.e. Jaak Panksepp's core consciousness, which is a primitive, almost unconscious level of consciousness). Understanding brain structures and pathways might help us understand what forms of consciousness are possible, even if we can never measure it.

Brief Comments

LeDoux seems to draw a pretty narrow definition around consciousness, but then shows the clear evolutionary history of *aspects* of consciousness along the way, and really advocates for a more subtle use of the term. I'll present my own subjective labelling system for all this at the end of the series (because we sure could use another!), but hopefully the contents of facts within that system will be uncontroversial, and they will surely draw on LeDoux's work.

Like Damasio, whose strange inversion was that emotions preceded feelings, LeDoux's first crazy idea is his own inversion, where he says cognition preceded emotion. In one respect, these guys are actually saying the same thing, that the "subjective experience of moods" came

last. But Damasio calls that "feelings" while LeDoux calls it "emotion". Clearly there is a split here between the chemical changes that cause behaviour, and the subjective experience of these changes, but it's frustrating that the field hasn't settled on consistent terminology yet of what's on each side of this divide, which makes discussing these ideas so much more difficult. (It's another good example of the value that philosophers of science can be to scientists.)

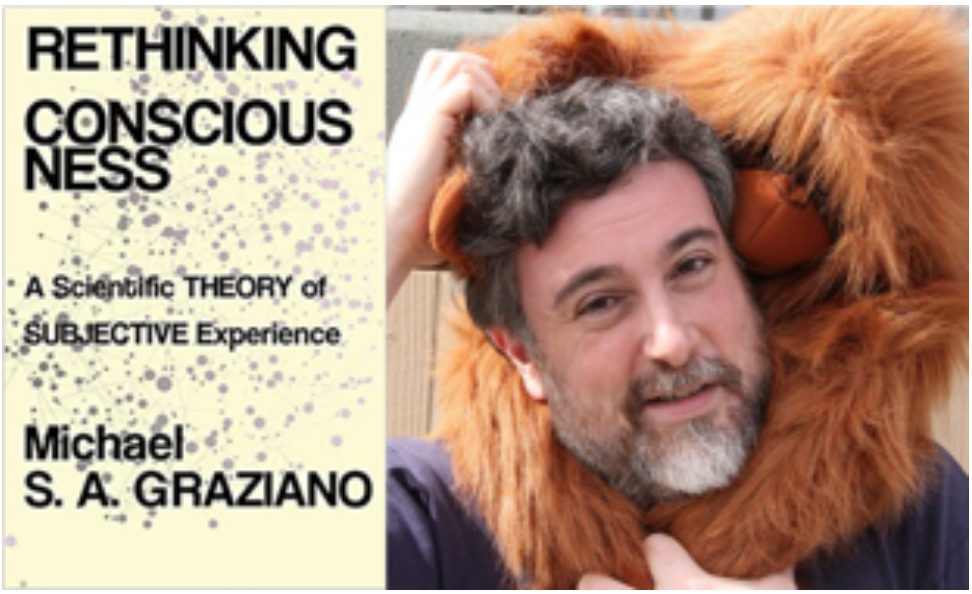
What I don't see from LeDoux in this crazy idea is any discussion of affect or value. The amygdala may be able to non-consciously produce behaviour in response to stimuli. It may even learn to do this differently throughout a lifetime. But it could only do so (successfully) by valuing some responses positively and others negatively. Since LeDoux does state that valence goes all the way back to the beginning of life, maybe he just lumps this in as part of "cognition", which then looks even more like Damasio's "emotions", which both men claim came first during evolution.

As for LeDoux's second crazy idea, it's hard for me to see how he can advocate for the need for Cognitive Behavioural Therapy to regulate emotional feelings, but then suggest that these emotional feelings weren't initially a product of natural selection. Perhaps it comes down to how narrowly one defines "initially" but if CBT can improve one's life, then it sure seems plausible that the advent of emotional feelings would have provided an advantage that could have been selected for. Maybe I'm just being overly critical of anyone quoting Gould, though, since I'm of the opinion that he generally lost **the Darwin Wars**.

Finally, as an evolutionary thinker, I note that LeDoux offers a really good critique of anthropomorphism and the role that Darwin may have played in going down that path. Such attributions to non-human animals can obviously be taken too far. But so can anthropodenial (**as Franz de Waal has coined it**) for the people who go in the other direction and tout human exceptionalism. I really appreciate LeDoux's openness about this and his search for hard evidence. I also like his recognition that it would be better for us to treat animals *as if* they had valuable internal experiences, since we are currently faced with the barrier that we may never know about that. So, one form of human exceptionalism that exists may just be that we are profoundly ignorant of life … except for what we can know about ourselves. Perhaps it would be better to pay attention sometimes to that wide ignorance rather than any narrow knowledge.

What do you think? Are LeDoux's two crazy ideas really that crazy? What else jumped out at you from his deep history of ourselves?

13 — (Rethinking) The Attention Schema



Graziano with his ventriloquist puppet Kevin. Consciousness studies sure draw renegades.

8 April 2020

In the **last post**, I noted that Dr. Ginger Campbell conducted one-on-one interviews with three prominent neuroscientists during the final episodes of her **Brain Science** podcast series on consciousness. We've already covered the first interview with Joseph LeDoux. Today, I'm going to go over **the second interview** with Michael Graziano about his book ***Rethinking Consciousness: A Scientific Theory of Subjective Experience.*** Graziano is currently a professor of Psychology and Neuroscience at Princeton University where he has had a lab studying consciousness since 2010. Here are the highlights from his interview:

- In 10 years of lab work, I have worked to put my ideas into an evolutionary context (i.e. how they developed), in order to give us an idea of the components that go into this thing we call consciousness.
- More and more, people in the science of consciousness are beginning to coalesce around a coherent set of ideas. My work fits into this growing standard model of consciousness. This core set of scientists realise that we are machines and the brain is an information processing machine that thinks it has magic inside it because it builds somewhat imperfect models of the world inside it. This includes Higher Order Thought Theory, Global Workspace Theory, and even some Illusionists who talk of consciousness as an illusion. My theory is not a rival to these. We are moving past rivalry and towards an integrating picture of it all.
- The realisation is coming that everything you think derives from information. No claims can be put out by the brain without information upon which to base it. This is just basic logic. The question then is how and why did the brain construct a particular piece of information? The brain can construct all sorts of seemingly crazy ideas (e.g. "I have a squirrel in my head instead of a brain.")
- I study movement control, which requires a whole model. If the brain wants to control the arm, it needs a model of the arm. It needs an internal model, a simulation of what an arm is and where it is at any one time. This is an engineering perspective, which is useful for the study of consciousness. Similar to the moving arm, the brain is continually shifting

its focus of attention. So, how do you control that? The same way as the arm. The brain needs a model or simulation of attention, of what it means to focus resources on something.

- This is called "attention schema theory", which follows the "body schema" developed 100 years ago. Phantom limbs are good examples of "body schema". By analogy, there must be a schema for attention—the brain's model for seeing information and processing it deeply.
- Like all complex traits, you can go back very, very far and see this gradual transition where it becomes impossible to draw a line and say "the trait exists after this but not before this." For example, you couldn't draw clear lines in evolution for hands, feet, and flippers. Consciousness is the same.
- I start with attention—a basic ability of a nervous system to focus on a few things at a time and process them deeply. Some forms of this attention go back possibly all the way to the beginnings of nervous systems. Attention is at the root of intelligence. At the heart of intelligence is a very pragmatic problem: you only have so much energy and space for a brain, but you need to use it as efficiently as possible to process deeply and intelligently. How do you do that? Don't occupy the brain with processing all of the million and a half things going on around you. Focus on one or two things at a time. Without that level of attention, any kind of intelligence is impossible.
- Attention comes in very early in evolution, and over time it becomes more and more complex. There's central attention, sensory attention, more cognitive kinds of attention, and they emerge gradually over this sweep of history from about half a billion years ago up to the present. Piggybacking off of this, what people call consciousness also emerged, and also as a gradual process.
- Attention can be separable from consciousness. At what point might it be consciousness?
- Bodies have been involved from the beginning. Schemas only came once nervous systems were capable of building models of these bodies. A body schema stands hierarchically above the body. It isn't the same thing, and they can be dissociated (e.g. phantom limbs). Similarly, this is the relationship between attention and consciousness. Attention is literally what the brain is focusing its resources on. The Attention Schema is what the brain thinks it is focusing its resources on, what the brain thinks focusing is, and what the brain thinks the consequences of focusing are. And those are dissociable too. Typically, they don't. Typically, they track quite well (like the body schema), but you can trick them and get them to peel off from one another.
- Global Workspace Theory is basically a theory about attention. How do you become conscious of an apple you are looking at? GWT says you attend to the signals. They become stronger from your visual system at the expense of other signals. At some point, the signals become so strong that they reach a state called "ignition" when they can then influence wide networks around the brain. Now attention has been reached, you can talk about it, you can move toward it, you can remember it later. The apple information reaches the global workspace and becomes available all around the brain systems. GWT says that is consciousness. The weakness of GWT is that it doesn't explain why we claim to have a subjective experience. It doesn't say why I have an inner experience of the apple.
- The attention schema says great for GWT, but you need one more component—a system in the brain that says "Ah, I am attending to the apple. I have a global workspace that has taken in that apple information." You need something in the brain that can model itself and build some kind of self-description. GWT is the attention. Attention Schema is the consciousness riding on top of that.
- To control something, you need a model of it. But an overly complicated one is wasteful. A "cartoonish" one is good enough.

- Why does it feel non-physical? This is one of the most successful points about the Attention Schema. The brain models itself, but it doesn't need to include little physical details. It doesn't need to know anything about the little implementation details. Therefore, the brain's self-models depict something that has no physical components. It depicts a vague non-physical thing that has a kind of location within us, but that's the only physical property it has. Efficiency dictates the models be as stripped down as possible. This is why introspection, informed by internal models, tells us there is something inside us but it feels like a non-physical essence.
- With this Attention Schema, we don't need another explanation for the philosopher's qualia because there it is. Chalmers, after the Hard Problem, now talks about the Meta Problem. The Hard Problem is how do we get qualia, or that inner subjective feeling. The Meta Problem is why do we think there is a Hard Problem? The Attention Schema solves the Meta Problem. It explains why people think there is this magical non-physical thing inside us. It does an end run around the Hard Problem.
- The ability to attribute consciousness to others is important. In this evolutionary process, we start out evolving an ability to model and keep track of ourselves, which helps make predictions about ourselves and control our behaviour. At some point, as social interactions become more sophisticated, we develop the ability to use the same machinery to model others. This social use probably came in very early in evolution. There is a lot of sophistication in reptiles, birds, and mammals. We not only keep track of and model our own attention, but we keep track of and model others' attention. That allows me to predict your behaviour.
- Ventriloquist dummies are great examples of our souped-up drive to model conscious minds in the world around us.
- We seem to model attention as if it were a fluid flowing out of their eyes, which explains all kinds of folk beliefs about feeling eyes on the back of the neck, telekinesis, the Force in Star Wars, the evil eye, etc., etc.
- Integrated Information Theory is kind of the opposite of this. IIT belongs to theories where you start with an axiomatic assumption. IIT starts with "consciousness exists" stating there is this non-physical feely thing inside us. The magical thing is there, so how does it emerge and under what conditions? So right from the outset there is a divergence. On my end, the starting point is that the brain cannot put out a claim unless there is information for that claim on which it is based. There is no reason to assume this information is accurate. When people feel they have magic, the job of scientists isn't to find out how the brain produces magic; it's to find out why the brain builds that model to describe itself. IIT is a fundamentally magical theory.
- According to IIT, consciousness arises from information and everything in the universe has some information in it. So, you end up with panpsychism that consciousness exists in everything and everywhere. That seems like you've used faulty logic to paint yourself into a corner. If everything is conscious, what does consciousness even mean anymore?

(Not So Brief) Brief Comments

When Graziano opened his interview talking about putting consciousness into an evolutionary context, he had me hooked. When he stated the field was coalescing around a growing standard model of consciousness that brought together Higher Order Thought Theory, Global Workspace Theory, and even some Illusionists, I got excited because those were the theories I most agreed with in the prior posts in this series. When Graziano said this core set of scientists think that we are machines and the brain is an information processing machine that thinks it has magic inside it because it builds somewhat imperfect models of the world inside it, this made a lot of sense. But when Graziano tried to offer his picture to integrate all of this, he finally lost me. To see why, let me go through some of his points one

by one.

>>> *"No claims can be put out by the brain without information upon which to base it."*

This is an excellent place to start. I'll use this later in the series when making connections between the evolution of consciousness and evolutionary epistemology, which charts the way knowledge-gathering has grown incrementally over evolutionary history.

>>> *"If the brain wants to control the arm, it needs a model of the arm. It needs an internal model, a simulation of what an arm is and where it is at any one time. This is an engineering perspective, which is useful for the study of consciousness. Similar to the moving arm, the brain is continually shifting its focus of attention. So, how do you control that? The same way as the arm. The brain needs a model or simulation of attention, of what it means to focus resources on something. ... By analogy, there must be a schema for attention —the brain's model for seeing information and processing it deeply."*

I believe Graziano is making a poor analogy here. When an arm moves, it moves through space and time by contracting muscles that cannot see anything. When a focus of attention shifts, no such physical movement or navigation issues occur. I think it's a mistake to think of models being required to control both of these different things in the same kind of way.

>>> *"Attention is at the root of intelligence. At the heart of intelligence is a very pragmatic problem: you only have so much energy and space for a brain, but you need to use it as efficiently as possible to process deeply and intelligently. How do you do that? Don't occupy the brain with processing all of the million and a half things going on around you. Focus on one or two things at a time. Without that level of attention, any kind of intelligence is impossible."*

This isn't the way evolution works. It doesn't start with information about "a million and a half things" and then pare back from that. Early nervous systems would have begun by sensing just one or a few things, with lots of trial and error going on about which few things. The most successful senses would have been naturally selected for, and then gone on to (blindly) experiment with adding a few new bits of information to sense and process. This evolution never stops, but it only gets as far as it needs to in order to remain alive and reproduce. As **Michael Ruse wrote in *The Oxford Handbook of Philosophy of Biology***, "Consider the much-discussed example of the frog, which snaps at anything suitably small, dark, and moving, regardless of whether it is frog food. A frog cannot discriminate between moving flies and small plastic pellets tossed in front of it no matter how many pass its way."

So, contrary to Graziano's claims, attention is NOT at the root of intelligence. And intelligence IS possible without attention. Intelligence can be very slowly built up by very narrow increments of additional information. Attention — the way that Graziano is using it — is really another word for choice, i.e. choosing which stimuli to "pay attention" to. But such choices do not need control; they can be made non-consciously by simply responding to the loudest signals, where evolutionary trials and errors shape what "loud signals" actually are. Think of the bees flying back from explorations for nectar and doing their wiggle dance to "convince" others to "listen" to them. It's just the most excited dances that "get paid attention to" by the rest of the hive. That doesn't require conscious choice. So, it's not obvious to me that attention is what consciousness is or is required for.

>>> *"A body schema stands hierarchically above the body. It isn't the same thing, and they can be dissociated (e.g. phantom limbs). Similarly, this is the relationship between attention and consciousness. Attention is*

literally what the brain is focusing its resources on. The Attention Schema is what the brain thinks it is focusing its resources on, what the brain thinks focusing is, and what the brain thinks the consequences of focusing are."

I think there is an excellent point here about body schemas and brain schemas both being separate from the actual bodies and brains. I just don't think attention is at the heart of it.

>>> "Global Workspace Theory is basically a theory about attention. How do you become conscious of an apple you are looking at? GWT says you attend to the signals. They become stronger from your visual system at the expense of other signals. At some point, the signals become so strong that they reach a state called "ignition" when they can then influence wide networks around the brain. Now attention has been reached, you can talk about it, you can move toward it, you can remember it later. The apple information reaches the global workspace and becomes available all around the brain systems. GWT says that is consciousness. The weakness of GWT is that it doesn't explain why we claim to have a subjective experience. It doesn't say why I have an inner experience of the apple."

>>> "The attention schema says great for GWT, but you need one more component—a system in the brain that says "Ah, I am attending to the apple. I have a global workspace that has taken in that apple information." You need something in the brain that can model itself and build some kind of self-description. GWT is the attention. Attention Schema is the consciousness riding on top of that."

See. Graziano unwittingly contradicts himself here by describing GWT as the attention without the consciousness. All of the choices of attention can be made (through evolutionarily-learned ignition) without a schema sitting on top of it and controlling it. Again, I think he's right that a schema is needed, but it isn't about attention alone.

>>> To control something, you need a model of it. But an overly complicated one is wasteful. A "cartoonish" one is good enough.

I think this may be a big source of Graziano's errors on this. He is thinking like an engineer who is concerned with top-down "control" rather than thinking like an evolutionary biologist who sees bottom-up emergence. There is no top-down control or design in nature.

>>> "Why does it feel non-physical? This is one of the most successful points about the Attention Schema. The brain models itself, but it doesn't need to include little physical details. It doesn't need to know anything about the little implementation details. Efficiency dictates the models be as stripped down as possible."

This is more thinking like an engineer. Nature doesn't strip down; it builds up. And if more building provides an advantage, then that building up gets selected for. Why wouldn't an Attention Schema ever build up these little physical details? Graziano raises an excellent point, but I think there's a better answer just ahead.

>>> "The ability to attribute consciousness to others is important. In this evolutionary process, we start out evolving an ability to model and keep track of ourselves, which helps make predictions about ourselves and control our behaviour. At some point, as social interactions become more sophisticated, we develop the ability to use the same machinery to model others. This social use probably came in very early in evolution. There is a lot of sophistication in reptiles, birds, and mammals. We not only keep track of and model our own attention, but we keep track of and model others' attention. That allows me to predict your behaviour."

Making models is vital, but I think Graziano has it backwards here. Life wouldn't have started with models of itself; it would have started with models of the outside world, with models of others. As we saw in **my post about Antonio Damasio**, "Valence / value evolved much

earlier. Even bacteria can go toward food and away from danger." What is a model other than a set of if / then rules? What rules would a bacteria have in place about itself before it developed rules for going towards food and away from danger? I can't think of any.

Graziano says that "at some point, as social interactions become more sophisticated, we develop the ability to model others." But long before social interactions mattered, the predator / prey relationship would have dominated the natural selection of minds that could make models of others. And here is a big realisation. Those models … would not have had any physical inputs for them! To say it like a philosopher, I cannot know what it feels like to be a bat, but I may need to know how a bat might attack or elude me, so I will build a model in my head of that bat, even though I have no physical inputs into that model. In more philosophical jargon, the epistemic barrier created by living in a physical world where mental phenomena do not just leap across organisms is exactly the reason why our theories of minds have to feel non-physical.

[I feel like I hit on something big there.]

By the time our model-building of others could turn inwards, these models would have experienced a runaway arms race between predators and prey that shaped them into sophisticated, but non-physical, models. Such sophisticated external models would do just fine for understanding our internal selves, so there would be no need to develop a new model using all of the internal physical processes going on. In fact, there would likely be evolutionary harm to even try because the resources expended on such a project would be wasted with no chance to catch up to the existing model-making skill. (Note: even if the internal models were being built at the same time, the external ones would have faced much stiffer competition and developed more rapidly.)

>>> *"With this Attention Schema, we don't need another explanation for the philosopher's qualia because there it is. Chalmers, after the Hard Problem, now talks about the Meta Problem. The Hard Problem is how do we get qualia, or that inner subjective feeling. The Meta Problem is why do we think there is a Hard Problem? The Attention Schema solves the Meta Problem. It explains why people think there is this magical non-physical thing inside us. It does an end run around the Hard Problem."*

As we saw in **my post about Chalmers**, that's not an accurate description of the Hard and Meta problems. You can't make an "end run" around the Hard Problem. Chalmers doesn't consider the Meta Problem to be beyond it. (He called it another "easy" problem about behaviour.) I think my explanation works better as to why this magical thing inside of us feels non-physical. And it's an impossible question to ever answer all the whys behind the Hard Question.

>>> *"We seem to model attention as if it were a fluid flowing out of their eyes, which explains all kinds of folk beliefs about feeling eyes on the back of the neck, telekinesis, the Force in Star Wars, the evil eye, etc., etc."*

I think Graziano is mixing up the possible uses of attention here. His Attention Schema is about choosing to pay attention to *some* senses rather than others. Modelling the attention of another being is about modelling *everything* that that being can see. We model the fluid as if it were on all the time, not as if it were being paid attention to only occasionally. My idea — let's call it an ExteroSchema for now — may still build its model of vision as a fluid flowing out of others' eyes. That might be the easiest way to do it and it's a cool explanation of that range of folk beliefs.

>>> "IIT is a fundamentally magical theory."

Finally, Graziano finishes with a critique of Integrated Information Theory that sounds pretty dismissive. Our next post will be all about IIT though, so I look forward to diving into it and seeing how it is presented by a strong proponent.

What do you think? Do you agree with me that Graziano has some evolutionary ideas backwards? Does my explanation of modelling others first make more sense? I'd love to hear what you think of this in the comments.

14 — Integrated Information Theory

IIT. Simple summary. Devil in the details.

11 April 2020

We're finally here! The end of my literature review on consciousness. In the **last post**, we heard Michael Graziano lump the work of all of the other neuroscientists I've profiled into one "growing standard model." This is by no means comprehensive for the entire field, so there are still people working outside of this model, but there was one particularly glaring omission that Graziano went out of his way to exclude — Integrated Information Theory (IIT). In the final interview in her four-part series on consciousness, Dr. Ginger Campbell spoke with one of the leading proponents of IIT, **Christof Koch**, about his latest book ***The Feeling of Life Itself: Why Consciousness is Widespread but Can't Be Computed***. There's a lot to consider here so let's get to the highlights:

- My background is in physics and philosophy. I worked with Francis Crick after his Nobel Prize. We looked for "the neural correlates of consciousness," i.e. what are the minimal physical / biophysical neuronal mechanisms that are jointly necessary for any one conscious perception? What is necessary for me to "hear" that voice inside my head? Not necessarily to sense it, or process it, but to have *that* experience.
- We now know it's really the cortex—the outer-most shell of the brain, size and thickness of a pizza, highly convoluted, left and right hemispheres, the most complex and highly organised piece of matter in the known universe—which gives rise to consciousness.
- This study of the neural correlates of consciousness is fantastic. For example, whenever you activate such and such neurons, you see your mom's face or hear her voice. And if you artificially stimulate them, you will also have some vague feeling of these things. There is no doubt that scientists have established this close one-to-one relationship between a particular experience and a particular part of the brain.
- Correlates don't, however, answer why we have this experience. Or how. Or whether something like a bee can be conscious. For mammals it's easy to see the similarity to ourselves. But what about the further away you go? Or what about artificial intelligence? Or how low does it go? Panpsychism has said it is everywhere. Maybe it is a fundamental part of the universe.
- To answer these questions, we need a fundamental theory of consciousness.
- I've been working on this theory with Giulio Tononi, which is called the Integrated Information Theory.

- IIT goes back to Aristotle and Plato. In science, something exists to the extent that it exerts causal power over other things. Gravity exists because it exerts power over mass. Electricity exists because it exerts power over charged particles. I exist because I can push a book around. If there is no causal power over anything in the universe, why postulate they exist
- IIT says fundamentally what consciousness is, is the ability of any physical system to exert causal power over itself. This is an Aristotelian notion of causality. The present state of my brain can determine one of the trillion future states of my brain. One of the trillion past states of my brain can have determined my current state so it has causal power. The more power the past can exert over the present and future, the more conscious the thing that we are talking about is.
- In principle, you can measure this system. The exact causal power, a number we call *phi*, is a measure of how much things exist for themselves, and not for others. My consciousness exists for itself; it doesn't depend on you, it doesn't depend on my parents, it doesn't depend on anybody else but me.
- *Phi* characterises the degree to which a system exists for itself. If it is zero, the system doesn't exist. The bigger the number, the more the system exists for itself and is conscious in this sense. Also the type and quality of this conscious experience (e.g. red feels different from blue) is determined by the extent and the quality of the causal power that the system has upon itself.
- Look for the structure within the brain, or the CPU, that has the maximal causal power, and that is the structure that ultimately constitutes the physical basis of consciousness for that particular creature.
- How does this relate to panpsychism? They share some intuitions, but also differ. One of the great philosophical problems with panpsychism is the superposition problem. I'm conscious. You are conscious. Panpsychism says there should be an uber-consciousness that is you and me. But neither of us have any experience of that. Also, every particle of my body has its own consciousness, and there is the consciousness of me and the microphone, or my wife and whatever, or even me and America. But there isn't anything of what it feels like to be America. This is the big weakness of panpsychism.
- IIT solves the superposition problem by saying only the maximum of this measure of IIT exists. Locally, there is a maximum within my brain or your brain. But the amount of causal interaction between me and you is minute compared to the massive causality within. Therefore, there is you and there is me.
- If we ran wires between two mice or two humans, IIT predicts some things. For example, between my left and right hemispheres there are connections called the corpus callosum. If you cut them, you get split brain syndrome—two conscious entities. If you could do the opposite, you would build an artificial corpus callosum between my brain and your brain. If you added just a few, I would slowly start to see some things that you see, but there would be no confusion as to who is who. As more wires are added, though, IIT says there is a precise point in time when the *phi* across this system will exceed the information within either single brain, and at that point, the individuals will disappear and the new conscious entity will arise.
- What is right about this as opposed to the Global Neuronal Workspace Theory or other approaches? GNWT only claims to talk about those aspects of consciousness that you can actually speak about. This is called "access consciousness." Once information reaches the level of consciousness, all areas of the brain can use it. If it remains non-conscious, only certain parts of the brain use it.
- There is an "adversarial collaboration" just beginning where IIT and GNWT proponents have agreed on a large set of experiments to see which theory is supported by fMRI, EEG, subjective reporting, etc. In principle this will be great, but practically, we will see.

- Where the theories really disagree is the fundamental nature of consciousness. GNWT embodies the dominant zeitgeist (Anglo-Saxon philosophy, scientists, Silicon Valley, sci-fi, etc), which says if you build enough intelligence into a machine, if you add feedback, self-monitoring, speaking, etc, sooner or later you will get to a system that is not only intelligent, but also conscious. Ultimately, consciousness is all about behaviour. It's a descendent of behaviourism saying behaviour is all we can talk about.
- The other view says no, consciousness is not magical, it's a natural property of certain systems, but it's about causal power. To the extent you can build something with causal power, that will be conscious, but you cannot simulate it. E.g. weather simulations don't cause your computer to get wet. The same thing holds for perfect simulations of the human brain. The simulation will say it is conscious, but it will all be a deep behavioural fake. What you have to do is build a computer in the image of a brain with massive overlapping connectivity and inputs. In principle, this could give rise to consciousness.
- Could a single cell or an atom be conscious? In the limit, it may well feel like something to be a bacterium. It doesn't have a psychology, feel fragile, or hungry, etc. But there are already a few billion molecules and a few thousand proteins. We haven't yet modelled this, but yes, most biological systems may feel like something.
- Has any consciousness of my mitochondria been subsumed into my own? Yes. On its own, mitochondria has *phi*, but IIT says that once it is put together with something else, that consciousness dissolves. If your brain is disassembled, for example when you die, there may be a few fleeting moments where each part again feels like something. In each case you have to ask what is the system that maximises the integrated information. Only that system exists for itself, is a subject, and has some experience. The other pieces can be poked and studied, but they aren't conscious.
- The zap and zip technique is being used to look for consciousness in patients who may be locked in or anesthetised irregularly. You zap the brain, like striking a bell, and look at the amount of information that reverberates around the brain. A highly compressed response, one that is "zipped up" so there is almost no information response, is more unconscious (or even dead if there is no response) than one where much response around the brain is noted. This is progress in the mind-body problem. (Note, you don't have to believe in IIT or GNWT to use this.)
- Right now, we don't have strong experimental evidence to think that quantum physics has anything to do with the function of brain systems. Classical physics is enough to model everything so far, but you still have to keep an open mind since we don't understand all causations.

Brief Comments

Although I found this interview to be a good overview, it still left me with a lot of questions about IIT. So, before I make any comments, I want to share a bit more research that I found helpful.

From the **Wikipedia Entry on Integrated Information Theory**:

- If we are ever going to make the link between the subjective experience of consciousness and the physical mechanisms that cause it, IIT assumes the properties of the physical system must be constrained by the properties of the experience.
- Therefore, IIT starts by attempting to identify the essential properties of conscious experience (called "axioms"), and then moves on to the essential properties of the physical systems underneath that consciousness (called "postulates").

- Every axiom should apply to every possible experience. The most recent version of these axioms states that consciousness has: 1) intrinsic existence, 2) composition, 3) information, 4) integration, and 5) exclusion. These are defined below.
- 1) Intrinsic existence — By this, IIT means that consciousness exists. Indeed, IIT claims it is the only fact I can be sure of immediately and absolutely, and this experience exists independently of external observers.
- 2) Composition — Consciousness is structured. Each experience has multiple distinctions, both elementary and higher-order. For example, within one experience I may distinguish a book, a blue color, a blue book, the left side, a blue book on the left, and so on.
- 3) Information — Consciousness is specific. Each experience is the particular way that it is because it is composed of a specific set of possible experiences. The experience differs from a large number of alternative experiences I could have had but am not actually having.
- 4) Integration — Consciousness is unified. Each experience is irreducible and cannot be subdivided. I experience a whole visual scene, not the left side of the visual field independent of the right side (and vice versa). Seeing a blue book is not reducible to seeing a book without the colour blue, or the colour blue without the book.
- 5) Exclusion — Consciousness is definite. Each experience is what it is, neither less nor more, and it flows at the speed it flows, neither faster nor slower. For example, the experience I am having is of seeing a body on a bed in a bedroom, a bookcase with books, one of which is a blue book. I am not having an experience with less content (say, one lacking colour), or with more content (say, with the addition of feeling blood pressure).
- These axioms describe regularities in conscious experience, and IIT seeks to explain these regularities. What could account for the fact that every experience exists, is structured, is differentiated, is unified, and is definite? IIT argues that the existence of an underlying causal system with these same properties offers the most parsimonious explanation. The properties required of a conscious physical substrate are called the "postulates" because the existence of the physical substrate is itself only postulated. (Remember, IIT maintains that the only thing one can be sure of is the existence of one's own consciousness).

From two articles (**1**,**2**) about the "adversarial collaboration" between IIT and Global Workspace Theory (GWT):

- Both sides agree to make the fight as fair as possible: they'll collaborate on the task design, pre-register their predictions on public ledgers, and if the data supports only one idea, the other acknowledges defeat.
- Rather than unearthing how the brain brings outside stimuli into attention, the fight focuses more on where and why consciousness emerges.
- The GWT describes an almost algorithmic view. Conscious behavior arises when we can integrate and segregate information from multiple input sources and combine it into a piece of data in a global workspace within the brain. According to Dehaene, brain imaging studies in humans suggest that the main "node" exists at the front of the brain, or the prefrontal cortex, which acts like a central processing unit in a computer.
- IIT, in contrast, takes a more globalist view where consciousness arises from the measurable, intrinsic interconnectedness of brain networks. Under the right architecture and connective features, consciousness emerges. IIT believes this emergent process happens at the back of the brain where neurons connect in a grid-like structure that hypothetically should be able to support this capacity.
- Koch notes, "People who have had a large fraction of the frontal lobe removed (as it used to happen in neurosurgical treatments of epilepsy) can seem remarkably normal." Tononi

added, “I’m willing to bet that, by and large, the back is wired in the right way to have high Φ, and much of the front is not. We can compare the locations of brain activity in people who are conscious or have been rendered unconscious by anesthesia. If such tests were able to show that the back of the brain indeed had high Φ but was not associated with consciousness, then IIT would be very much in trouble.”

- Another prediction of GWT is that a characteristic electrical signal in the brain, arising about 300-400 milliseconds after a stimulus, should correspond to the “broadcasting” of the information that makes us consciously aware of it. Thereafter the signal quickly subsides. In IIT, the neural correlate of a conscious experience is instead predicted to persist continuously while the experience does. Tests of this distinction, Koch says, could involve volunteers looking at some stimulus like a scene on a screen for several seconds and seeing whether the neural correlate of the experience persists as long as it remains in the consciousness.
- It may also turn out that no scientific experiment can be the sole and final arbiter of a question like this one. Even if only neuroscientists adjudicated the question, the debate would be philosophical. When interpretation gets this tricky, it makes sense to open the conversation to philosophers.

Great! So let's get on with some philosophising.

Right off the bat, the first axiom of IIT is problematic. It is trying to build upon the same bedrock that Descartes did. But that is an **infamously circular** argument that rested on first establishing that we are created by an all-perfect God rather than an evil demon. Descartes said this God wouldn't let him be deceived about seeing things “clearly and directly,” which led to his claim that therefore, I am. Now, the first axiom of IIT claims consciousness is the only fact one can be sure of “immediately and absolutely.” This is the same argument, and it still doesn't hold up. The study of illusions and drug-altered states of experience shows us that consciousness is not perceived immediately and absolutely. And as Keith Frankish pointed out in **my post about illusionism**, once that wedge of doubt is opened up, it cannot be closed.

Regardless, let's grant that the subjective experience each of us thinks we are perceiving does actually constitute a worthwhile data point. (Even if this isn't a certain truth, it's a pretty excellent hypothesis.) Talking to one another about all of our individual data points is how IIT comes up with its five axioms. But would it follow from that that ALL conscious experiences have the same five characteristics? No! That would be an enormous leap of induction from a specific set of human examples to a much wider universal rule.

However, despite the universal pretensions of IIT and its definition of *phi* that could theoretically (**though not currently**) be calculated for any physical system, when Koch is talking about consciousness, he occasionally is only referring to the very restricted human version of it that requires awareness and self-report. This makes him confusing at times, but that's certainly the consciousness he's talking about for the upcoming “adversarial collaboration” that will test predictions about consciousness by proponents of IIT and GWT. It's great to see such falsifiable predictions being made and tested, and of course the human report of consciousness is where we have to start our scientific studies of consciousness, but it's hard to see how these tests will actually end the debate any time soon. Why? Because as we have seen throughout this series, we just don't have a settled definition for the terms being used in this debate. One camp's proof of consciousness is another camp's proof of something else. They could all seemingly just respond to one another, “but that's not *really* consciousness.”

So, what does IIT say consciousness *really* is? Koch reports:

>>> *"IIT says fundamentally what consciousness is, is the ability of any physical system to exert causal power over itself."*

I've heard Dan Dennett say that vigorous debates occur about whether tornadoes fit this kind of definition about consciousness. Their prior states influence their current and future states. That's a kind of causal power. They are also a physical system that acts as one thing even though none of the constituent parts act the way the system as a whole does. But does anyone really think a tornado is conscious? Koch continues:

>>> *"My consciousness exists for itself; it doesn't depend on you, it doesn't depend on my parents, it doesn't depend on anybody else but me."*

This isn't strictly true, of course. Everything is interrelated. We have no evidence of any **uncaused causes** in this universe, so Koch's consciousness clearly depends on lots of outside factors. If I shouted that at him, would his consciousness be able to stop him from hearing it? I imagine that's not exactly what Koch meant, but between this and the similarity to Descartes' argument using God to see the world clearly and directly, IIT strikes me as practically a religious viewpoint. Tellingly enough, I found out that it is.

In an essay at Psychology Today titled, "**Neuroscience's New Consciousness Theory Is Spiritual**", there was this passage:

- Most rational thinkers will agree that the idea of a personal god who gets angry when we masturbate and routinely disrupts the laws of physics upon prayer is utterly ridiculous. Integrated Information Theory doesn't give credence to anything of the sort. It simply reveals an underlying harmony in nature, and a sweeping mental presence that isn't confined to biological systems. IIT's inevitable logical conclusions and philosophical implications are both elegant and precise. What it yields is a new kind of scientific spirituality that paints a picture of a soulful existence that even the most diehard materialist or devout atheist can unashamedly get behind.

I'll let the "inevitability" of IIT's logical conclusions slide for now, but is this "sweeping mental presence" just another form of **idealism**, which **George Berkeley** used to argue that the mind of God was everywhere and caused all things? It's not from the same source or for exactly the same reason, but it's related. As an essay at the Buddhist magazine *Lion's Roar* points out, "**Leading neuroscientists and Buddhists agree: 'Consciousness is everywhere'**." Here we find that:

- Buddhism associates mind with sentience. The late Traleg Kyabgon Rinpoche stated that while mind, along with all objects, is empty, unlike most objects, it is also luminous. In a similar vein, IIT says consciousness is an intrinsic quality of everything yet only appears significantly in certain conditions — like how everything has mass, but only large objects have noticeable gravity."
- In his major work, the *Shobogenzo*, Dogen, the founder of Soto Zen Buddhism, went so far as to say, "All is sentient being." Grass, trees, land, sun, moon, and stars are all mind, wrote Dogen.
- Koch, who became interested in Buddhism in college, says that his personal worldview has come to overlap with the Buddhist teachings on non-self, impermanence, atheism,

and panpsychism. His interest in Buddhism, he says, represents a significant shift from his Roman Catholic upbringing. When he started studying consciousness — working with Nobel Prize winner Francis Crick — Koch believed that the only explanation for experience would have to invoke God. But, instead of affirming religion, Koch and Crick together established consciousness as a respected branch of neuroscience and invited Buddhist teachers into the discussion.

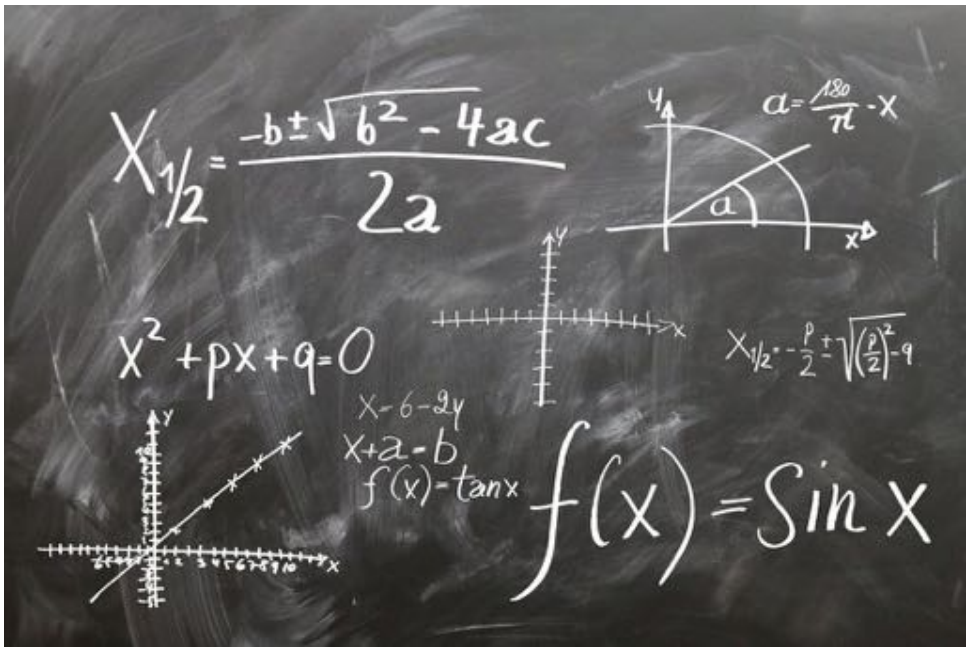
- At Drepung Monastery, the Dalai Lama told Koch that the Buddha taught that sentience is everywhere at varying levels, and that humans should have compassion for all sentient beings. Until that point, Koch hadn't appreciated the weight of his philosophy. "I was confronted with the Buddhist teaching that sentience is probably everywhere at varying levels, and that inspired me to take the consequences of this theory seriously," says Koch. "When I see insects in my home, I don't kill them."

These religious motivations don't necessarily mean that the motivated reasoning behind IIT is unsound. But it sure makes me skeptical. The cracks I see in IIT's logic—e.g. starting with seeing consciousness immediately and absolutely, making leaps from human experience to all experience, seeing islands of uncaused causes everywhere—are enough to give me pause. Despite all the fancy math plastered on top of these ideas, I'm still fundamentally unconvinced that consciousness is the integration of information, yet somehow "**can't be computed and is the feeling of being alive**." As for what I think consciousness really is, it's finally time for me to say. Hope I can get it down clearly!

What do you think? Is IIT flawed to you too? What useful concepts or calculations might it offer?

AN EVOLUTIONARY THEORY OF CONSCIOUSNESS

15 — What is a Theory?



16 April 2020

In the **last post**, I finished my series of reviews about what I consider to be the best theories and data about consciousness that are currently available from philosophers and scientists. I was planning to start laying out my own thoughts about this subject in today's post, but as luck would have it, I happened to come across an amazing lecture last night that I thought would be helpful as a transition and setup before I continue.

A few days into this coronavirus lockdown, I stumbled across an app called **Kanopy** that lets you log into it using your local library account, and then watch stuff online that you could normally check out of your library. All for free! It's such a great idea. As it happens, my wife's university library account also gave us free access to **The Great Courses**, which is a real treasure trove of university-level lectures. For reasons I don't need to go into now, I started watching a class called **An Introduction to Formal Logic** by Professor Steven Gimbel of Gettysburg College. Last night, I made it through lesson 7 on inductive reasoning. (Quick recap: deductive reasoning narrows down from a big rule to small facts, while inductive reasoning grows out from small observances to general rules. Of course, the **problem of induction** is well known as "the glory of science and the scandal of philosophy.")

Towards the end of this lecture, Gimbel went over the difference between using inductive reasoning for a theory versus using it for a hypothesis. This ended up being one of the best passages I've seen for explaining why Darwin's great idea is called the theory of evolution rather than the fact of evolution. This will also come in handy for anyone who wants to put together a theory of consciousness. Enjoy.

Take Newton's theory of gravity, which is comprised of three laws: 1) the law of inertia; 2) the force law; and 3) the action-reaction law. Put them together, and you have a full theory of motion. But what we have here are three general propositions, not specific observable claims. These general laws are then combined to form a system from which we can derive specific

cases by plugging in the conditions of the world.

These proposed laws of nature, which function as the axioms of the theory, should not be confused with hypotheses. Hypotheses are proposed individual statements of possible truth. They are more specific than the axioms, and we get evidence for them individually. The axioms work together as a group. We may be able to derive hypotheses when working within the theory, but the parts of the theory themselves are not hypotheses.

For example, a hypothesis would be, "If I drop a 10-pound bowling ball and a 16-pound bowling ball off the roof of my house, they will land at the same time." I could test this with a ladder and two bowling balls. Hypotheses are open to such direct testing. The purported laws of nature in Newton's theory, however, are different. Consider Newton's First Law. If I have an object, and there's no external force applied to it, then it will move in a straight line at a constant speed. At first glance, this seems like it should be just as testable as the hypothesis about the bowling ball. But the problem is that there can be no such object without an external force applied to it! As soon as there's any other object in the universe, the object we're examining would feel the pull of gravity, which is an external force. So, Newton's law of inertia, a vital part of his theory of motion, holds for no actual object. If we treat it like we do hypotheses, it would be kind of like having a biological law about unicorns. So, we have to have different inductive processes for hypotheses and for theories.

[Karl Popper gave us the idea that hypotheses must be falsifiable. Hypotheses are tested using independent and dependent variables, i.e. the things we adjust and the things we measure.]

What about theories? Here, the philosopher **Hans Reichenbach** drew a distinction between discovery and justification. What this distinction has come to mean is that there is a difference between the context in which scientists come up with their theories, and the context in which they provide good reasons to believe those theories are true. The context of discovery is genuinely thought to be free. There's no specific logic of discovery, no turn-the-crank method for coming up with scientific theories. The great revolutionaries are considered geniuses because they were able to not only think rigorously, but also creatively in envisioning a different way the world could work. There's no logic that tells scientists what to consider when coming up with new theories.

While there's no set method, surely there is induction in there somewhere. Scientists are working from their experiences and their data. They have a question about how a system works, they consider what they know, and they make inductive leaps. They look for models and analogies where the system could be thought to work like a different system that is better understood. So, while there's no set means of using induction in the context of discovery, it usually is playing some kind of role.

The most important place in scientific reasoning that we find induction is in the context of justification. Once a theory has been proposed, why should we believe it? Theories are testable. They have effects, results, and predictions that come from them. These observable results of a theory are determined deductively. That is, if a theory is true, then, in some given situation, let's say that observable consequence O should result. We go to the lab, set up the situation, and see if we observe O as expected. If not, then the theory has failed, and, as it stands, it is not acceptable. It will either have to be rejected or fixed. But, if the theory says to expect O, and we actually do observe O, now we have evidence in favour of the theory. That evidence is inductive. It may be that theory T1 predicts O, but there will also be other theories, like T2, which is different from T1, which is also supported by O. As such, neither

T1 nor T2 are certain. (To the degree that inductive inferences could be anyway.)

How then do we go from supporting evidence (which makes a theory more likely), to conclusive evidence (which makes a theory probably true)? We need lots of evidence. We also need evidence of different types. It's good for a theory if it can account for everything we already know. We call this *retrodiction*. This is particularly true if everything we knew was previously unexplained. For example, before Einstein's theory of general relativity, we knew that not only did Mercury orbit the Sun, but each time Mercury would make it around the Sun, the farthest point in its orbit would be in a different place. In other words, Mercury did not make the same exact trip around the Sun every time. But we had no idea why! Once Einstein gave us a new theory of gravitation, this effect was naturally explained. The fact that it solved the mystery was taken as strong inductive evidence.

Even better than explaining what we already know, *prediction* is also taken as strong evidence. Newton's theory predicted that a comet would appear around Christmastime in 1758. When this unusual sight appeared in the sky on Christmas day, the comet (named for Newton's close friend Edmund Halley) was taken as very strong evidence for his theory.

Beyond even prediction, the best evidence for a theory can bring forth what **William Whewell** termed *consilience*. Whewell was a philosopher of science, an historian of science, and also a scientist. In fact, he was the person who coined the term scientist. Consilience is when a theory that is designed to account for phenomena of type A, turns out to also account for phenomena of type B. If you set out to explain one thing, and are also able to explain something completely different, then that is extremely strong evidence that your theory is probably true.

The reigning champ in this realm is Darwin's theory of evolution. It accounts for biodiversity. It accounts for fossil evidence. It accounts for geographical population distribution. There's just a huge range of all sorts of observations that evolution makes sense of. This is stunning, and stands as extremely strong evidence for its likely truth.

This consilience is no accident. In his college days, Darwin was a student of Whewell's. When he later began to develop his ideas, Darwin was extremely nervous about them. He knew how explosive his view was, so he spent many, many years accumulating a broad array of different sources of evidence in order to demonstrate his theory's consilience. Some people today contend that evolution is not proven. Well of course it isn't! The only things that are proven are the results of deductive logic. Darwin's theory—like everything else in science—is confirmed by inductive logic, which never gives proof, but which offers high probability, and thereby firm grounds, for rational belief.

What do you think? Does this understanding of a theory help you see how science can actually posit ideas that cannot be tested on their own, yet still help us make sense of the world? Are we ready for a theory of consciousness that uses analogies from things we understand to explain everything we know, make some predictions, and offer a consilient view of a wide variety of observations? And might it fit in with the theory of evolution too? Maybe not 100% ready, but I'm going to sketch out a new theory next time and give this all a go.

16 — A (Sorta) Brief History of the Definitions of Consciousness

25 May 2020

(Quick Note: Sorry for the long delay on this. I am fine; no coronavirus infections yet. I posted the first 15 parts of this series roughly every other day, but it's taken a little over a month now for this one. Basically, I've been doing loads of research. I had a sketch in mind for my personal thoughts about consciousness but fleshing out the details took a lot longer than I expected. I put together 37 pages of research for the first 15 posts in this series, but I had to gather another 80 pages (!) for these final posts, of which I expect there to be 8. Don't let that put you off, though. The first 15 posts were basically transcriptions of that research, but these final ones will be highly summarised. Anyway, back to the series!)

In the **last post**, I went over what a scientific theory is and is not. I asked if we were ready for just such a scientific theory of consciousness, one that uses analogies from things we understand to explain everything we know, makes some predictions, and offers a consilient view of a wide variety of observations. Before fleshing out my own take on this, I knew I ought to take a little more care in reviewing the history of how other people have grappled with this big, tangled concept. Whole books have already been written about this, and I don't intend to duplicate the details there, but a useful sketch can be drawn from the following sources that I found particularly helpful:

- the wikipedia entry on **Consciousness**
- the Stanford Encyclopedia of Philosophy entry on **Consciousness**
- a 2012 paper by the British philosopher Peter Hacker titled "**The Sad and Sorry History of Consciousness: Being, Among Other Things, A Challenge to the 'Consciousness-Studies Community'**"
- three papers from Dan Dennett: "**The Unimagined Preposterousness of Zombies**" (1995); "**Who's on First? Heterophenomenology Explained**" (2003); and "**Darwin and the Overdue Demise of Essentialism**" (2016)

I'll use these sources and some details from the previous posts in this series to slot ideas about

consciousness into three categories: philosophical, scientific, and dictionary.

Philosophical Considerations of Consciousness

- Descartes introduced the term 'conscious' into philosophy in 1640, although it was only in passing as part of his writing about thoughts. Descartes defined the term 'thought' (*pensée*) as "all that we are conscious as operating in us." This included everything passing in our minds—thinking, sensing, understanding, wanting, and imagining. He held these things to be private, infallible, and beyond doubt, leading to his famous "I think therefore I am" argument (which **is deeply flawed**). Descartes was also a 'substance dualist' who asserted the existence of both physical and non-physical substances as components of nature. Such Cartesian dualism has largely been dropped from philosophy now.
- Fifty years later in 1690, John Locke is credited with the first modern concept of 'consciousness' which he defined as "the perception of what passes in a Man's own Mind."
- In 1714, Leibniz made the first distinction between 'perception' ("the representation of that which is outside") and 'apperception' ("consciousness, or the reflective knowledge of this internal state"). Leibniz also famously argued that a mechanical explanation of consciousness would be impossible for it would be like going into a windmill and claiming the moving parts explained the phenomenon.
- In the 1780's, Kant took these ideas to their "baroque culmination" by developing a rich structure of mental organisation. Kant called the components of this structure fundamental 'intuitions', which include 'object', 'shape', 'quality', 'space', and 'time'. Kant's category of 'quality' (aka *qualia*, e.g. redness, pain, etc.) has proven particularly difficult for philosophers to explain in physical terms. Some claim these 'raw feels' are ineffable and incapable of being reduced to component processes. There are psychologists and neuroscientists who reject this, however.
- "It was not until the middle of the nineteenth century that 'consciousness' came to be used to signify wakefulness as opposed to being unconscious. Thenceforth one could speak of losing and regaining consciousness." (Hacker 2012)
- Phenomenology arose in the early 20th century in the works of Husserl, Heidegger, Sartre, and Merleau-Ponty. These phenomenologists studied the structures of consciousness as experienced from the first-person point of view. The experiences they considered ranged from perception, thought, memory, imagination, emotion, desire, and volition, to bodily awareness, embodied action, and social activity, including linguistic activity. This typically involved what Husserl called 'intentionality'—the directedness of experience toward things in the world.
- In 1933 (*The Physical Dimensions of Consciousness*), the psychologist E. G. Boring originated the idea of 'type-identity' physicalism, aka the 'identity theory of mind'. Boring wrote, "To the author, a perfect correlation is identity. Two events that always occur together at the same time in the same place, without any temporal or spatial differentiation at all, are not two events but the same event." Several versions of this developed over the following decades but all share the central idea that the mind is identical to something physical.
- In 1949 (*The Concept of Mind*), Gilbert Ryle argued that traditional beliefs about consciousness were based on Cartesian dualism, which improperly separated minds from bodies. Ryle proposed we instead ought to talk about individuals acting in the world, and thus, 'consciousness' was not something separate from behaviour. (This paralleled B.F. Skinner's behaviourism in psychology in the 1930's.) As part of these arguments, Ryle coined the terms 'ghost in the machine' as well as 'category mistake'. He provided robust distinctions between 'knowing-how' and 'knowing-that', as well as between 'thin' and 'thick' descriptions (i.e., observations only and providing context for them). Ryle also identified 'topic-neutral terms' such as 'if', 'or', 'not', 'because', and 'and'. Ryle said his

philosophical arguments "are intended not to increase what we know about minds but to rectify the logical geography of the knowledge we already possess." Former Ryle student Daniel Dennett has said that recent trends in psychology such as embodied cognition and discursive psychology have provoked a renewed interest in Ryle's work.

- Two major schools in the philosophy of mind developed in the post-war years — representationalism and functionalism.
- Direct representationalism (aka naïve realism) argues that we perceive the world directly. Indirect realism/representationalism states that we do not and cannot perceive the external world as it really is; we can only know our ideas and our interpretations of the way the world is. This is roughly the accepted view of perception in the natural sciences.
- Functionalism was first put forth by Hilary Putnam in the 1960s. This theory of mind states that mental states (beliefs, desires, being in pain, etc.) are constituted solely by their functional role. It developed largely as an alternative to the identity theory of mind and behaviourism. An important part of some arguments for functionalism is the idea of 'multiple realisability', which asserts that mental states can be realised in multiple kinds of systems, not just brains.
- The term 'folk psychology' is used to characterise the human capacity to explain and predict the behaviour and mental state of other people. This has primarily focused on intentional states described in terms of everyday language rather than technical jargon, and includes concepts such as 'beliefs', 'desires', 'fear', and 'hope'.
- Eliminative materialism is the claim that folk psychology is false and should be discarded (or eliminated). It is a materialist position in the philosophy of mind. Some supporters of eliminativism argue that no coherent neural basis will be found for many everyday psychological concepts such as belief or desire, since they are poorly defined. The main roots of eliminative materialism can be found in the writings of mid-20th century philosophers Wilfred Sellars, W.V.O. Quine, Paul Feyerabend, and Richard Rorty.
- In 1962 ("Philosophy and the Scientific Image of Man"), Wilfrid Sellars coined a distinction between the 'manifest image' and the 'scientific image' of the world. The manifest image includes intentions, thoughts, and appearances. The scientific image describes the world in terms of the theoretical physical sciences such as causality, particles, and forces. Sellars is also known for describing the task of philosophy as explaining how things, in the broadest sense of term, 'hang together'.
- In 1974 ("What is it like to be a bat?"), Thomas Nagel published the paper that Dan Dennett called "the most widely cited and influential thought experiment about consciousness." In it, Nagel defended three theses: 1) An experience is a conscious experience if and only if there is something it is like for the subject of the experience to have that very experience. 2) A creature is conscious or has conscious experience if and only if there is something it is like for the creature to be the creature it is. 3) The subjective character of the mental can be apprehended only from the point of view of the subject. Nagel used these theses to argue that "materialist theories of mind omit the essential component of consciousness." (In **my response to this thought experiment**, I argued that it is actually entirely consistent with a materialist/physicalist worldview.)
- In 1980 ("Minds, Brains and Programs"), John Searle first published his Chinese Room thought experiment in which a man who does not understand Chinese, stays inside a room, takes in requests written in Chinese characters, consults a complete book for responses, and simply returns whatever characters the book tells him to. This experiment challenged the functionalist view that it is possible for a computer running a program to have a 'mind' and 'consciousness' in the same sense that people do, since this man would have no understanding of the Chinese function being performed. This was part of Searle's 'biological naturalism' which states that consciousness requires the specific biological machinery that is found in brains. Searle argues that this machinery (known to

neuroscience as the 'neural correlates of consciousness') must have some as yet unspecified 'causal powers' that give us our experience of consciousness. (In **my response to this thought experiment**, I noted that Searle's dismissal of the notion that the Chinese Room 'system' gains consciousness chimes with what theoretical evolutionary biologists John Maynard Smith and Eros Szathmary said in *The Origins of Life* in their analysis of ecosystems (emphasis added): "There is a massive amount of information in the system, but it is information specific to individuals. *There is no additional information concerned with regulating the system as a whole.* It is therefore misleading to think of an ecosystem as a super-organism." However, I also went through a list of behaviours that might give an AI system enough of the appearance of consciousness to get us to pragmatically treat it as if it did. Once computers become unique individuals that have changed their goals and understanding due to irreplaceable, learned experiences, then they will similarly attain the infinite value that any life has.)

- In 1982 ("Epiphenomenal Qualia") and 1986 ("What Mary Didn't Know"), Frank Jackson published and then clarified his 'knowledge argument' about a neuroscientist named Mary who learns "all there is to know" about the colour red while being confined to a black and white existence. Her discovery of 'something new' when she sees red for the first time is intended to show that consciousness must contain non-physical elements since she already supposedly knew every physical fact about red. (In **my response to this thought experiment**, I noted that a physical universe would preclude Mary from having every fact about red because mental imaginings are not enough to move the physical atoms in the nerves of our eyes and brain synapses.)
- In 1991 (*Consciousness Explained*), Dan Dennett put forward his 'multiple drafts model' of consciousness, claiming there is no single central place (a 'Cartesian theatre') where conscious experience occurs. Dennett's view of consciousness is that it is the apparently serial account of the brain's underlying parallelism. Dennett says that only a theory that explained conscious events in terms of unconscious events could explain consciousness at all. He says, "To explain is to explain away."
- Robert Kirk first introduced the idea of philosophical zombies—unconscious beings who are physically and behaviourally identical to human beings—in 1974 ("Zombies v. Materialists"). However, this idea gained much more traction in the mid-1990's with the publications of essays by Todd Moody ("Conversations with Zombies" 1994), Owen Flanagan and Thomas Polger ("Zombies and the Function of Consciousness" 1995), Dan Dennett ("The Unimagined Preposterousness of Zombies" 1995), and David Chalmers (*The Conscious Mind* 1996). If philosophical zombies existed, this would show that consciousness has non-physical properties. Robert Kirk eventually reversed his earlier position about zombies, but in 2019 wrote a Stanford Encyclopedia of Philosophy entry on zombies that ended by saying, "In spite of the fact that the arguments on both sides have become increasingly sophisticated—or perhaps because of it—they have not become more persuasive. The pull in each direction remains strong." (In **my response to this thought experiment**, I noted that the argument takes our *uncertainty* about the existence of zombies and uses that to claim *certainty* that physicalism is false. That's a logical error. We just don't know yet and speculations about the possibility of zombies or zoombies (beings who are *non-physically* the same as zombies but are conscious) can actually be used to argue for or against physicalism in either direction.)
- In 1995 ("Facing Up to the Problem of Consciousness"), David Chalmers introduced the 'hard problem' of consciousness to ask why some internal states are subjective, felt states, rather than non-subjective, unfelt states, as in a thermostat or a toaster. Chalmers contrasted this with the 'easy problems' of explaining the neural basis for abilities to discriminate, integrate information, report mental states, focus attention, and so forth. Easy problems are (relatively) easy because "all that is required for their solution is to

specify a mechanism that can perform the function." The existence of the hard problem is controversial, with many philosophers and neuroscientists on both sides of the argument. (In **an earlier post in this series**, I said it is only hard because it can keep retreating to an impossible problem.)

- In 1996 (*Consciousness and Experience*), William Lycan argued that at least eight clearly distinct types of consciousness can be identified: 1) organism consciousness; 2) control consciousness; 3) consciousness of; 4) state/event consciousness; 5) reportability; 6) introspective consciousness; 7) subjective consciousness; and 8) self-consciousness.
- In 1998 ("On a Confusion About a Function of Consciousness"), Ned Block wrote that consciousness "is a mongrel concept: there are a number of very different 'consciousnesses'." In particular, Block proposed a distinction between two types of consciousness that he called phenomenal (P-consciousness) and access (A-consciousness). P-consciousness is simply raw experience: it is moving, coloured forms, sounds, sensations, emotions, and feelings with our bodies. These experiences can be called qualia. A-consciousness, on the other hand, is when information in our minds is accessible for verbal report, reasoning, and the control of behaviour. Information *about* what we perceive is access conscious; information *about* our thoughts is access conscious; information *about* the past is access conscious, and so on. Some philosophers, such as Daniel Dennett, have disputed the validity of this distinction. David Chalmers has argued that A-consciousness can in principle be understood in mechanistic terms but understanding P-consciousness is the hard problem.
- In 2003 ("Who's on First? Heterophenomenology Explained"), Dan Dennett further elucidated the methodology used for studying consciousness, which he calls 'heterophenomenology' (the phenomenology *of another*, not oneself). Dennett says this is a straightforward extension of objective science that covers *all* the realms of human consciousness without having to abandon the experimental methods that have worked so well in the rest of science. Heterophenomenology is a way to take the first-person point of view as seriously as it can be taken. Social sciences are almost entirely conducted in this way already, so the methods are well understood. Consider two possible sources of data: (a) 'conscious experiences themselves' and (b) beliefs about these experiences. If you have conscious experiences you *don't* believe you have, then those extra conscious experiences are just as inaccessible *to you* as to external observers. On the other hand, if you believe you have conscious experiences that you *don't in fact* have, then it is your beliefs that we need to explain, not the non-existent experiences! Either way, this demonstrates the need to collect the data of (b), and those beliefs can be shared and studied objectively. In contrast, 'lone-wolf autophenomenology', in which the subject and experimenter are one and the same person, is a foul because it isn't science until you turn your self-administered pilot studies into heterophenomenological experiments. Whatever insights one may garner from first-person investigations fall happily into place in third-person heterophenomenology. Heterophenomenology is, therefore, the beginning of a science of consciousness, not the end. And nobody has yet pointed to any variety of data that are inaccessible to heterophenomenology.
- Other philosophical explorations of consciousness talk of components such as:
  - Four main pieces: 1) knowledge in general; 2) intentionality; 3) introspection; and 4) phenomenal experience.
  - Streams of thought, as in the experience of thinking 'in words' or 'in images'.
  - Creature consciousness—an animal, person, or other cognitive system may be conscious in a number of ways: sentience, wakefulness, self-consciousness, what it is like, subject of conscious states, or transitive consciousness (being conscious of).
  - State consciousness—there are six major options for distinct, though perhaps interrelated, types of this: 1) states one is aware of (meta-mentality); 2) qualitative

states (raw sensory feels, qualia); 3) phenomenal states (not only sensory ideas and qualities but complex representations of time, space, cause, body, self, world, and the organized structure of lived reality); 4) what-it-is-like states (similar to 2 and 3, but coming from Nagel); 5) access consciousness (info generally available for use); and 6) narrative consciousness (serial episodes of a self).

- Current schools of philosophy about consciousness largely fall into two main camps: property dualism and physicalism.
- Property dualists assert the existence of conscious properties that are neither identical with nor reducible to physical properties, but which may nonetheless be made up of the same stuff as physical things. There are: 1) fundamental property dualists (consciousness is a basic part of the universe, much like fundamental physical properties such as electromagnetism); 2) emergent property dualists (consciousness arises in a radically new way from physical stuff, but only once it reaches a certain complexity); 3) neutral monist property dualists (physical and mental properties are both derived from something even more basic in reality); and 4) panpsychists (all parts of reality have both physical and mental properties).
- Physicalists assert that reality is only composed of physical objects and the fundamental forces acting upon them. There are: 1) eliminativists (the existence or distinction for some or all features of consciousness are denied in either modest or radical ways); 2) identity theorists (conscious properties just *are* physical processes, usually neurophysiological processes, and so no further causes or explanations are necessary). Most physicalists acknowledge the reality of consciousness but say that it *supervenes* on the physical, is *composed of* the physical, or is *realised by* the physical.
- In January 2020, when asked if he had a simple definition of consciousness, Dan Dennett said, "No. But that's okay. That's the way science works too. There's no perfect definition of time or energy, but scientists get on with it."

That's obviously not everything written by philosophers about consciousness, but it's a pretty good summary of the modern timeline. In my previous posts in this series, I already covered how some prominent scientists do "get on with" consciousness research, but let's look at some of the main definitions used there.

Scientific Considerations of Consciousness

- In 1890 (*The Principles of Psychology*), William James wrote that introspection "means, of course, the looking into one's own mind and reporting there what we discover" and the use of this inner sense is the way we become conscious. He said this inner sense is just like an outer sense, only: 1) without a sense organ; 2) its successful exercise is independent of observation conditions; 3) it never fails us, but always yields knowledge; and so therefore 4) we know the mind better than the material world. While some philosophers still seem beholden to such a Cartesian view of infallibility and indubitability, all four of these characteristics of consciousness have been shown to be faulty. James also considered the ways the unity of consciousness might be explained by known physics and found no satisfactory answer. He coined the term 'combination problem', in the context of a 'mind-dust theory' in which a full human conscious experience is proposed to be built up from proto- or micro-experiences in the same way that matter is built up from atoms. James claimed that such a theory was incoherent, since no causal physical account could be given of how distributed proto-experiences would 'combine'. Today, some prominent philosophers and neuroscientists (e.g., Dan Dennett and Bernard Baars) disagree that this combination problem even exists, claiming consciousness is not unified in the way James described it. Evidence from recall experiments and change blindness support this.

- It was not known that neurons are the basic units of the brain until approximately 1900 (Santiago Ramón y Cajal). The concept of chemical transmission in the brain was not known until around 1930 (Henry Hallett Dale and Otto Loewi). In the 1950s, we began to understand the basic electrical phenomenon that neurons use to communicate—the action potential (Alan Lloyd Hodgkin, Andrew Huxley and John Eccles). We became aware of how neuronal networks code stimuli in the 1960s, which showed how the formation of concepts is possible (David H. Hubel and Torsten Wiesel). The molecular revolution swept through US universities in the 1980s. And it was only in the 1990s that molecular mechanisms of behavioural phenomena became widely known (Eric Richard Kandel).
- Starting in the 1980s, an expanding community of neuroscientists and psychologists have associated themselves with a field called 'Consciousness Studies'. This created a stream of experimental work, which was published in books and journals such as *Consciousness and Cognition*, *Frontiers in Consciousness Research*, *Psyche*, and the *Journal of Consciousness Studies*. Regular conferences were also organised by groups such as the Association for the Scientific Study of Consciousness, and the Society for Consciousness Studies.
- Seven types of specific detailed theories have emerged from Consciousness Studies about the nature of consciousness. This is not comprehensive, but it helps to indicate the main range of options. They are: 1) higher-order theories, 2) representational theories, 3) interpretative narrative theories, 4) cognitive theories, 5) neural theories, 6) quantum theories, and 7) nonphysical theories. These are described below.
- 1. Higher-order (HO) theories analyse the notion of a conscious mental state in terms of reflexive meta-mental self-awareness. Unconscious mental states are unconscious precisely because we lack higher-order states about them.
- 2. Representational theories attempt to explain the various phenomena of consciousness in terms of representation. A mental representation is a hypothetical internal cognitive symbol or process that represents external reality. Mental representation is the mental imagery of things that are not actually present to the senses. A mental representation is one of the prevailing ways of explaining and describing the nature of ideas and concepts. Mental representations enable representing things that have never been experienced as well as things that do not exist. Although visual imagery is more likely to be recalled, mental imagery may involve representations in any of the senses.
- 3. According to narrative interpretive theories, consciousness is dependent on interpretative judgments. Dan Dennett's 'Multiple Drafts Model' is a prominent example of this. MDM says that at any given moment many types of content are being generated throughout the brain. What makes some of these contents conscious is not that they occur in a privileged spatial or functional location—the so called 'Cartesian Theatre'—but, rather, it is a matter of what Dennett calls 'cerebral celebrity'. MDM says the self emerges from the roughly serial narrative that is constructed out of the various contents in the system.
- 4. Cognitive theories associate consciousness with a distinct cognitive architecture or a special pattern of activity within that structure. For example, Global Workspace Theory describes consciousness in terms of a competition among processors and outputs to 'broadcast' information for widespread access and use.
- 5. Neural theories of consciousness come in many forms, though most in some way concern the so called 'neural correlates of consciousness' or NCCs. A sampling of recent neural theories includes models that appeal to:
    - global integrated fields (Kinsbourne)
    - binding through synchronous oscillation (Singer 1999, Crick and Koch 1990)
    - NMDA channels in neurons (Flohr 1995)
    - patterns of cortical activation modulated by the thalamus (Llinas 2001)

- re-entrant cortical loops (Edelman 1989)
- comparator mechanisms that engage in continuous action-prediction-assessment loops between frontal and midbrain areas (Gray 1995)
- left hemisphere based interpretative processes (Gazzaniga 1988)
- emotive somatosensory hemostatic processes based in the frontal-limbic nexus (Damasio 1999) or in the periaqueductal gray (Panksepp 1998)

  It is possible for several of these to be true, with each contributing some partial understanding to the links between all the diverse forms of consciousness and the brain activity that occurs in many different levels of complex organization and structure.

- 6. According to quantum theories, the nature and basis of consciousness cannot be adequately understood within the framework of classical physics but must be sought within the alternative picture of physical reality provided by quantum mechanics.
- 7. Those who reject physicalist descriptions of consciousness look for ways of modelling it as a non-physical aspect of reality. For example, David Chalmers (1996) has offered an admittedly speculative version of panpsychism which appeals to the notion of information not only to explain synchrony between psycho and physical events, but also to possibly explain the existence of the physical itself as derived from information (i.e., an "it from bit" theory).
- Dr Ginger Campbell, host of the Brain Science podcast, notes that while theories of consciousness do have their differences, there are still three concepts that the most prominent scientific ones all share: 1) consciousness requires a brain; 2) consciousness is a product of evolution; and 3) consciousness is embodied.
- The 'Global Neuronal Workspace Theory' states that consciousness is global information broadcasting within the cortex.
- Antonio Damasio defines consciousness as: mind + self. To him, a 'mind' emerges from the brain when an animal is able to create images and to map the world and its body. Consciousness requires the addition of self-awareness. This begins at the level of the brain stem, with 'primordial feelings'. The 'self' is built up in stages starting with the proto-self made up of primordial feelings, affect alone, and feeling alive. Then the core self is developed when the proto-self can interact with objects and images such that they are modified and there is a narrative sequence. Finally comes the autobiographical self, which is built from the lived past and the anticipated future.
- Feinberg and Mallatt say their theory of 'Neurobiological Naturalism' rests on three principles: 1) life—consciousness is grounded in the unique features of life; 2) neural features—consciousness correlates with neural activity; and 3) naturalism—nothing supernatural is needed. To F&M, the defining features of consciousness are organised in three levels. Level 1) General Biological Features—life, embodiment, processes, self-organising systems, emergence, teleonomy, and adaption. Level 2) Reflexes of animals with nervous systems. Level 3) Special Neurobiological Features—complex hierarchy of networks, nested and non-nested processes (aka recursive), isomorphic representations, mental images, affective states, attention, and memory.
- Joseph LeDoux prefers higher-order representations from among the different theories of consciousness. LeDoux seems to draw a pretty narrow definition around consciousness, but then shows the clear evolutionary history of *aspects* of consciousness along the way, and really advocates for a more subtle use of the term.
- Michael Graziano sees a growing standard model of consciousness whose core set of scientists realise that we are machines and the brain is an information processing machine that thinks it has magic inside it because it builds somewhat imperfect models of the world inside it. This brings together Higher Order Thought Theory, Global Workspace Theory,

and even some Illusionists who talk of consciousness as an illusion. His 'Attention Schema Theory' attempts to provide an integrative picture of these.

- Integrated Information Theory says fundamentally what consciousness is, is the ability of any physical system to exert causal power over itself. This is an Aristotelian notion of causality. For example, the present state of my brain can determine one of the trillion future states of my brain. One of the trillion past states of my brain can have determined my current state, so it has causal power. The more power the past can exert over the present and future, the more conscious the thing that we are talking about is.
- These neuroscientific theories can be summed up into **two main camps**: global and local.
- Global theories describe modules for: balance and coordination; memory; emotion; language; writing; attention, planning, organisation, and reasoning; emotional affect and adaptability; motor / sensory; listening and decoding; reading and interpretation; visual-spatial and visual recognition. There may be specific pathways through each of these modules, e.g. dorsal visual stream, but for general connection between multiple modules there may also be a global workspace. This global workspace coordinates inputs from evaluative systems (value), attentional systems (focusing), long-term memory (past), and perceptual systems (present), into motor control outputs (future). Information in the global workspace is available from all modules and can be seen by each module.
- Local theories say vision, for example, just needs to trigger the right kind of activity patterns in the visual module to be consciously perceived. (E.g. Victor Lamme's local recurrence theory.) Activity that is forward-focused only (from stimulus to response) is unconscious. Feedback activity is required for consciousness. One thing common to all local theories is they say that "activity in frontal and parietal cortices is not absolutely needed for conscious perception to occur."

Phew! That is a heck of a lot of history and detail about this subject.

So, what is consciousness?

We still can't say! And from all of this, you can probably see why there are still so many different dictionary definitions of consciousness. Let's add them to the list of this research too.

Dictionary Definitions of Consciousness

- (*Wikipedia*)—the English word 'conscious' originally derived from the Latin *conscius* (*con-* 'together' and *scio* 'to know'), but the Latin word did not have the same meaning as our word, it meant 'knowing with' or 'having joint or common knowledge with another'
- (*Diderot and d'Alembert's 1753 Encyclopédie*)—the opinion or internal feeling that we ourselves have from what we do
- (*The Oxford Living Dictionary*)—the state of being aware of and responsive to one's surroundings; a person's awareness or perception of something; the fact of awareness by the mind of itself and the world
- (*Cambridge Dictionary*)—the state of understanding and realizing something
- (*Merriam-Webster*)—awareness or sentience of internal or external existence
- (*Webster's Third New International Dictionary*)—1) awareness or perception of an inward psychological or spiritual fact; intuitively perceived knowledge of something in one's inner self; inward awareness of an external object, state, or fact; concerned awareness: interest, concern—often used with an attributive noun; 2) the state or activity that is characterized by sensation, emotion, volition, or thought; mind in the broadest possible sense; something in nature that is distinguished from the physical; 3) the totality in psychology of

sensations, perceptions, ideas, attitudes, and feelings of which an individual or a group is aware at any given time or within a particular time span

Finally, I want to note the hierarchy of consciousness that Mike Smith (aka **Self Aware Patterns**) has developed from his very extensive reading about all of this. To him, consciousness involves:

1. reflexes and fixed action patterns
2. perceptions, representations of the environment, expanding the scope of what the reflexes are reacting to
3. volition, goal directed behaviour, allowing or inhibiting reflexes based on simple valenced cause and effect predictions
4. deliberative imagination, sensory-action scenario simulations assessed on valenced reactions
5. introspection, recursive metacognition, and symbolic thought.

Brief Thoughts

So, attributions of consciousness stretch all the way from it being something as small as the private, ineffable, special feeling that only we rational humans have when we think about our thinking, right on down to it being a fundamental force of the universe that gives proto-feelings to an electron of what it's like to be that electron. Wow. What a mess. As the Wikipedia entry on consciousness notes:

"The level of disagreement about the meaning of the word indicates that it either means different things to different people, or else it encompasses a variety of distinct meanings with no simple element in common."

The Stanford Encyclopedia of Philosophy entry on consciousness comes to a similar conclusion:

"A comprehensive understanding of consciousness will likely require theories of many types. One might usefully and without contradiction accept a diversity of models that each in their own way aim respectively to explain the physical, neural, cognitive, functional, representational, and higher-order aspects of consciousness. There is unlikely to be any single theoretical perspective that suffices for explaining all the features of consciousness that we wish to understand. Thus, a synthetic and pluralistic approach may provide the best road to future progress."

Once again, however, I am drawn to use the 'universal acid' of evolutionary thinking that Dan Dennett described in his 1995 book *Darwin's Dangerous Idea*. If anything stands a chance to usefully provide "a single theoretical perspective" on consciousness, I think it's likely to be that. For a helpful start, consider these passages from Dennett's 2016 paper "Darwin and the Overdue Demise of Essentialism."

"Ever since Socrates pioneered the demand to know what all Fs have in common, in virtue of which they are Fs, the ideal of clear, sharp boundaries has been one of the founding principles of philosophy."

"When Darwin came along with the revolutionary discovery that the sets of living things were not eternal, hard-edged, in-or-out classes but historical populations with fuzzy boundaries, the main reactions of philosophers were to either ignore this hard-to-deny fact or treat it as a challenge: Now how should we impose our cookie-cutter set theory on this vague and meandering portion of reality?"

"We should quell our desire to draw lines. We can live with the quite unshocking and unmysterious fact that there were all these gradual changes that accumulated over many millions of years."

"The demand for essences with sharp boundaries blinds thinkers to the prospect of gradualist theories of complex phenomena, such as life, intentions, natural selection itself, moral responsibility, and consciousness."

Indeed. So, we're looking for a gradualist theory of the complex phenomena of consciousness. We've got a pretty good idea of what we're looking for, based on all the definitions from philosophers, scientists, and dictionaries shown above, but it could be anywhere, and it could have got started at any time. To really spot an emergence and development of consciousness, in order to try to then characterise it, we'll have to look at the history of everything that has ever existed. So, I'll give that a go in the next post. That shouldn't take too long.

17 — From Physics to Chemistry to Biology

23 June 2020

In my last post—**a (sorta) brief history of consciousness**—we saw the enormous range of definitions for consciousness that have existed throughout history among philosophers, scientists, and dictionaries. This led to my conclusion that I ought to go back and look for consciousness "in everything that has ever existed." As David Chalmers **said** about this,

"My background is in mathematics, computer science, and physics, so my first instincts are materialist. To try to explain everything in terms of the processes of physics: e.g. biology in terms of chemistry and chemistry in terms of physics. This is a wonderful great chain of explanation, but when it comes to consciousness, this is the one place where that great chain of explanation seems to break down."

Does it really? For a philosopher like myself who sees the hypothesis of physicalism still holding up, I thought I ought to go through the "great chain of explanation" to see precisely where it does break down. Now, the details of the physicalist picture of the universe is not complete. And it never really can be either since we can't get outside of the universe to know for sure what might be "out there" that we just don't know yet. But we sure know a lot more about the universe now than we did when Descartes kicked this discussion off with the first philosophical usage of the word conscious in 1640. Major mysteries still exist, but I'd like to sketch in what we currently have a good picture of and see how consciousness best fits in there. As I do this, please be generous about what I'm calling a "sketch" for this simple blog post, but hopefully even these faintest outlines will prove helpful.

Physics

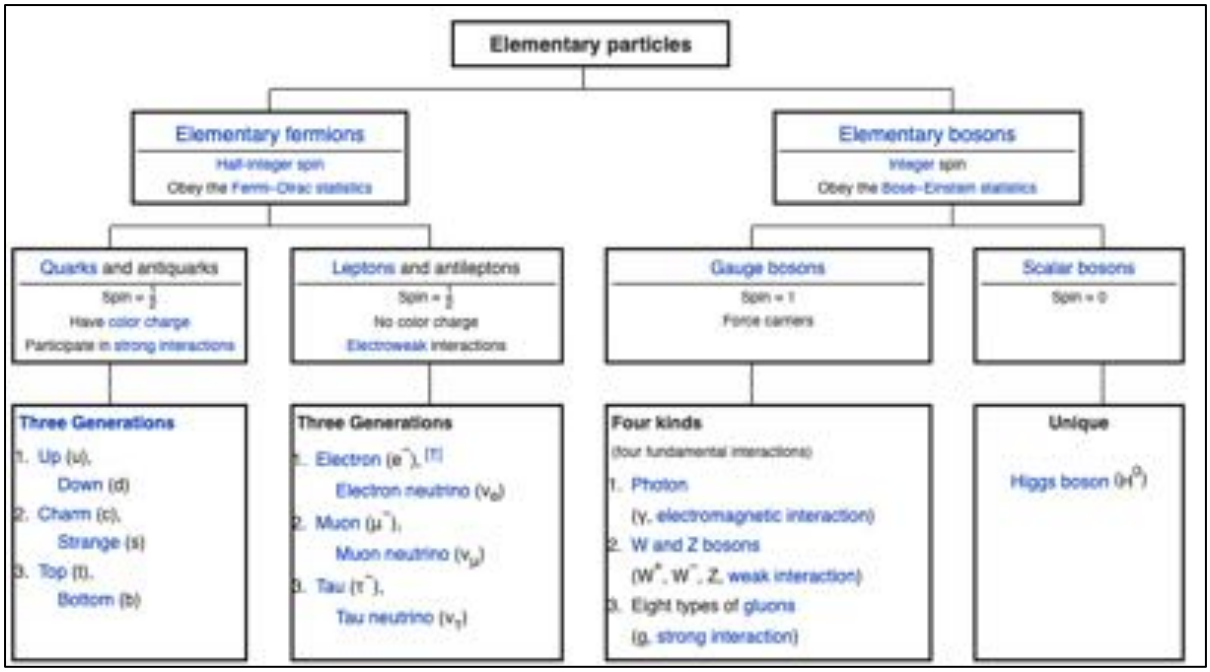
Everything we know about what has existed stretches back to the Big Bang origins of our universe. That's not the whole story, but it's a pretty big one. Among **my basic tenets**, the third one describes a bit about this size:

The universe is composed of trillions and trillions of stars and is currently expanding after a Big Bang and 13-14 billion years of evolutionary processes. We are just another species of animal life on a single planet*

orbiting one of the stars in the universe. (The best current estimate of the age of the universe is 13.75 ± 0.11 billion years. The best current estimate of the number of stars in the universe is from 3 to 100 × 10^22 or between 30 sextillion and 30 septillion.*

The most successful theory describing the basic makeup of this universe is known as the **Standard Model** of particle physics. There are some fundamental physical phenomena that are currently **beyond the Standard Model** such as dark matter, dark energy, matter-antimatter asymmetry, and gravity (which is best described by Einstein's Theory of General Relativity). Plus, Richard Feynman is also famously quoted as saying, "If you think you understand quantum mechanics, you don't understand quantum mechanics." However, no experimental results have definitively contradicted the Standard Model at the five-sigma level, and we are only working on a series about consciousness. Other than in the most extreme panpsychist views, consciousness doesn't appear to operate at the quantum scales of quantum theory. Throughout Ginger Campbell's podcast series on consciousness (**Brain Science episodes 160—163**), she noted that physicists and neuroscientists believe the human body is too warm for quantum computing and the speed and scale is all wrong. At the other extreme, conventional ideas about consciousness don't think of it as operating over cosmic scales either, where dark matter and dark energy show themselves. So, let's not worry too much about the frontiers still being explored in physics. Here, then, in my usual bullet-point format, are a few highlights about the Standard Model of physics. (Sources throughout this article are generally from well-cited Wikipedia entries unless otherwise noted.)

- The Standard Model of particle physics was developed in stages throughout the latter half of the 20th century through the work of many scientists around the world. The current formulation was finalized in the mid-1970s upon experimental confirmation of the existence of quarks.
- The Standard Model is the theory that classifies all known elementary particles and describes three of the four known fundamental forces in the universe—the electromagnetic, weak, and strong interactions, but not the gravitational force.
- The fundamental interactions, also known as fundamental forces, are the interactions that do not appear to be reducible to more basic interactions. The gravitational and electromagnetic interactions produce significant long-range forces whose effects can be seen directly in everyday life. The strong and weak interactions produce forces at minuscule, subatomic distances and govern nuclear interactions. Although the electromagnetic force is far stronger than gravity (gravity is 10x-36 of electromagnetism at the scale of protons/neutrons), it tends to cancel itself out within large objects, so over large distances (on the scale of planets and galaxies), gravity tends to be the dominant force.
- The Standard Model includes 12 elementary particles of spin 1/2, known as fermions. The model also includes gauge bosons, which are force carriers that mediate the strong, weak, and electromagnetic fundamental interactions. The Higgs boson explains why the photon and gluons are massless, and why the other elementary particles have mass.



- In physics, interactions are the ways that particles influence other particles. Gauge bosons are the force carriers that mediate the strong, weak, and electromagnetic fundamental interactions. The Standard Model explains such forces as resulting from matter particles exchanging other particles, generally referred to as force-mediating particles. (Gravitons have been hypothesised as force-mediating particles for gravity but have so far been undetected and mathematically problematic. Einstein's description of the curvature of spacetime remains the best explanation of gravity.) When a force-mediating particle is exchanged, the effect at a macroscopic level is equivalent to a force influencing both of them, and the particle is therefore said to have mediated (i.e., been the agent of) that force.
- Quarks form composite particles called hadrons that contain either mesons (a quark and an antiquark) or baryons (three quarks). The most familiar baryons are protons and neutrons, which make up most of the mass of the visible matter in the universe, as well as forming the components of the nucleus of every atom. The first-generation charged particles do not decay, hence all ordinary (baryonic) matter is made of such particles. Specifically, all atoms consist of electrons orbiting around atomic nuclei, which are constituted of up and down quarks.

These sub-atomic particles and fundamental forces interact in many various ways but are governed by the three laws of thermodynamics. Let's describe those briefly too.

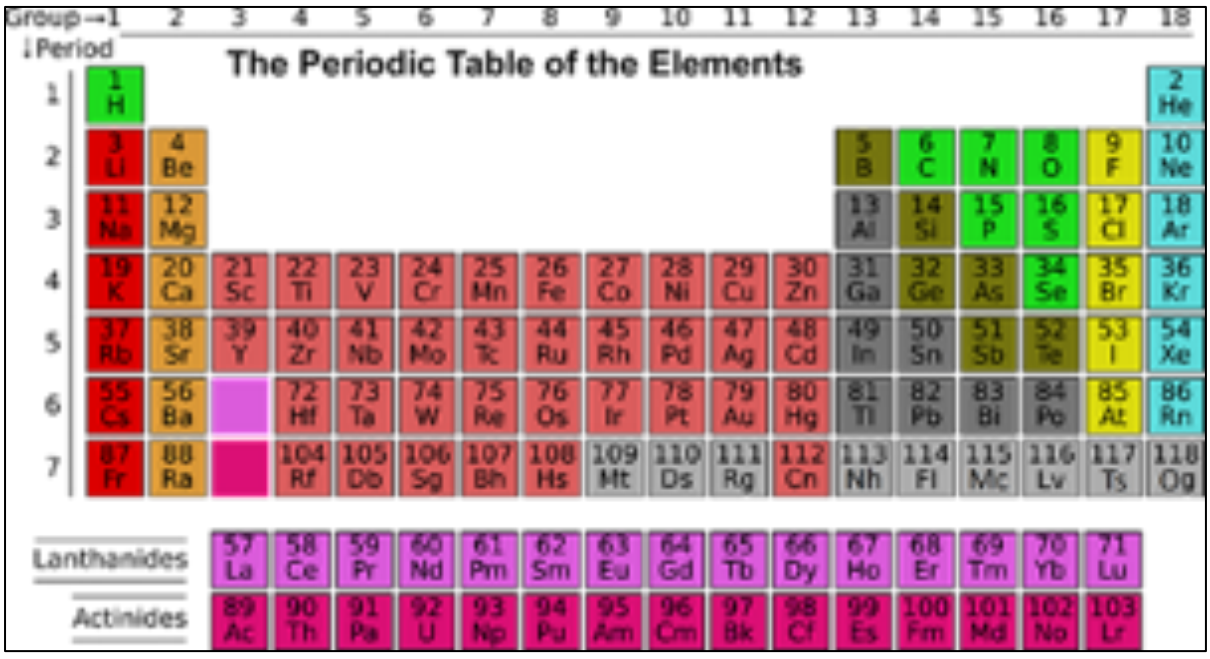
- The laws of thermodynamics define physical quantities, such as temperature, energy, and entropy, which characterise systems at equilibrium. The laws describe the relationships between these quantities, and they form a basis for precluding the possibility of certain phenomena, such as perpetual motion. The three fundamental laws are:
  1. Conservation of Energy — The total energy of an isolated system is constant; energy can be transformed from one form to another but can be neither created nor destroyed. When energy passes, as work, as heat, or with matter, into or out of a system, the system's internal energy changes by the corresponding amount.
  2. Entropy — The total entropy of an isolated system can never decrease over time. (Entropy can be described as "the number of possible configurations of a system's components that is consistent with the state of the system as a whole.")

3. Zero —The entropy of a system approaches a constant value as the temperature approaches absolute zero.

Chemistry

Once these physics particles combine into atoms, we arrive at the field of chemistry. Here are some highlights from that field which contribute to this journey.

- Chemistry is the scientific discipline involved with elements and compounds composed of atoms, molecules, and ions, as well as their composition, structure, properties, behaviour, and the changes they undergo during a reaction with other substances.
- Traditional chemistry starts with the study of elementary particles, atoms, molecules, substances, metals, crystals, and other aggregates of matter. Matter can be studied in solid, liquid, gas, and plasma states, in isolation or in combination. The interactions, reactions, and transformations that are studied in chemistry are usually the result of interactions between atoms, leading to rearrangements of the chemical bonds which hold atoms together.
- The atom is the basic unit of chemistry. It consists of a dense core called the atomic nucleus surrounded by a space occupied by an electron cloud. The nucleus is made up of positively charged protons and uncharged neutrons, while the electron cloud consists of negatively charged electrons which orbit the nucleus.
- A chemical element is a pure substance which is composed of a single type of atom, characterized by its particular number of protons in the nuclei of its atoms, known as the atomic number. The standard presentation of the chemical elements is in the periodic table, which orders elements by this atomic number.



- A molecule is the smallest indivisible portion of a pure chemical substance that has its unique set of chemical properties allowing it to undergo a certain set of chemical reactions with other substances.
- A compound is a pure chemical substance composed of more than one element. The properties of a compound bear little similarity to those of its elements.
- Molecules are held together by covalent bonds, which involve the sharing of electron pairs between atoms. Covalent bonding occurs when these electron pairs form a stable balance

between attractive and repulsive forces between atoms. Covalent bonding does not necessarily require that the two atoms be of the same elements, only that they be of comparable electronegativity.

- Intermolecular forces are the forces which mediate interactions between molecules and other types of neighbouring particles such as atoms or ions. They are weak relative to the intramolecular forces of covalent bonding which hold a molecule together.
- Intermolecular forces are electrostatic in nature; that is, they arise from the interaction between positively and negatively charged molecules. The four key intermolecular forces are: 1) Ionic bonds; 2) Hydrogen bonding; 3) Van der Waals dipole-dipole interactions; and 4) Van der Waals dispersion forces.
- The investigation of intermolecular forces starts from macroscopic observations which indicate the existence and action of forces at a molecular level. Information on intermolecular forces is obtained by macroscopic measurements of properties like viscosity, pressure, volume, and temperature data.
- A chemical reaction is a transformation of some substances into one or more different substances. Chemical reactions usually involve the making or breaking of chemical bonds. Chemical reactions happen at a characteristic reaction rate at a given temperature and chemical concentration. Typically, reaction rates increase with increasing temperature because there is more thermal energy available to reach the activation energy necessary for breaking bonds between atoms.
- There are hundreds or even thousands of specific types of chemical reactions. Oxidation, reduction, dissociation, acid-base neutralization, and molecular rearrangement are some of the commonly used kinds of chemical reactions.
- If you are asked to name the main 4, 5, or 6 types of chemical reactions, here is how they are categorized. The main four types of reactions are synthesis (A + B --> AB), decomposition (AB --> A + B), single replacement (AB + C --> AC + B), and double replacement (AB + CD --> AC + BD). If you're asked for the five main types of reactions, it is these four and then either acid-base or redox (combustion) depending who you ask.
- Chemical reactions are governed by many laws, which have become fundamental concepts in chemistry. Some of them are: Avogadro's law, Beer–Lambert law, Boyle's law, Charles's law, Fick's laws of diffusion, Gay-Lussac's law, Henry's law, Hess's law, Law of definite composition, Law of multiple proportions, Raoult's law.

So, in physics, I noted that exchanges of particles (from the Standard Model), governed by discovered laws (of Thermodynamics), led to a description of fundamental forces being exerted on matter. In chemistry, we see something analogous: exchanges of elements (from the Periodic Table), governed by discovered laws (Avogadro's, Boyle's, Hess's, etc.), leading to descriptions of intra- and inter-molecular forces exerted on matter. Might the same pattern hold for biology?

The Origins and Definitions of Life

In order to get there, we'll have to traverse one of the other great mysteries of science. Besides the mysteries of quantum physics, dark matter, dark energy, and (of course) consciousness, the mystery of how life arose is still a major gap in our knowledge. How exactly did biology arise out of mere chemistry? Wherever gaps in our knowledge occur, supernatural explanations abound. But they offer no actual explanatory power. However, let's take a look at one of the leading natural hypotheses of the origin of life (known technically as abiogenesis), and see how explanatory that might be. Here are some highlights from the transcript of a short video called **The Origin of Life**, which is about Dr. Jack Szostak (who happens to have won a **Nobel Prize** for his work on telomeres) and his work on abiogenesis at the Harvard

Medical School.

(*Note: Unless your biochemistry is very strong, I recommend watching the 10-minute video instead of reading these transcript highlights. The simple diagrams really help understand what is going on.*)

- We know from experiments and observations in the fields of astronomy, chemistry, geology, and meteorology that the early pre-biotic Earth was filled with organic molecules, the building blocks of life. Organic molecules are actually quite common in space. We also know that early life must have been extremely simple, meaning no complex protein machinery. Modern cells separate themselves from the environment with a lipid bilayer (internally hydrophobic, externally hydrophilic). The problem with modern phospholipids is that they are too good at what they do. They form a nearly impenetrable barrier. Modern cells must use proteins to move molecules through their surface. But life didn't have to start with modern chemicals!
- The pre-biotic environment contained many simple fatty acids. Under a range of pH, they spontaneously form stable vesicles (fluid-filled bladders). And they are permeable to small organic molecules, meaning no complex proteins are required to get stuff in. When a vesicle encounters free fatty acids in solution, it will incorporate them. Eating and growth are driven purely by thermodynamics. When a vesicle grows, it adopts a tubular branched shape (because surface area grows faster than volume), which is easily divided by mechanical forces (e.g. waves, currents, rocks, etc.). During mechanical division, none of the contents of the vesicle are lost.
- So far, with naturally occurring simple fatty acids, we have a vesicle that can spontaneously grow and divide. So, what about the genetic material? Again, modern nucleotides are too stable and require complex protein machinery to replicate. The pre-biotic environment contained hundreds of types of different nucleotides (not just DNA and RNA). All it took was for one to self-polymerize. Recent experiments have shown that some of these are capable of spontaneous polymerization where monomers will base pair with a single stranded template and self ligate. In other words, strings X (e.g. AGGTACA) bond with specific strings Z (e.g. CTTGCAC) using hydrogen bonds for each base pair and covalent bonds for further ligation. They can also polymerize in solution and spontaneously form new templates or extend existing templates. No special sequences are required. It's just chemistry.
- So far, we have lipid vesicles that can grow and divide, and nucleotide polymers that can self-replicate, all on their own. But how does it become life? Here's how. Our fatty acid vesicles are permeable to nucleotide monomers, but not polymers. (Single chains can get in; bonded ones can't get out.) Once spontaneous polymerization occurs within the vesicle, the polymer is trapped. Floating though the ocean, the polymer-containing vesicles will encounter convection currents such as those set up by hydrothermal vents. (Fatty acid vesicles are stable under near boiling conditions.) The high temperatures will separate the polymer strands and increase the membrane's permeability to monomers. Once the temperature cools, spontaneous polymerization can occur. And the cycle repeats. Here's where it gets cool.
- The polymer, due to surrounding ions, will increase the osmotic pressure within the vesicle, stretching its membrane. A vesicle with more polymer, through simple thermodynamics, will "steal" lipids from a vesicle with less polymer. This is the origin of competition. They eat each other. A vesicle that contains a polymer that can replicate faster will grow and divide faster, eventually dominating the population.
- Let's review: Monomers diffuse into a fatty acid vesicle. Monomers spontaneously polymerize and copy any template. Heat separates strands and increases membrane permeability to monomers. Polymer backbones attract ions, increasing osmotic pressure.

Pressure on the membrane drives its growth at the expense of nearby vesicles containing less polymer. Vesicles grow into tubular structures. Mechanical forces cause vesicles to divide. Daughter vesicles inherit polymers from the parent vesicle. Polymer sequences that replicate faster will dominate the population. Thus beginning evolution!

- Early genomes were completely random and therefore contained no information. It was their ability to spontaneously replicate, irrespective of sequence, that drove growth and division of the fatty acid vesicles. Any mutation that increases the rate of polymer replication would be selected for. And, as we know, mutation + natural selection = increased information. Early beneficial mutations would include: "change in sequence to contain only the most common nucleotides"; "don't form secondary structures that block replication"; "form sequences that are stable yet separate easily"; and "form secondary structures that show some enzymatic activity".
- Just like RNA, early nucleotides could both store information and function as enzymes. Early polymer enzymes would enhance replication, use high energy molecules in the environment (near thermal vents) to recharge monomers, synthesize lipids from other molecules in the environment, modify lipids so they don't leave a membrane, and that's it. That's it! A simple 2-component system that spontaneously forms in the pre-biotic environment can eat, grow, contain information, replicate, and evolve, simply through thermodynamic, mechanical, and electrical forces. No ridiculous improbability, no supernatural forces, no lightening striking a mud puddle. Just chemistry.
- For much more on this RNA world hypothesis, see the **video series** with Dr. Jack Szostak.

We can't go back and empirically observe this formation of life. Nor can we run an experiment over millions of years to see if it could happen again. But this sure sounds like a plausible theory for the leap from chemistry to biology that kicks off evolution and the development of life from there. When can we say life first arose? That's impossible to say. Through an evolutionary lens life is a gradually emerging phenomenon with no currently clear dividing line or definition, although there are some characteristics that slowly took root and are now generally well established and accepted as defining life. Let's see those.

- The definition of life has long been a challenge for scientists and philosophers, with many varied definitions put forward. This is partially because life is a process, not a substance. Most current definitions in biology are descriptive. Life is considered a characteristic of something that preserves, furthers, or reinforces its existence in the given environment. According to this view, life exhibits all or most of the following traits:

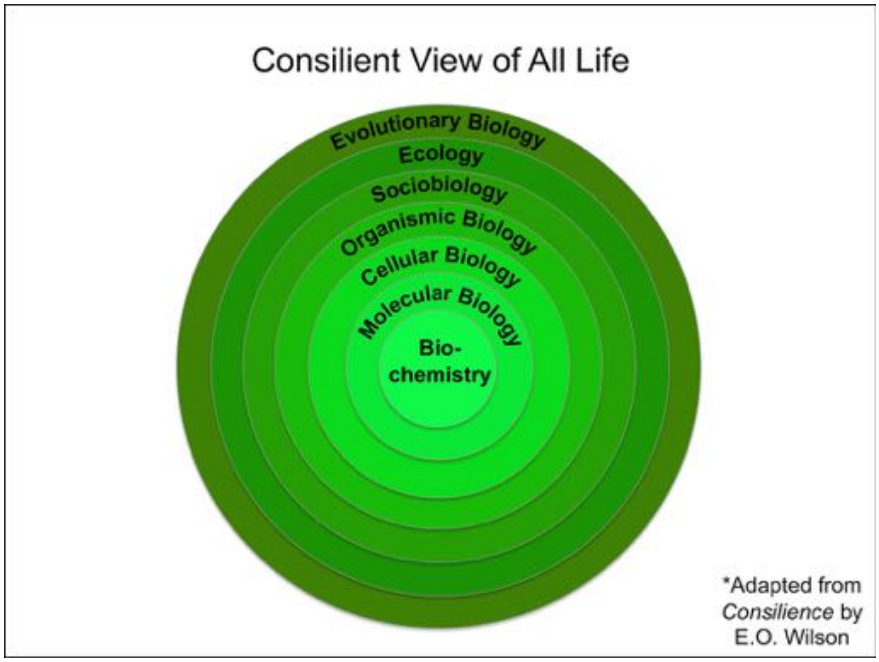
    1. Homeostasis: regulation of the internal environment to maintain a constant state; for example, sweating to reduce temperature.
    2. Organization: being structurally composed of one or more cells—the basic units of life.
    3. Metabolism: transformation of energy by converting chemicals and energy into cellular components (anabolism) and decomposing organic matter (catabolism). Living things require energy to maintain internal organization (homeostasis) and to produce the other phenomena associated with life.
    4. Growth: maintenance of a higher rate of anabolism than catabolism. A growing organism increases in size in all of its parts, rather than simply accumulating matter.
    5. Adaptation: the ability to change over time in response to the environment. This ability is fundamental to the process of evolution and is determined by the organism's heredity, diet, and external factors.

6. Response to stimuli: a response can take many forms, from the contraction of a unicellular organism to external chemicals, to complex reactions involving all the senses of multicellular organisms. A response is often expressed by motion; for example, phototropism (the leaves of a plant turning toward the sun), and chemotaxis (movement of a motile cell or organism, or part of one, in a direction corresponding to a gradient of increasing or decreasing concentration of a particular substance).
7. Reproduction: the ability to produce new individual organisms, either asexually from a single parent organism or sexually from two parent organisms.

- These complex processes, called physiological functions, have underlying physical and chemical bases, as well as signalling and control mechanisms that are essential to maintaining life.
- From a physics perspective, living beings are thermodynamic systems with an organized molecular structure that can reproduce itself and evolve as survival dictates.
- Thermodynamically, life has been described as an open system which makes use of gradients in its surroundings to create imperfect copies of itself. Hence, life is a self-sustained chemical system capable of undergoing Darwinian evolution. A major strength of this definition is that it distinguishes life by the evolutionary process rather than its chemical composition.
- Whether or not viruses should be considered as alive is controversial. They are most often considered as just replicators rather than forms of life. They have been described as "organisms at the edge of life" because they possess genes, evolve by natural selection, and replicate by creating multiple copies of themselves through self-assembly. However, viruses do not metabolize, and they require a host cell to make new products. Virus self-assembly within host cells has implications for the study of the origin of life, as it may support the hypothesis that life could have started as self-assembling organic molecules.
- The study of artificial life imitates traditional biology by recreating some aspects of biological phenomena. Scientists study the logic of living systems by creating artificial environments—seeking to understand the complex information-processing that defines such systems. While life is, by definition, alive, artificial life is generally referred to as data confined to a digital environment and existence.

Biology

So, once physical and chemical processes have self-assembled and evolved into having these characteristics, we get life, the study of which is called biology. In his book ***Consilience: The Unity of Knowledge***, E.O. Wilson proposed seven categories to integrate all of the biological sciences. His seven categories describe the study of life in totality, from the smallest atomic building blocks, to the billions of years of life-history that they have all constructed. Therefore, the simple diagram below of these mutually exclusive, collectively exhaustive categories is actually an astonishingly broad vision of all of the life that has ever existed or will ever exist.



So, to recap where we are, we had sub-atomic particles in physics and the four fundamental forces that affect them, which are governed by the laws of thermodynamics. In chemistry, we had elements from the periodic table and the fundamental bonding forces that hold them together or cause exchange reactions that can be described by many laws. And now we have the material elements of all of life in biology. The obvious holes left would be an account of the fundamental forces that act on life and the laws that describe the various interactions that thereby arise. Note that when we talked about forces in physics, they were described this way:

"When a force-mediating particle is exchanged, the effect at a macroscopic level is equivalent to a force influencing both of them, and the particle is therefore said to have mediated (i.e., been the agent of) that force."

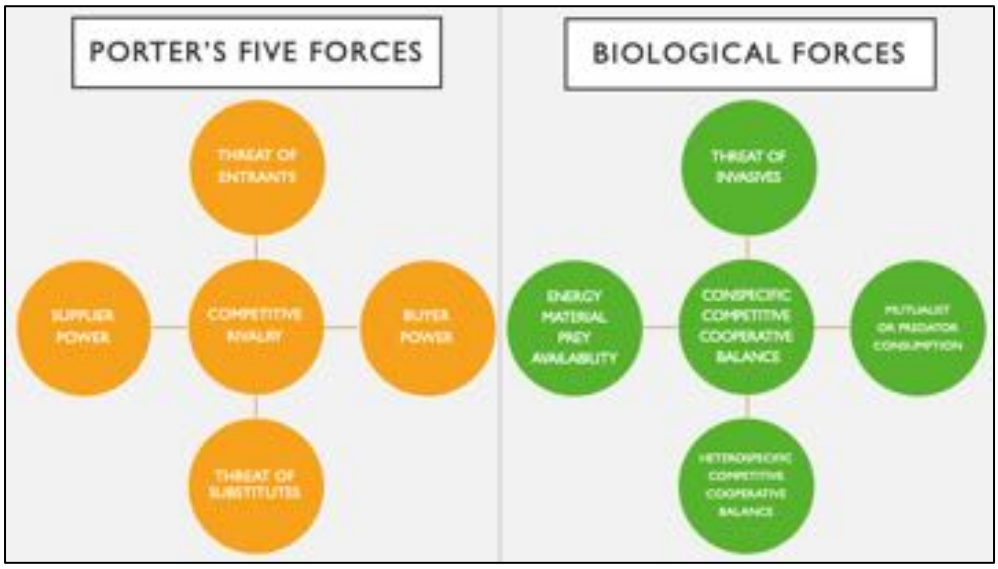
And when we talked about forces in chemistry, they were described this way:

"The investigation of intermolecular forces starts from macroscopic observations which indicate the existence and action of forces at a molecular level."

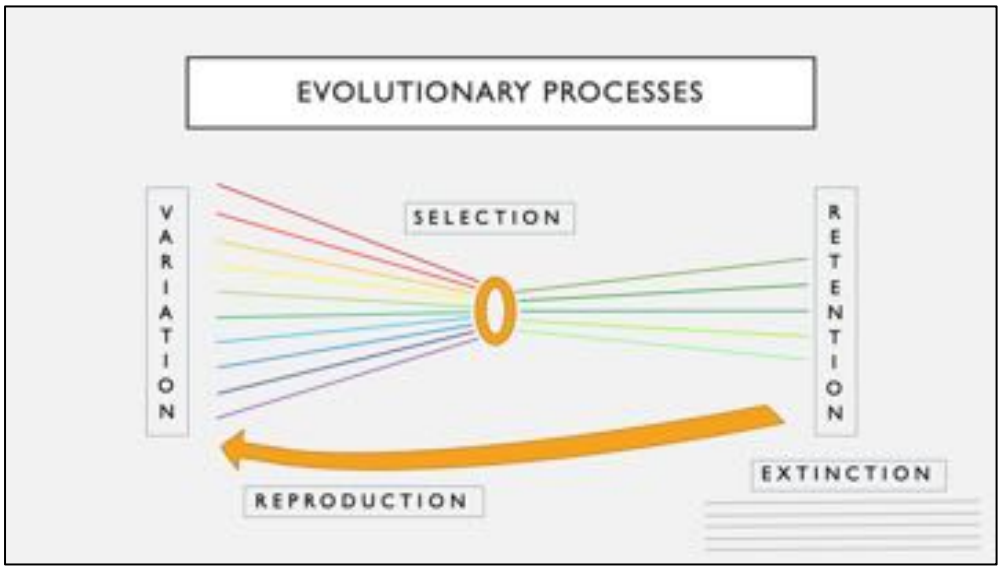
In other words, it is the effect at a macroscopic level that we describe as equivalent to a force. This reminds me of **Porter's Five Forces** in the field of strategic management. Harvard business school professor Michael Porter noted that you could map the competitive environment of any industry in order to understand the industry's attractiveness in terms of profitability. Porter's five forces are exerted by: 1) suppliers (supplier power), 2) buyers (buyer power), 3) entrants (threat of new entrants), 4) substitutes (threat of substitution), and 5) competitors (competitive rivalry). These forces are the influences that change the behaviour of businesses. Strategic analysts can rate their relative strengths in order to predict profitability for a firm and then guide actions to improve a firm's chances for success. Calculations are far too complicated to put stable coefficients in front of formulas to calculate these forces and their combined interactions, but we generally grasp them and can see how they work.

Similarly, there are forces at work in the competitive environment of biological life. However, instead of driving towards the profits that allow a business to survive, biological forces drive life towards lots of actions that aid survival. One significant difference between these forces is that in the business world, cooperation between separate legal entities can often be ruled as

illegal collusion, so B-school graduates tend to focus only on competition. In biology, of course, cooperation plays a major role in the collective struggle for life to survive. Within any ecological niche, however, the same dynamics play out as in the business world. (This makes sense, of course, because the business world is just another ecological niche.) In biology, there is 1) consumption of upstream inputs of energy, material, or prey (suppliers); 2) consumption of downstream outputs by mutualists, micro- or macroscopic predators (buyers); 3) potentially invasive species (threat of entrants); 4) current niche competitors from heterospecifics in other species (substitutes); and 5) the balance between competition and cooperation among conspecifics from the same species (competitive rivalry). In the great interrelated web of life, any individual or species can play any of these parts depending on how you define the circle around an ecosystem for analysis. (We all get eaten at some point I like to say.) And just as the complexity in the system makes Porter's Five Forces impossible to calculate with precision, the same is also true for these biological forces. Yet, we can illustrate them and discuss their relative strengths to aid in analysis.



Are there any laws that describe the results of these forces? Yes. A review of evolutionary processes shows there are two fundamental ones that govern the ultimate goal of survival. As a reminder, here is how evolution works:



Many different rules or laws can be used to describe how proximate goals are reached. (Think, for example, of Elinor Ostrom's Nobel prize-winning **principles for common pool resources**.) But the two orange bottlenecks in the picture above give us the two most fundamental laws that govern biology—natural selection and sexual selection. (Asexual reproducers are, of course, only confined by the first law.) Here are some summary highlights of these two evolutionary laws.

- Natural selection is the differential survival and reproduction of individuals due to differences in phenotype. It is a key mechanism of evolution (which is defined as the change in the heritable traits that are characteristic of a population over generations). Charles Darwin popularised the term 'natural selection', contrasting it with artificial selection, which in his view is intentional, whereas natural selection is not.
- Natural selection acts on the phenotype, the characteristics of the organism which actually interact with the environment, but the genetic (heritable) basis of any phenotype that gives that phenotype a reproductive advantage may become more common in a population. Over time, this process can result in populations that specialise for particular ecological niches (microevolution) and may eventually result in speciation (the emergence of new species, macroevolution).
- Darwin defined natural selection as the "principle by which each slight variation [of a trait], if useful, is preserved."
- In a letter to Charles Lyell in September 1860, Darwin regretted the use of the term Natural Selection, preferring the term Natural Preservation [which sounds less directed and more emergent].
- With the early 20th century integration of evolution via Mendel's laws of inheritance (the so-called Modern Synthesis), scientists generally came to accept natural selection.
- Ernst Mayr recognised the key importance of reproductive isolation for speciation in 1942. W. D. Hamilton conceived of kin selection in 1964. This synthesis cemented natural selection as the foundation of evolutionary theory, where it remains today.
- A second synthesis was brought about at the end of the 20th century by advances in molecular genetics, creating the field of evolutionary developmental biology ('evo-devo'), which seeks to explain the evolution of form in terms of the genetic regulatory programs which control the development of the embryo at the molecular level. Natural selection is here understood to act on embryonic development to change the morphology of the adult body.
- Selection can be classified in several different ways, such as by its effect on a trait, on genetic diversity, by the life cycle stage where it acts, by the unit of selection, or by the resource being competed for.
- Selection has different effects on traits. 'Stabilizing selection' acts to hold a trait at a stable optimum, and in the simplest case all deviations from this optimum are selectively disadvantageous. 'Directional selection' favours extreme values of a trait. The uncommon 'disruptive selection' also acts during transition periods when the current mode is sub-optimal but alters the trait in more than one direction.
- Alternatively, selection can be divided according to its effect on genetic diversity. 'Purifying' or 'negative selection' acts to remove genetic variation from the population (and is opposed by '*de novo* mutation', which introduces new variation). In contrast, 'balancing selection' acts to maintain genetic variation in a population by negative frequency-dependent selection. One mechanism for this is heterozygote advantage, where individuals with two different alleles have a selective advantage over individuals with just one allele. The polymorphism at the human ABO blood group locus has been explained in this way.

- Another option is to classify selection by the life cycle stage at which it acts. Some biologists recognise just two types: 'viability selection', which acts to increase an organism's probability of survival, and 'fecundity selection', which acts to increase the rate of reproduction, given survival.
- Selection can also be classified by the level or unit of selection. 'Individual selection' acts on the individual, in the sense that adaptations are for the benefit of the individual and result from selection among individuals. 'Gene selection' acts directly at the level of the gene. In 'kin selection', gene-level selection provides a more apt explanation of the underlying process. 'Group selection', if it occurs, acts on groups of organisms, on the assumption that groups replicate and mutate in an analogous way to genes and individuals.
- Finally, selection can be classified according to the resource being competed for. 'Sexual selection' results from competition for mates. Sexual selection typically proceeds via fecundity selection, sometimes at the expense of viability. 'Ecological selection' is natural selection via any means other than sexual selection, such as kin selection, competition, and infanticide. Following Darwin, natural selection is sometimes defined as ecological selection, in which case sexual selection is considered a separate mechanism.
- How life originated from inorganic matter remains an unresolved problem in biology. One prominent hypothesis is that life first appeared in the form of short self-replicating RNA polymers. On this view, life may have come into existence when RNA chains first experienced the basic conditions, as conceived by Charles Darwin, for natural selection to operate. These conditions are: 1) heritability, 2) variation of type, and 3) competition for limited resources. The three primary adaptive capacities could therefore logically have been: 1) the capacity to replicate with moderate fidelity (giving rise to both heritability and variation of type), 2) the capacity to avoid decay, and 3) the capacity to acquire and process resources.
- By analogy to the action of natural selection on genes, the concept of memes has arisen as units of cultural transmission, or culture's equivalents of genes undergoing selection and recombination. Memes were first described in this form by Richard Dawkins in 1976 and were later expanded upon by philosophers such as Daniel Dennett as explanations for complex cultural activities, including human consciousness.

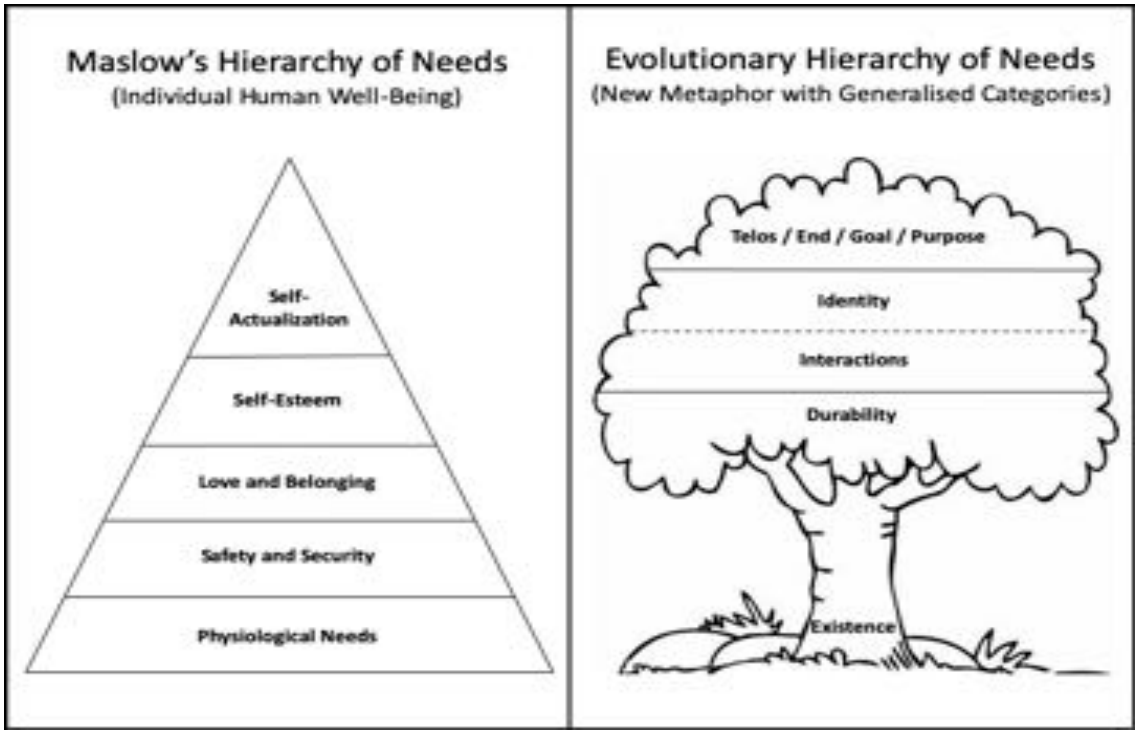
- Sexual reproduction is the most common life cycle in multicellular eukaryotes, such as animals, fungi, and plants. Sexual reproduction does not occur in prokaryotes (organisms without cell nuclei), but they have processes with similar effects such as bacterial conjugation, transformation, and transduction, which may have been precursors to sexual reproduction in early eukaryotes.
- Sexual selection is a mode of natural selection in which some individuals out-reproduce others of a population because they are better at securing mates for sexual reproduction.
- Sexual selection was first proposed by Charles Darwin in *The Origin of Species* (1859) and developed in *The Descent of Man and Selection in Relation to Sex* (1871), as he felt that natural selection alone was unable to account for certain types of non-survival adaptations.
- Darwin's ideas on sexual selection were met with scepticism by his contemporaries and not considered of great importance until the 1930s when biologists decided to include sexual selection as a mode of natural selection. Only in the 21st century have they become more important in biology; the theory is now seen as generally applicable and analogous to natural selection.
- One factor that can influence the type of competition observed is *the population density of males*. Another factor that can influence male-male competition is *the value of the resource to competitors*. Male-male competition can pose many risks to a male's fitness, such as high energy expenditure, physical injury, lower sperm quality, and lost paternity. The risk of

competition must therefore be worth the value of the resource. A third factor that can impact the success of a male in competition is *winner-loser effects*. The winner effect is the increased probability that an animal will win future aggressive interactions after experiencing previous wins, while the loser effect is the increased probability that an animal will lose future aggressive interactions after experiencing previous losses. The outcomes of winner and loser effects help develop and structure hierarchies in nature and is used to support the game theory model of aggression.

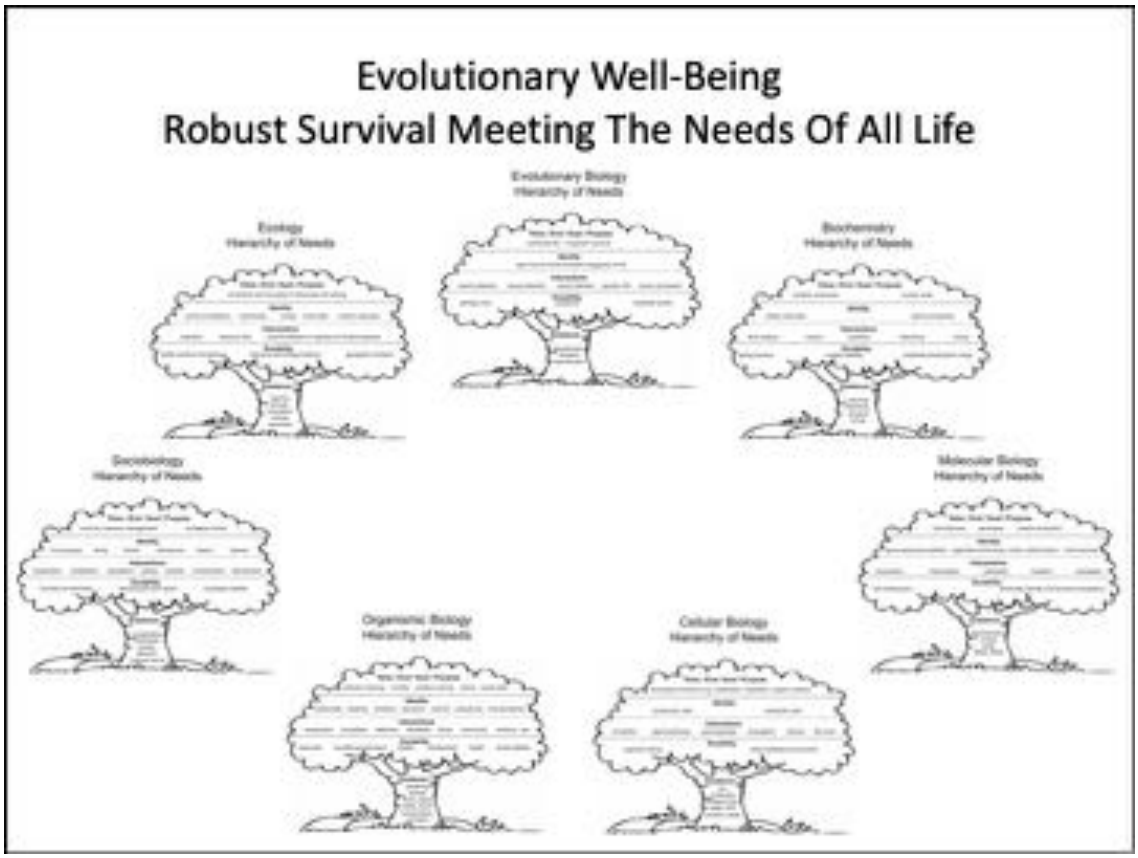
So, we've identified the most fundamental forces and laws affecting life on Earth. There are, of course, many ways that the ultimate question of survival can be determined, and life has been slowly learning to sense and understand these over billions of years. For example, there are so many things that can kill you, your genes, your kin, or your species, and they can all do so in the immediate, medium, or very long term. Living organisms that can sense and respond to more and more of these threats are the ones that will last and emerge over time. Such organisms will sense many, many *needs* to meet all of the threats and exploit all of the opportunities in its environment. Each living organism's unique genetic, environmental, and evolutionary histories are constantly leading to changes in the relative strengths of these needs, but at no point does something outside of the physical realm enter into the equation. All of these needs can be described through physical properties, even if the magnitude of their felt force cannot yet be calculated.

The ever-growing list of threats and opportunities is why the needs of life are ever-growing too. The psychologist Abraham Maslow studied these for individual humans and produced his famous Hierarchy of Needs. In **a 2017 article**, I generalised these and adapted them to apply to all of life, thereby producing something I call an Evolutionary Hierarchy of Needs. Here are some details from that work:

- Maslow's Hierarchy of Needs
  1. Physiological Needs — breathing, food, water, sex, sleep, homeostasis, excretion
  2. Safety and Security — resources, property, employment, health, social stability
  3. Love and Belonging — friendship, family, intimacy
  4. Self-Esteem — confidence, achievement, mutual respect, uniqueness
  5. Self-Actualisation — meaning, purpose, morality, creativity, spontaneity, problem solving
- The evolutionary perspective of our diverse and ever-changing web of life transforms Maslow's hierarchy. Starting at the bottom of the pyramid, we see that the 'physiological' needs of the human are merely the brute ingredients necessary for 'existence' that any form of life might have. In order for that existence to survive through time, the second-level needs for 'safety and security' can be understood as promoting 'durability' in living things. The third-tier requirements for 'love and belonging' are necessary outcomes from the unavoidable 'interactions' that take place in our deeply interconnected biome of Earth. The 'self-esteem' needs of individuals could be seen merely as ways for organisms to carve out a useful 'identity' within the chaos of competition and cooperation that characterizes the struggle for survival. And finally, the 'self-actualization' that Maslow struggled to define could be seen as the end, goal, or purpose that an individual takes on so that they may (consciously or unconsciously) have an ultimate arbiter for the choices that have to be made during their lifetime. This is something Aristotle called '*telos*'.



- Maslow and other psychologists say that individual humans have a need to care for their kin, but what does that really mean once science teaches us that *all of life* is our kin? Rather than just trying to understand and meet the hierarchy of needs for our fellow human individuals, we could collectively spend much more time considering such details *for each realm of E.O. Wilson's consilient view of life*.



Evolutionary Hierarchy of Needs for Human Individuals

1. Existence — breathing, food, water, senses, sleep, touch, homeostasis, excretion
2. Durability — resources, bountiful environment, shelter, employed, health, social stability
3. Interactions — cooperation, competition, defences, friendship, family, community, intimacy / sex
4. Identity — personality, creativity, emotions, decisions, memory, uniqueness, transcendence
5. Telos / End / Goal / Purpose — ultimate meaning, morality, problem solving, culture, social roles

- Evolutionary Hierarchy of Needs for Evolutionary Biology to Occur
  1. Existence — biochemistry, variation, reproduction
  2. Durability — geologic time, adaptation, habitable worlds
  3. Interactions — natural selection, sexual selection, group selection, genetic drift, cosmic processes
  4. Identity — each and all of the consilient categories of the tree of life
  5. Telos / End / Goal / Purpose — continued life, long term survival

- The most important takeaway from a quick pass through the collection of hierarchies is the fact that they are all related. Each level of biology requires a healthy and stable lower level to provide the ingredients for its existence. Each level also needs a healthy and stable level above it to provide a durable habitat for its existence. And the top-most level of evolutionary biology can only kick off (as far as we know from the history of Earth) after the formation of biochemistry in the lowest level. In other words, no matter how much you focus on one seemingly individual tree, it is actually part of an interwoven forest of life.
- This broad perspective is not a luxury for the philosophically minded alone. It is a necessity. If we are to consider needs at all, we must enlarge our circle of concern as far as it will go. If I held that the flourishing of Ed Gibney was the absolute highest priority, others would find me selfish and stop working with me. They might even imprison me depending on my acts of callous selfishness. Only a lack of power and opportunity would stop me from acting for myself by exploiting others. If, instead, the flourishing of my family were the highest priority, I would provoke feuds with clans or mafias around me. If the flourishing of my community were the highest priority, ideological crusades and genocides would be eventual outcomes after intractable disagreements. If the flourishing of my nation were the highest priority, wars would be the result. If the flourishing of my species were the highest priority, we would commit ecocide without a second thought. If my ecosystem were the highest priority, our invasive species would produce monocultures with little resilience in the face of change. It's only when our absolute highest priorities are concerned with the evolution of life in general that we can find ways for all of life to flourish together and ensure its long-term survival.
- And so, it is incumbent upon us, for individual and collective reasons, to not only understand Maslow and other psychologists' hierarchies of human needs, but we must also expand these hierarchies and adapt them to portray a wider and fully evolutionary view as well. As Darwin himself said, there is grandeur in this view of life.

Brief Comments

Phew! That concludes my (very) brief history of everything that has ever empirically existed. I've gone from the appearance of sub-atomic particles and fundamental forces after the Big Bang up to the longest-term view of all of the needs required for the evolution of life. This

gives us an outline of the "great chain of explanation" that Chalmers described at the top of this article as "biology in terms of chemistry and chemistry in terms of physics." All along the way, we see exchanges of particles defining changes in forces that affect matter according to natural laws that are regular and can be studied empirically.

Where might consciousness fit into all of this??

In **my last post** about the history of philosophical and scientific studies of consciousness, I noted an etymological root of the word that I think offers some help. Wikipedia noted that the English word 'conscious' originally derived from the Latin *conscius* where *con* meant 'together' and *scio* meant 'to know'. According to this literal interpretation, to be conscious would be 'to know', which requires a knower. And to 'know together, this conscious thing would need to know at least two things.

Do sub-atomic particles feeling fundamental forces meet these criteria? No. Do elements from the periodic table feeling intermolecular forces meet these criteria? Also no. Do living things feeling biological forces meet these criteria? Yes. Once chemistry makes the jump to biology, the resulting proto-life forms have a defined self AND they begin to compete for resources with other potential entrants, substitutes, or conspecifics in order to self-replicate and survive. They know (from an outsiders' perspective) what they are AND what they need. A radical panpsychist might claim that a quark can feel the strong nuclear force, or a hydrogen atom can feel the covalent bond in H2O, but I think a more natural joint to carve a philosophical place for consciousness is in the biological realm where life responds to biological forces to survive. Could artificial life also respond to these forces and be declared conscious? I think yes, although the "feeling of what it is like" to be such life would be very different from current biological life forms that are built from organic chemistry. We already believe the feeling of what is like to be a bat is likely very different from that of a cuttlefish, so the difference would be even greater for artificial life given the much larger change in underlying mechanisms. Yet both could be considered conscious in my definition.

As Mark Solms wrote in **The Hard Problem of Consciousness and the Free Energy Principle**, "There cannot be any objects of consciousness without a *subject* of consciousness. You cannot experience objects unless *you* are there to experience them." The earliest forms of life were the first such subjects who experienced a need. As these lifeforms evolved to sense more and more needs, their consciousness grew in quantity and quality of varieties. I acknowledge that this view of consciousness—as an evolved trait of living things sensing and responding to biological forces—raises a lot of questions. To try and answer them—at least as well as the current state of science allows—we'll need a comprehensive understanding of this position. In my next post, I'll introduce a framework that can help lead us through that kind of comprehensive explanation. After that, I'll step through the framework item-by-item until I can finally arrive at my full evolutionary theory of consciousness (for now).

18 — Tinbergen's Four Questions

3 July 2020

In **my last (long) post**, I noted that "I think a more natural joint to carve a philosophical place for consciousness is in the biological realm where life responds to biological forces in order to survive. … I acknowledge that this view of consciousness raises a lot of questions. To try and answer them —at least as well as the current state of science allows—we'll need a comprehensive understanding of this position." And I said that "In my next post, I'll introduce a framework that can help lead us through that kind of comprehensive explanation."

Where should we look for such a framework? We could turn to Aristotle since his empirical studies of plants and animals led him **to be considered the founder of the science of biology**. During his observations and classifications, Aristotle developed a framework known as the **four causes**. And he said, "we do not have knowledge of a thing until we have grasped its why, that is to say, its cause." The four kinds of causes are **described briefly** as:

1. The **efficient cause**. This is something outside of the thing that is under consideration, which is responsible for the origination of that thing. For a table, this is a carpenter.
2. The **material cause**. This is the material "out of which" a thing is composed. For a table, this might be wood.
3. The **formal cause**. This is the form or shape of a thing which makes up the general definition of that thing. For a table, this could be its blueprint.
4. The **final cause**. This is the goal or the purpose (*telos* in Greek) for which a thing originated and at which it aims. For a table, this could be dining.

These causes—arguably better translated as "**explanations**"—weren't considered to be separable and mutually exclusive things that all operated on their own. Rather, they are just different aspects that work together to explain something in its full context. This line of thinking about biological organisms remained successful for about 2000 years. It was developed and used by Neoplatonists in antiquity, **Averroes** and **Aquinas** in the Middle Ages, and right on through to the 19th century. An elderly **Charles Darwin** even famously **said**, "Linnaeus and Cuvier have been my two gods, though in very different ways, but they were mere school-boys to old Aristotle."

Looking at these four causes now, however, we see that Darwin undermined them completely. The four causes are static, they lack any sense of evolutionary history, and the final *telos* cause has been flipped upside down by Darwin's "**strange inversion of reasoning**." As the world got to grips with that, Julian Huxley (who was the grandson of "Darwin's Bulldog" **Thomas Huxley**) gave us an updated evolutionary framework in his 1942 book *Evolution: The Modern Synthesis*. That looked at three major aspects of biological facts: 1) mechanistic-physiological, 2) adaptive-functional, and 3) evolutionary or historical aspects.

This is much better, but the final framework that evolutionary biologists still use today came a few decades later from **Nicholaas Tinbergen**. Tinbergen won a Nobel Prize in 1973 for his contributions as one of the founders of ethology (the study of animal behaviour), and his

1963 paper "On aims and methods of Ethology" **has** "become a classic that gives evolutionary students of behaviour a basic framework for their agenda. Tinbergen proposes, in what has subsequently become known as **Tinbergen's Four Questions**, that to achieve a complex understanding of a particular phenomenon, we may ask different questions which are mutually non-transferable."

What that phrase 'mutually non-transferable' really means in this case is your classic 2x2 matrix with 2 options for each of 2 different variables. In this case, Tinbergen considered static vs. dynamic views as well as proximate vs. ultimate views. The static view looks at the current form of an organism. The dynamic view looks at the historical sequence that led to it. The proximate view considers how an individual organism's structures function, whereas the ultimate view asks why a species evolved the structures that it has. Setting up this 2x2 matrix yields the following four areas for consideration:

1. **Mechanism (causation)**. This gives mechanistic explanations for how an organism's structures currently work. (Static + Proximate)
2. **Ontogeny (development).** This considers developmental explanations for changes in individuals, from their original DNA to their current form. (Dynamic + Proximate)
3. **Function (adaptation)**. This looks at a species trait and how it solves a reproductive or survival problem in the current environment. (Static + Ultimate)
4. **Phylogeny (evolution)**. This examines the entire history of the evolution of sequential changes in a species over many generations. (Dynamic + Ultimate)

This framework adds the important consideration of ontogeny to Huxley's three major aspects of biology. (How did he tell the story of a frog without the story of a tadpole?) The fact that Tinbergen arrived at four considerations is perhaps why "it has been repeatedly pointed out that this concept is derived from Aristotle's Four Causes." A paper called "**Was Tinbergen an Aristotelian? Comparison of Tinbergen's Four Whys and Aristotle's Four Causes**" thinks that "in general, they parallel very well" but honestly that feels a bit forced to me. (Try to match them up to see for yourself.) Tinbergen apparently never mentioned Huxley or Aristotle, but regardless of his inspiration, he now has the "standard framework in the behavioural sciences."

If that's the case, then why hasn't consciousness already been considered using this framework? I was sure someone would have done this already, but I couldn't find it. In fact, one of the top search results for "Tinbergen and Consciousness" was a paper called "**The Mind-Evolution Problem: The Difficulty of Fitting Consciousness in an Evolutionary Framework**" written by Yoram Gutfreund in 2018 in *Frontiers in Psychology*. I'll say more about that later but let me quickly trace the history of this difficulty.

Firstly, Tinbergen wrote in the 1950's and 1960's at the height of the behaviourist movement in psychology which tried to get rid of cognitive studies. Perhaps because of this, Tinbergen himself thought his framework did not apply to consciousness. As he **wrote**, "Psychology does not come into contact with objective study of the lowest levels such as the reflex level, because introspection does not reach them. At the higher level, introspection brings us into contact with an aspect of behaviour that is out of reach of objective study. … As scientists, we have to recognise the duality of our thinking and to accept it." Duality?! Well I think I see a fatal flaw in his thinking about this.

Even Dan Dennett, however, apparently subscribed to this separation of consciousness from ethology. In his doctoral thesis "**Content and Consciousness**", published in 1969,

Dennett is **described** as saying that "in the intentional case the antecedent (intention) cannot be described or defined independently from the consequent (action), it [therefore] cannot be properly regarded in terms of cause and effect in the natural sciences' sense; antecedent-consequent relationships in the behaviouristic or ethological tradition (stimulus-response relationships) however can. These approaches are incompatible." To untangle that jargon, Dennett meant that we couldn't grasp our intentions as they arise and separate them into a standard model of cause and effect. I believe this is a problem we can solve now, but even the man who later called evolution a "**universal acid**" didn't originally see how it could eat into this particular method of studying consciousness.

In fact, **The Oxford Companion to Consciousness** notes, "Conspicuously absent from [Tinbergen's] 'classical' ethology were issues involving consciousness. Thus, in ethology as well as in behaviouristic experimental and comparative psychology, questions of animal consciousness and related ones involving emotion and subjective experiences in general largely became taboo. To recognise a broadened view of ethology that encompassed cognitive, emotional, and conscious processes, Burghardt (1997) added a fifth aim, the study of private experience."

When neuroscientists finally broke through these taboos in the 1980's and developed the field of **Consciousness Studies**, they kept some of this separation intact. As the founder of animal cognition studies Donald Griffin **noted**: "Crick and Koch (1998), leaders in the renewal of scientific studies of consciousness, take it for granted that monkeys are conscious. But they prefer to defer investigating nonhuman consciousness because they claim that 'when one clearly understands, both in detail and in principle, what consciousness involves in humans, then will be the time to consider the problem of consciousness in much simpler animals.'" This might seem like prudent behaviour, but I agree with Griffin who further **said**, "Restricting scientific investigation to the most complex of all known brains may be unwise, however, for insofar as consciousness can be identified and analysed in a variety of animals, certain species might turn out to be especially suitable for investigating its basic attributes."

Many neuroscientists have indeed studied the evolutionary origins of consciousness (see especially **Feinberg and Mallat**, and **LeDoux**), but not using Tinbergen as far as I can tell. In the paper I mentioned earlier about "**The Difficulty of Fitting Consciousness in an Evolutionary Framework**", we can partly see why. The author Yoram Gutfreund noted that "the question of how the mind emerged in evolution (the mind-evolution problem) is tightly linked with the question of how the mind emerges from the brain (the mind-body problem). It seems that the evolution of consciousness cannot be resolved without first solving the 'hard problem'. Until then, I argue that strong claims about the evolution of consciousness based on the evolution of cognition are premature and unfalsifiable." But I already dismissed the worst of the hard problem in **my post about it**.

So, scientists remain unable, unwilling, or uninterested in tackling the philosophical problems of consciousness. They have also, in my view, drawn too small or too large a circle around the term to accurately describe it. What about philosophers? Can they attack these problems from their side?

In a lecture series from The Great Courses about **Mind-Body Philosophy**, the professor Patrick Grim of SUNY Stony Brook made a sketch of this in his final lecture titled "**A Philosophical Science of Consciousness?**" He proposed that we need an integration of philosophy, brain science, and AI in order to stand a better chance of grasping consciousness. He didn't have this new science worked out yet, but he offered: "a sketch of a speculative

plan. First, figure out what consciousness is by figuring out what consciousness is for. What does it do that other cognitive processing could not? Second, analyse the process in abstract terms. What function is needed to produce the process we've identified as what consciousness is for? Then move to concrete specifics. How does the brain perform that function?" He finished by imagining that AI researchers could then build simulated brains using these discoveries and see where that got us in a kind of looped project with all sides informing one another for further progress. That iterative scientific process sounds great, but Grim entirely missed out on 2 of Tinbergen's 4 questions. He only proposed we look at the 1st (mechanism) and 3rd (function) from my list above. He entirely ignored the 2nd (ontogeny) and 4th (phylogeny), which provides a striking example of the lack of evolutionary thinking among philosophers.

Well, let's rectify that as best as I can. In **my last post**, I said consciousness involves living organisms, governed by the laws of natural and sexual selection, sensing and responding to biological forces. With that in mind, I think we can gain a lot of detail about this general definition by stepping through Tinbergen's four questions one at a time. So, that's what I'll do with my next four posts. Please bear with me as I work on that for a while.

19 — The Functions of Consciousness

30 July 2020

Here goes. In **my last post**, I reiterated my concept of consciousness as involving living organisms, governed by the laws of natural and sexual selection, sensing and responding to biological forces. With that in mind, I said we could gain a lot of detail about this general definition by stepping through Tinbergen's four questions one at a time. As a quick reminder, I listed those as: 1) mechanism (causation), 2) ontogeny (development), 3) function (adaptation), and 4) phylogeny (evolution).

So, which one should we tackle first? I believe we have to start by trying to nail down the functions of consciousness. Without that, how would we even know what to look for in terms of mechanisms, ontogeny, and phylogeny? This is a big task though. Remember that Tinbergen carved out the biological view of function in his 2x2 matrix as static (the current form of an organism) and ultimate (why a species evolved the structures that it has). This "static + ultimate" view of a function means that we are looking for a species trait that solves a reproductive or survival problem in the current environment. The problem with trying to do this for "consciousness" is that it is such a multi-faceted complex concept, there are therefore many, many aspects of consciousness that solve many, many reproductive and survival problems. And if we are trying to do so for a general definition of consciousness that applies across all organisms, then we have an even bigger set of possibilities that needs to encompass all of the evolutionary history of life. It's a good thing we already covered **a brief history of everything that has ever existed**!

Since a review of the functions of consciousness can quickly get unwieldy, I'm going to write it in a way that helps us (i.e. me) hold onto the thread of the plot. I'm going to write a series of numbered statements (42 in all) with the justification for each one coming after the statement. You can just quickly read the statements to get the gist of the argument if you like. Or you can dip into the rest of the ~10,000 words to find any details you might want. Hopefully that will work well for a variety of readers with lots of differing backgrounds. Let's begin.

1. Naming an evolved function for consciousness has proven to be very difficult and there is no widely accepted position on this.

- If consciousness exists as a complex feature of biological systems, then its adaptive value is likely relevant to explaining its evolutionary origin, though of course its present function, if it has one, need not be the same now as when it first arose. (**Consciousness Entry in Stanford Encyclopedia**)
- Why did evolution result in creatures who were more than just informationally sensitive? [Instead, they are 'experientially sensitive' too.] There are, to the best of our knowledge, no good theories about this. … Surely we jest, the reader might think. There must be good theories for why consciousness evolved. Well we have looked far and wide and no credible theories emerge. … There are as yet no credible stories about why subjects of experience emerged, why they might have won—or should have been expected to win—an evolutionary battle against very intelligent zombie-like information sensitive organisms. At least this has not been done in a way that provides a respectable theory for why subjects of experience gained hold in this actual world—for why we are not zombies. (**Flanagan and Polger**)

2. A common way to express this difficulty is to ask what life would be like without consciousness? Would life as a "zombie" look any different?

- Zombie thought experiments highlight the need to explain why consciousness evolved and what function(s) it serves. This is the hardest problem in consciousness studies. … If systems "just like us" could exist without consciousness, then why was this ingredient added? Does consciousness do something that couldn't be done without it? (**Flanagan and Polger**)
- Why doesn't all this information-processing go on "in the dark," free of any inner feel? Chalmers (1995) insists that consciousness cannot be explained in functional terms. He claims that reducing consciousness (as we experience it) to a functional mechanism will *never* solve the hard problem. (**Solms**)

3. "Zimboes" show the preposterousness of these zombie claims.

- Todd Moody [notes that] although "it is true that zombies who grew up in our midst might become glib in the use of our language, including our philosophical talk about consciousness [and other mentalistic concepts], a world of zombies could not originate these exact concepts." … Zombies, lacking the inner life that is the referent for our mentalistic terms, will not have concepts such as 'dreaming', 'being in pain', or 'seeing'. This, Moody says, will reveal itself in the languages spoken on Zombie Earth, where terms for conscious phenomena will never be invented. The inhabitants of Zombie Earth won't use the relevant mentalistic terms and thus will show "the mark of zombiehood". (**Flanagan and Polger**)
- Todd Moody's (1994) essay on zombies, and Owen Flanagan and Thomas Polger's commentary on it, vividly illustrate a point I have made before, but now want to drive home: when philosophers claim that zombies are conceivable, they invariably under-estimate the task of conception (or imagination), and end up imagining something that violates their own definition. … Only zimboes could pass a demanding Turing Test, for instance. … Zimboes think they are conscious, think they have qualia, think they suffer pains—they are just 'wrong' (according to this lamentable tradition), in ways that neither they nor we could ever discover! … Zimboes are so complex in their internal cognitive architecture that whenever there is a strong signal in either the pain or the lust circuitry, all these 'merely informational' effects, (and the effects of those effects, etc.) are engendered. That's why zimboes, too, wonder why sex is so sexy for them [but not for simpler zombies, such as insects] and why their pains have to 'hurt'. If you deny that

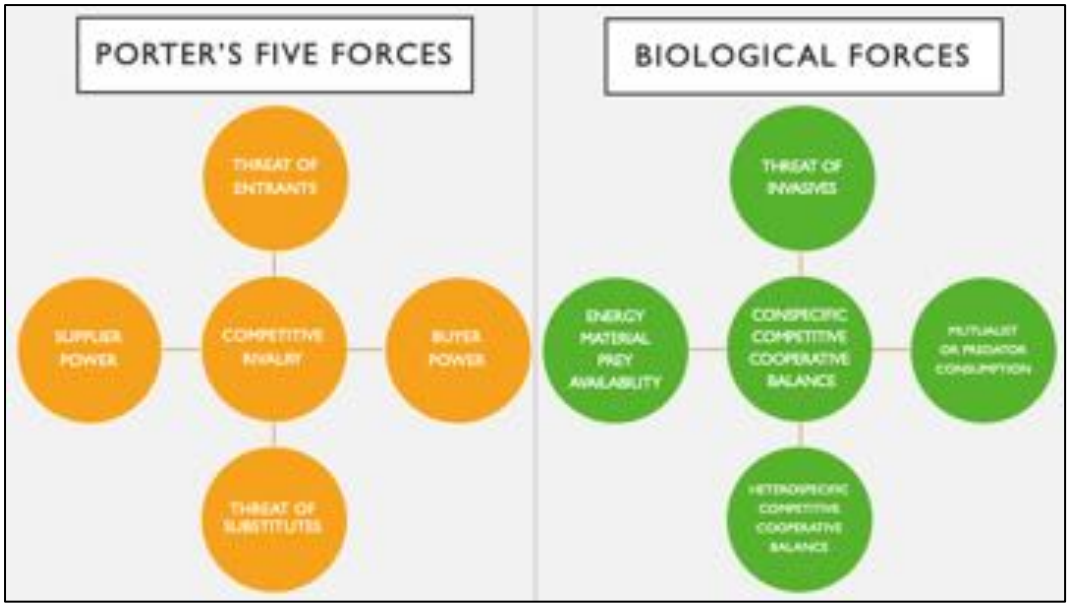
zimboes would wonder such wonders, you contradict the definition of a zombie. … Zombies would pull their hands off hot stoves, and breed like luna moths, but they wouldn't be upset by memories or anticipations of pain, and they wouldn't be apt to engage in sexual fantasies. No. While all this might be true of simple zombies, zimboes would be exactly as engrossed by sexual fantasies as we are, and exactly as unwilling to engage in behaviours they anticipate to be painful. If you imagine them otherwise, you have just not imagined zombies correctly. (**Dennett**)

4. So, what is the purpose of consciousness? That's a poorly formed question because there is no one thing that consciousness is, so there is no one purpose that it is for.

- The question of adaptive advantage, however, is ill-posed in the first place. If consciousness is (as I argue) not a single wonderful separable thing ('experiential sensitivity') but a huge complex of many different informational capacities that individually arise for a wide variety of reasons, there is no reason to suppose that 'it' is something that stands in need of its own separable status as fitness-enhancing. It is not a separate organ or a separate medium or a separate talent. To see the fallacy, consider the parallel question about what the adaptive advantage of health is. (**Dennett**)
- I am not suggesting that there are not numerous empirical problems about the various forms of consciousness. We should like to understand, not what consciousness is for, but rather what sleep is for. It is of interest to know the neural mechanisms involved in perceptual consciousness (i.e. of having one's attention caught by something in one's field of perception). It is important to discover how the brain maintains intransitive consciousness. And so on. My point is merely that the so-called 'hard problem' of consciousness, and the battery of related questions often cited by philosophers are merely conceptual confusions masquerading as empirical questions. (**Hacker**)

5. We must drop the essentialist language of consciousness. Consciousness isn't a thing that just turns on. It involves the slow accrual of many properties. I defined it around detecting and responding to biological forces.

- I think a more natural joint to carve a philosophical place for consciousness is in the biological realm where life responds to biological forces in order to survive. (**Post 17**)
- In the field of strategic management. Harvard business school professor Michael Porter noted that you could map the competitive environment of any industry in order to understand the industry's attractiveness in terms of profitability. Porter's five forces are exerted by: 1) suppliers (supplier power), 2) buyers (buyer power), 3) entrants (threat of new entrants), 4) substitutes (threat of substitution), and 5) competitors (competitive rivalry). (**Post 17**)
- In biology, there is 1) consumption of upstream inputs of energy, material, or prey (suppliers); 2) consumption of downstream outputs by mutualists, micro- or macroscopic predators (buyers); 3) potentially invasive species (threat of entrants); 4) current niche competitors from heterospecifics in other species (substitutes); and 5) the balance between competition and cooperation among conspecifics from the same species (competitive rivalry). (**Post 17**)



6. Defining consciousness this way implies that the processes of consciousness began with the origins of life. Our current best guess for how that occurred involves chemical and physical processes leading to simple constructions that were separable, stable, and replicable.

- The pre-biotic environment contained many simple fatty acids. Under a range of pH, they spontaneously form stable vesicles (fluid-filled bladders). When a vesicle encounters free fatty acids in solution, it will incorporate them. Eating and growth are driven purely by thermodynamics. … The pre-biotic environment contained hundreds of types of different nucleotides (not just DNA and RNA). All it took was for one to self-polymerize. … No special sequences are required. It's just chemistry. … So far, we have lipid vesicles that can grow and divide, and nucleotide polymers that can self-replicate, all on their own. But how does it become life? Here's how. Our fatty acid vesicles are permeable to nucleotide monomers, but not polymers. (Single chains can get in; bonded ones can't get out.) Once spontaneous polymerization occurs within the vesicle, the polymer is trapped. Floating though the ocean, the polymer-containing vesicles will encounter convection currents such as those set up by hydrothermal vents. The high temperatures will separate the polymer strands and increase the membrane's permeability to monomers. Once the temperature cools, spontaneous polymerization can occur. And the cycle repeats. Here's where it gets cool. The polymer, due to surrounding ions, will increase the osmotic pressure within the vesicle, stretching its membrane. A vesicle with more polymer, through simple thermodynamics, will "steal" lipids from a vesicle with less polymer. This is the origin of competition. They eat each other. A vesicle that contains a polymer that can replicate faster will grow and divide faster, eventually dominating the population. Thus beginning evolution! (**Post 17**)

7. These earliest structures satisfy at least 3 of the 7 major traits that currently define life: organisation, growth, and reproduction.

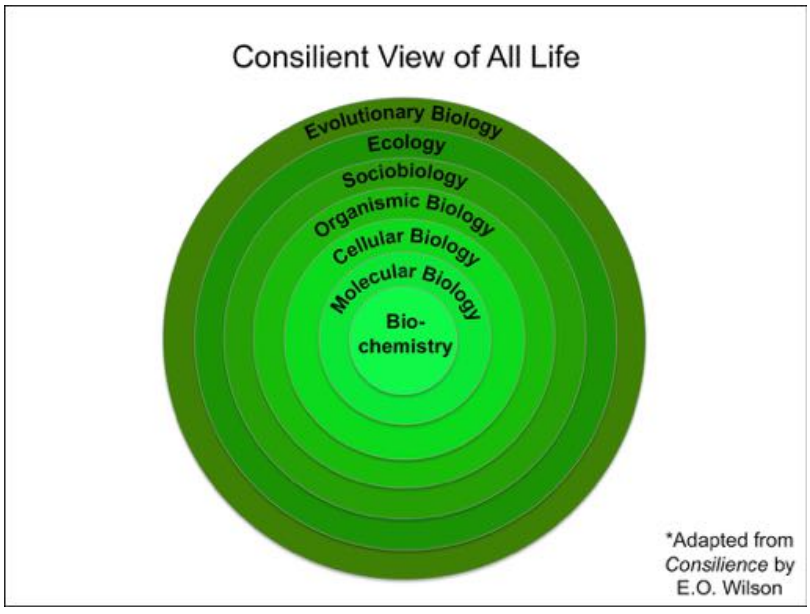
- The definition of life has long been a challenge for scientists and philosophers, with many varied definitions put forward. This is partially because life is a process, not a substance. Most current definitions in biology are descriptive. Life is considered a characteristic of

something that preserves, furthers, or reinforces its existence in the given environment. According to this view, life exhibits all or most of the following traits:

1. Homeostasis: regulation of the internal environment to maintain a constant state; for example, sweating to reduce temperature.
2. Organisation: being structurally composed of one or more cells—the basic units of life.
3. Metabolism: transformation of energy by converting chemicals and energy into cellular components (anabolism) and decomposing organic matter (catabolism). Living things require energy to maintain internal organization (homeostasis) and to produce the other phenomena associated with life.
4. Growth: maintenance of a higher rate of anabolism than catabolism. A growing organism increases in size in all of its parts, rather than simply accumulating matter.
5. Adaptation: the ability to change over time in response to the environment. This ability is fundamental to the process of evolution and is determined by the organism's heredity, diet, and external factors.
6. Response to stimuli: a response can take many forms, from the contraction of a unicellular organism to external chemicals, to complex reactions involving all the senses of multicellular organisms. A response is often expressed by motion; for example, phototropism (the leaves of a plant turning toward the sun), and chemotaxis (movement of a motile cell or organism, or part of one, in a direction corresponding to a gradient of increasing or decreasing concentration of a particular substance).
7. Reproduction: the ability to produce new individual organisms, either asexually from a single parent organism or sexually from two parent organisms. (**Post 17**)

8. Within E.O. Wilson's consilient view of all of life, this gets us from biochemistry to molecular biology.

- In his book ***Consilience: The Unity of Knowledge***, E.O. Wilson proposed seven categories to integrate all of the biological sciences. His seven categories describe the study of life in totality, from the smallest atomic building blocks, to the billions of years of life-history that they have all constructed. Therefore, the simple diagram below of these mutually exclusive, collectively exhaustive categories is actually an astonishingly broad vision of all of the life that has ever existed or will ever exist. (**Post 17**)



9. Any changes to these biological molecules would generate forces. These forces are exerted on singularly identifiable objects.

- Molecules are held together by covalent bonds, which involve the sharing of electron pairs between atoms. Covalent bonding occurs when these electron pairs form a stable balance between attractive and repulsive forces between atoms. Covalent bonding does not necessarily require that the two atoms be of the same elements, only that they be of comparable electronegativity. … Intermolecular forces are the forces which mediate interactions between molecules and other types of neighbouring particles such as atoms or ions. They are weak relative to the intramolecular forces of covalent bonding which hold a molecule together. … Intermolecular forces are electrostatic in nature; that is, they arise from the interaction between positively and negatively charged molecules. The four key intermolecular forces are: 1) Ionic bonds; 2) Hydrogen bonding; 3) Van der Waals dipole-dipole interactions; and 4) Van der Waals dispersion forces. (**Post 17**)

10. I propose that these chemical forces, once in service of biological needs, are the defined starting point for turning objects into subjects.

- Foundational to what we call psychology is the *subjective* observational perspective. The fact that self-organizing systems must monitor their own internal states in order to persist (that is, to exist, to survive) is precisely what brings *active* forms of subjectivity about. The very notion of selfhood is justified by this existential imperative. It is the origin and purpose of mind. (**Solms**)

11. As these living subjects evolve to survive and reproduce in accordance with the laws of natural and (later) sexual selection, any changes that occur in their makeup will induce chemical forces. Those that lead towards better survival and reproductive chances are objectively good for the subject and take on the affective valence of pleasure. The opposite is objectively bad and painful. Homeostasis is a comfortable stable state in between. These states of affective valence are fundamental components of consciousness.

- Consciousness is fundamentally *affective* (see Panksepp, 1998; Solms, 2013; Damasio, 2018). The arousal processes that produce what is conventionally called "wakefulness" constitute the experiencing subject. In other words, *the experiencing subject is constituted by affect*. (**Solms**)
- We have seen that minds emerge in consequence of the existential imperative of self-organizing systems to monitor their own internal states in relation to potentially annihilatory, entropic forces. Such monitoring is an inherently value-laden process. It is predicated upon the biological ethic (which underwrites the whole of evolution) to the effect that survival is "good." (**Solms**)
- Valence / value evolved much earlier. Even bacteria can go toward food and away from danger. (**Post 10**)
- The brainstem structures that generate conscious "state" are not only responsible for the degree but also for the core quality of subjective being. The primal conscious "state" of mammals is intrinsically affective. It is this realization that will revolutionize consciousness studies in future years. (**Solms and Panksepp**)
- Homeostasis is the primary mechanism driving life. Emotions are chemical reactions. The emotive response triggered by sensory stimuli are the qualia of philosophical tradition. This subjectivity is the critical enabler of consciousness. (**Post 10**)

- Affective qualia are accordingly claimed to work like this: deviation away from a homeostatic settling point (increasing uncertainty) is felt as unpleasure, and returning toward it (decreasing uncertainty) is felt as pleasure. There are many types (or "flavours") of pleasure and unpleasure in the brain (Panksepp, 1998). (**Solms**)
- Interoceptive consciousness is phenomenal; it "feels like" something. Above all, the phenomenal states of the body-as-subject are experienced affectively. Affects, rather than representing discrete external events, are experienced as positively and negatively valenced states. Their valence is determined by how changing internal conditions relate to the probability of survival and reproductive success. The empirical evidence for the feeling component are simply based on the highly replicable fact that wherever in the brain one can artificially evoke coherent emotional response patterns with deep brain stimulation, those shifting states uniformly are accompanied by "rewarding" and "punishing" states of mind. By attributing valence to experience—determining whether something is "good" or "bad" for the subject, within a biological system of values—affective consciousness (and the behaviours it gives rise to) intrinsically promotes survival and reproductive success. This is what consciousness is for. (**Solms and Panksepp**)
- The dumb id, in short, knows more than it can admit. Small wonder, therefore, that it is so regularly overlooked in contemporary cognitive science. But the id, unlike the ego, is only dumb in the glossopharyngeal sense. It constitutes the primary stuff from which minds are made; and cognitive science ignores it at its peril. We may safely say, without fear of contradiction, that were it not for the constant presence of affective feeling, conscious perceiving and thinking would either not exist or would gradually decay. This is just as well, because a mind unmotivated (and unguided) by feelings would be a hapless zombie, incapable of managing the basic tasks of life. (**Solms and Panksepp**)
- An explanation of experience will never be found in the function of vision—or memory, for that matter—or in any function that is not inherently experiential. The function of experience cannot be inferred from perception and memory, but it *can* be inferred from feeling. There is not necessarily "something it is like" to perceive and to learn, but who ever heard of an unconscious feeling—a feeling that you cannot feel? If we want to identify a mechanism that explains the phenomena of consciousness (in both its psychological and physiological aspects) we must focus on the function of feeling—the technical term for which is "affect." (**Solms**)

12. Affective valence can only be felt by the subject experiencing the physical and chemical changes. There is, therefore, a barrier to knowing "what it is like" to be another subject. However, affect will eventually lead to distinctive behaviour in complex animals that can be objectively observed.

- Behavioural criteria showing an animal has affective consciousness (likes and dislikes):
  1. Global operant conditioning (involving whole body and learning brand-new behaviours)
  2. Behavioural trade-offs, value-based cost-benefit decisions
  3. Frustration behaviour
  4. Self-delivery of pain relievers or rewards
  5. Approach to reinforcing drugs or conditioned place preference (**Feinberg and Mallatt**)

13. These affects become separable into three categories and seven basic emotions.

- Subcortical affective processes come in at least three major categorical forms: (a) the homeostatic internal bodily drives (such as hunger and thermoregulation); (b) the sensory affects, which help regulate those drives (such as the affective aspects of taste and feelings of coldness and warmth); and (c) the instinctual-emotional networks of the brain, which embody the action tools that ambulant organisms need to satisfy their affective drives in the outside world (such as searching for food and warmth). These instinctual "survival tools" include foraging for resources (SEEKING), reproductive eroticism (LUST), protection of the body (FEAR and RAGE), maternal devotion (CARE), separation distress (PANIC/GRIEF), and vigorous positive engagement with conspecifics (PLAY). (**Solms and Panksepp**)

14. Cognition is built alongside and on top of this affective valence to sense, remember, and know more and more about what is bad and good. This happens in an ever-evolving way, growing in time, space, and circles of concern.

- "Cognition is comprised of sensory and other information-processing mechanisms an organism has for becoming familiar with, valuing, and interacting productively with features of its environment in order to meet existential needs, the most basic of which are survival/persistence, growth/thriving, and reproduction." This specifies the adaptive value of cognition for an organism and has the additional virtue of differentiating cognition from metabolic functions such as respiration and photosynthesis, which arguably also employ mechanisms for acquiring, processing, and acting on information. … This proposed definition is consistent with Peter Sterling and Simon Laughlin's (2015) description in Principles of Neural Design of what brains do, including the human brain: "The brain's purposes reduce to regulating the internal milieu and helping the organism to survive and reproduce. All complex behaviour and mental experience—work and play, music and art, politics and prayer—are but strategies to accomplish these functions." (p. 11) (**Lyon**)
- What, then, does cortex contribute to consciousness? Although neocortex surely adds much to refined perceptual awareness, initial perceptual processing appears to be unconscious in itself (cf. blindsight) or it may have qualities that we do not readily recognize at the level of cognitive consciousness. … It is possible that perceptual and higher cognitive forms of consciousness emerged in the neocortex upon an evolutionary foundation of affective consciousness. (**Solms and Panksepp**)

15. Cognitive processing enables the interruption of affective reflexes in order to consider several things at once. Cognition thus gives stability to the fleeting nature of affective emotion. This stability allows for driven, intentional acts.

- It is argued here that cortex *stabilises* consciousness rather than generates it; i.e., that cortical functioning binds affective arousal, and thereby transforms it into conscious cognition. … The essential task of cognitive (cortical) consciousness is to *delay* motor responses to affective "demands made upon the mind for work." This delay enables thinking. The essential function of cortex is thus revealed to be stabilisation of non-declarative executive processes, which is the essence of what we call *working* memory. (**Solms**)
- The fundamental contribution of cortex to consciousness in this respect is stabilisation (and refinement) of the objects of perception and generating thinking and ideas. This contribution derives from the unrivalled capacity of cortex for representational forms of memory (in all of its varieties, both short- and long-term). To put it metaphorically, cortex

transforms the fleeting, fugitive, wave-like states of consciousness into mental solids. It generates objects. (Freud called them "object presentations".) Such stable representations, once established, can be innervated both externally and internally, thereby generating objects not only for perception but also for cognition. To be clear: the computations and memories underlying these representational processes are unconscious in themselves; but when consciousness is extended to them, it (consciousness) is transformed by them into something stable, something that can be thought, something in the nature of crystal clear perceptions that are transformed into ideas in working memory. (**Solms and Panksepp**)

16. Further cognition allows brains to become better reality simulators or prediction machines, which aid tremendously in prospects for survival.

- The evolutionary and developmental pressure to constrain incentive salience in perception through prediction-error coding (the "reality principle") places inhibitory constraints on action. The resultant inhibition requires tolerance of frustrated affects, but it secures more efficient drive satisfaction in the long run. (**Solms and Panksepp**)
- In this process, the organism must stay "ahead of the wave" of the biological consequences of its choices (to use the analogy that gave Andy Clark's (2016) book its wonderful title: *Surfing Uncertainty*): "To deal rapidly and fluently with an uncertain and noisy world, brains like ours have become masters of prediction—surfing the waves of noisy and ambiguous sensory stimulation by, in effect, trying to stay just ahead of the place where the wave is breaking (p. xiv)." (**Solms**)
- What I am claiming is something else: feeling enables complex organisms to register—and thereby to regulate and prioritize through thinking and voluntary action—deviations from homeostatic settling points *in unpredicted contexts*. This adaptation, in turn, underwrites learning from experience. In predictable situations, organisms may rely on automatized reflexive responses (in which case, the biologically viable predictions are made through natural selection and embodied in the phenotype; see Clark, 2016). But if the organism is going to make plausible *choices* in novel contexts (cf. "free will") it must do so via some type of here-and-now assessment of the relative *value* attaching to the alternatives (see Solms, 2014). (**Solms**)

17. The development and feelings of "precision" are an important part of how these predictions work.

- "Precision" is an extremely important aspect of active and perceptual inference; it is the *representation of uncertainty*. The precision attaching to a quantity estimates its reliability, or inverse variance (e.g., visual—relative to auditory—signals are afforded greater precision during daylight vs. night-time). Heuristically, precision can be regarded as the confidence afforded probabilistic beliefs about states of the not-system—or, more importantly, what actions "I should select." (**Solms**)
- The feeling of knowing ("I do know that") is a basic emotion like fear that is not under conscious control. (**Campbell**)

18. As cognitive predictions are tested, they take on valence where surprises and uncertainty are bad and therefore honed by evolution to improve. This cognitive valence is what philosophers seemingly refer to as qualia.

- Friston's model of the Bayesian brain (in terms of which prediction-error or "surprise", equated with "free energy") is minimized through the encoding of better models of the world leading to better predictions is therefore, in principle, entirely consistent with the

model outlined here. It is important to note that in this model, prediction error (mediated by the sensory affect of surprise), which increases incentive salience (and therefore conscious "presence" of the self) in perception, is a "bad" thing, biologically speaking. The more veridical the brain's generative model of the world, the less surprise (the less salience, the less consciousness, the more automaticity), the better. Freud called this the "Nirvana principle". [In simpler terms,] the goal of all learning is automatised mental processes, increased predictability, and reduced uncertainty or "surprise". (**Solms and Panksepp**)

- The inherently subjective and qualitative nature of this auto-assessment process explains "how and why" it feels like something to the organism, for the organism (cf. Nagel, 1974). Specifically, increasing uncertainty in relation to any biological imperative *just is* "bad" from the (first-person) perspective of such an organism—indeed it is an existential crisis—while decreasing uncertainty *just is* "good." (**Solms**)
- The proposal on offer here is that this imperative *predictive* function—which bestows the adaptive advantage of enabling organisms to survive in novel environments—is performed by feeling. On the present proposal, this is the causal contribution of qualia. (**Solms**)

19. Predictions about the intentions of others are particularly vital.

- I claim that our power to *interpret* the actions of others depends on our power—seldom explicitly exercised—to predict them. Where utter patternlessness or randomness prevails, nothing is predictable. The success of folk-psychological prediction, like the success of any prediction, depends on there being some order or pattern in the world to exploit. … Folk psychology provides a description system that permits highly reliable prediction of human (and much nonhuman) behaviour. (**Dennett**)
- Understanding of others' intentions is a critical precursor to understanding other minds because intentionality, or "aboutness", is a fundamental feature of mental states and events. The "intentional stance" has been defined by Daniel Dennett as an understanding that others' actions are goal-directed and arise from particular beliefs or desires. (**Theory of Mind Wikipedia**)
- Here is how it works: first you decide to treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do. (p.17) (**Dennett**)

20. By making cognitive connections between intentions, predictions, and internal affective feelings, the development of self-awareness slowly arises.

- [At first,] the external body is not a subject but an object, and it is perceived in the same register as other objects. Something has to be added to simple perception before one's own body is differentiated from others. This level of representation (a.k.a. higher-order thought) enables the subject of consciousness to separate itself as an object from other objects. We envisage the process involving three levels of experience: (a) the subjective or phenomenal level of the *anoetic* self as affect, a.k.a. first-person perspective; (b) the perceptual or representational level of the *noetic* self as an object, no different from other objects, a.k.a. second-person perspective; (c) the conceptual or re-representational level of the *autonoetic* self in relation to other objects, i.e., perceived from an external perspective,

a.k.a. third-person perspective. The self of everyday cognition is therefore largely an abstraction. That is why the self is so effortlessly able to think about itself in relation to objects, in such everyday situations as "I am currently experiencing myself looking at an object". (**Solms and Panksepp**)

21. Models of others and the self are made using the same mechanisms.

- In *The Ancient Origins of Consciousness*, Feinberg and Mallatt contend that consciousness is about creating image maps of the environment and oneself. But systems that do it with orders of magnitude less sophistication than humans can still trigger our intuition of a fellow conscious being. (**Post 11**)
- External body representation is made of the same "stuff" as the representation of other objects. The external bodily "self" is represented as a thing—"my body"—and is inscribed on the page of consciousness in much the same way as other objects. It is, in short, an external, stabilized, detailed representation of the subject of consciousness. It is not the subject itself. The subject of consciousness identifies itself with this external bodily representation in much the same way as a child might project itself into the animated figures that she controls in a computer game. The representations rapidly come to be treated as if they were the self, but in reality they are not. Here is some experimental evidence for the counterintuitive relation between the self and its external body. Petkova and Ehrsson reported a series of "body swap" experiments in which cameras mounted over the eyes of other people or mannequins, transmitting images from their viewpoint to goggles mounted over the eyes of experimental subjects, rapidly created the illusion in the experimental subjects that the other body or mannequin was their own body. This illusion was so compelling that it persisted even when the subjects (projected into the other bodies) shook hands with their own bodies. The existence of this illusion was demonstrated objectively by the fact that when the other (illusory own) body and one's own body were both threatened with a knife, the emotional reaction (measured by heart rate and galvanic skin response) was greater for the illusory body. The well-known "rubber hand illusion" demonstrates the same relation between the self and the external body, albeit less dramatically. So does the inverse "phantom limb" phenomenon. (**Solms and Panksepp**)

22. Studies have shown that conscious awareness is necessary for some types of learning that give organisms additional plasticity to respond to new and novel stimuli in their environment.

- Our pain and sex lives might be regulated by unconscious information, but organisms need to learn. It is this that consciousness is for. It confers, like nothing else could, plasticity. Innate responses to basic evolutionarily advantageous or disadvantageous things might get us to mate or avoid standard bad things, but they wouldn't get us to learn about the contingent features of our environment on which rests our ultimate success. (**Flanagan and Polger**) (*Note: F&P don't actually support this argument. They say "This argument won't work. Plasticity, learning, and the like need not be, indeed in our own case they often are not, conscious." However, the following studies show they are wrong for some types of learning.*)
- Robert Clark and Larry Squire published the results of a classical Pavlovian conditioning experiment in humans. Two different test conditions were employed both using the eye-blink response to an air puff applied to the eye but with different temporal intervals between the air puff and a preceding, predictive stimulus (a tone): in one condition the tone remained on until the air puff was presented and both coterminated ("delay conditioning"); in the other a delay (500 or 1000 ms) was used between the offset of the

tone and the onset of the air puff ("trace conditioning"). In both conditions experimental subjects were watching a silent movie while the stimuli were applied and questions regarding the contents of the silent movie and test conditions were asked after test completion. In the delay conditioning task, subjects acquired a conditioned response over 6 blocks of 20 trials: as soon as the tone appeared they showed the eye-blink response before the air puff arrived. This is a classical Pavlovian response in which a shift is noted from reaction to action, also known as specific anticipatory behaviour. This shift occurred whether subjects had knowledge of the temporal relationship between tone and air puff or not: both subjects who were aware of the temporal relationship — as judged by their answers to questions regarding this relationship after test completion — and subjects who were unaware of the relationship learned this experimental task. One could say that this type of conditioning occurs automatically, reflex-like, or implicitly. In contrast, the trace conditioning task required that the subjects explicitly knew or realized the temporal relationship between the tone and air puff. Only those subjects knowing this relationship explicitly — as judged by their answers to questions regarding this relationship — succeeded in performing the task; those that were not, failed. In other words, subjects had to be explicitly aware or have conscious knowledge of the task at hand in order to bring the shift about, that is, to respond after the tone and before the air puff. This is called explicit or declarative knowledge. … Clark and Squire (1998, p.79) suggested that "awareness is a prerequisite for successful trace conditioning": (i) when explicitly briefed before trace conditioning about the temporal relationship between tone and air puff, all subjects learned the task, and faster than those without briefing; (ii) when performing an attention-demanding task, subjects did not acquire trace conditioning. (**van den Bos**)

23. An awareness of internal "emotions" allow us to learn from "feelings".

- Feelings are mental experiences that are the conscious experience of emotions. (**Post 10**)
- Learning arises from associations between interoceptive drives and exteroceptive representations, guided by the feelings generated by the affective experiences aroused by those representations. This is why they become conscious; the embodied subject must evaluate them. (**Solms and Panksepp**)
- As the cognitive science of the late twentieth century is complemented by the affective neuroscience of the present, we are breaking through to a truly mental neuroscience, and finally understanding that the brain is not merely an information-processing device but also a sentient, intentional being. Our animal behaviours are not "just" behaviours; in their primal affective forms they embody ancient mental processes that we share, at the very least, with all other mammals. (**Solms and Panksepp**)

24. This learning can be repeated over and over in order to rebuild memories and learn new things from them in light of further information.

- The reversal of the memory consolidation process (*reconsolidation*; Nader et al., 2000) renders Long Term Memory-traces labile, through literal dissolution of the proteins that initially "wired" them (Hebb, 1949). This iterative feeling and re-feeling one's way through declarable problems is the function of the cognitive qualia which have so dominated contemporary consciousness studies. (**Solms**)

25. Conscious awareness of what is going on in our own minds goes hand in hand with developing awareness that others have minds.

- The study of which animals are capable of attributing knowledge and mental states to others, as well as the development of this ability in human ontogeny and phylogeny, has identified several behavioural precursors to theory of mind. Understanding attention, understanding of others' intentions, and imitative experience with other people are hallmarks of a theory of mind that may be observed early in the development of what later becomes a full-fledged theory. (**Theory of Mind Wikipedia**)
- Selfhood is impossible unless a self-organizing system monitors its internal state in relation to not-self dissipative forces. The self can only exist in contradistinction to the not-self. This ultimately gives rise to the philosophical problem of other minds. In fact, the properties of a Markov blanket *explain* the problem of other minds: the internal states of a self-organizing system can only ever register hidden external (not-system) states vicariously, via the sensory states of their own blanket. (**Solms**)

26. Once living organisms become aware of selves and others, simple forms of communication such as pointing develop.

- Joint attention refers to when two people look at and attend to the same thing; parents often use the act of pointing to prompt infants to engage in joint attention. The inclination to spontaneously reference an object in the world as of interest, via pointing, and to likewise appreciate the directed attention of another, may be the underlying motive behind all human communication. (**Theory of Mind Wikipedia**)

27. While sensory memory and pointing are enough for the self and for rudimentary communication, the development of language through abstract symbols allows for much greater scale and scope in cognition.

- On the "self-awareness being tied to language" note, I found this quote from Helen Keller interesting: "Before my teacher came to me, I did not know that I am. I lived in a world that was a no-world. I cannot hope to describe adequately that unconscious, yet conscious time of nothingness. (…) Since I had no power of thought, I did not compare one mental state with another." Hellen Keller, 1908: quoted by Daniel Dennett, 1991, Consciousness Explained. London, The Penguin Press. pg 227 (**Hiskey**)
- Interestingly, deafness is significantly more serious than blindness in terms of the effect it can have on the brain. This isn't because deaf people's brains are different than hearing people, in terms of mental capacity or the like; rather, it is because of how integral language is to how our brain functions. To be clear, "language" here not only refers to spoken languages, but also to sign language. It is simply important that the brain have some form of language it can fully comprehend and can turn into an inner voice to drive thought. (**Hiskey**)
- Recent research has shown that language is integral in such brain functions as memory, abstract thinking, and, fascinatingly, self-awareness. Language has been shown to literally be the "device driver", so to speak, that drives much of the brain's core "hardware". Thus, deaf people who aren't identified as such very young or that live in places where they aren't able to be taught sign language, will be significantly handicapped mentally until they learn a structured language, even though there is nothing actually wrong with their brains. The problem is even more severe than it may appear at first because of how important language is to the early stages of development of the brain. Those completely deaf people who are taught no sign language until later life will often have learning problems that stick with them throughout their lives, even after they have eventually learned a particular sign language. (**Hiskey**)

- Today I found out how deaf people think in terms of their "inner voice". It turns out, this varies somewhat from deaf person to deaf person, depending on their level of deafness and vocal training. Those who were born completely deaf and only learned sign language will, not surprisingly, think in sign language. What is surprising is those who were born completely deaf but learn to speak through vocal training will occasionally think not only in the particular sign language that they know, but also will sometimes think in the vocal language they learned, with their brains coming up with how the vocal language sounds. Primarily though, most completely deaf people think in sign language. Similar to how an "inner voice" of a hearing person is experienced in one's own voice, a completely deaf person sees or, more aptly, feels themselves signing in their head as they "talk" in their heads. (**Hiskey**)
- Interestingly, if you take a deaf person and make them grip something hard with their hands while asking them to memorize a list of words, this has the same disruptive effect as making a hearing person repeat some nonsense phrase such as "Bob and Bill" during memorization tasks. (**Hiskey**)

28. Language increases our ability to make sense of the world compared to working memory alone.

- Feeling only persists (is only required) for as long as the cognitive task at hand remains unresolved. Conscious cognitive capacity is an extremely limited resource (cf. Miller's law) which must be used sparingly. [Miller's law states that human beings are capable of holding seven-plus-or-minus-two units of information in working memory at any one point in time.] (**Solms**)
- Only consciousness allows us to entertain lasting thoughts. It also allows us to create algorithms, a step-by-step way of solving a problem. It allows for flexible routing of information and appears to be necessary for making a final decision. Consciousness is an important element of social information sharing. It condenses information, [making it easier to transfer]. (**Post 9**)
- If I ask you to picture a rope and climbing up it, you can do it. I specifically chose those objects and actions because it is exactly what a chimp in a zoo is familiar with. If I asked a chimp to do the same thing, could it? We don't know, but I suspect not, because you can't do it wordlessly. You need to be able to interact using language. Without language, I don't think you have the cognitive systems for self-simulation and self-probing that we have. … Language allows us to be conscious of things we otherwise wouldn't be able to be conscious of. (**Post 7**)

29. Language also vastly enlarges the recognition of patterns in the world, which is a vital part of our prediction abilities.

- Differences in knowledge yield striking differences in the capacity to pick up patterns. Expert chess players can instantly perceive (and subsequently recall with high accuracy) the total board position in a real game but are much worse at recall if the same chess pieces are randomly placed on the board, even though to a novice both boards are equally hard to recall. This should not surprise anyone who considers that an expert speaker of English would have much less difficulty perceiving and recalling: "The frightened cat struggled to get loose" than "Te serioghehnde t srugfcalde go tgtt ohle" which contains the same pieces, now somewhat disordered. Expert chess players, unlike novices, not only know how to *Play* chess; they know how to *read* chess—how to see the patterns at a glance. (**Dennett**)

30. Language enables deep and precise probing of the self.

- A particular human experience is where you know the experience is happening to you. We can't rule that out in other animals, but neurological evidence suggests that it's not happening. This "autonoetic consciousness" represents the view of the self as the subject. It enables mental time-travel (i.e. you can review past experiences and possible future states). Other animals can learn from the past, but in a simple way. (**Post 12**)

31. Language enables many more degrees of freedom. We may not have ultimately free will, but Libet's attempt to deny it is a misunderstanding of the difference between the core affective self and the represented self of cognition.

- Degrees of freedom is something I'm using more lately. It is an opportunity for control. Degrees of freedom can be clamped or locked down to be removed. How many degrees of freedom do humans have? Millions and millions of things we can think of. We have orders of magnitude more that we can think of than a bear does, even with roughly the same number of cells. So, our complexity is higher. The options a bear has are a vanishing subset of the options that we have. Learning to control these options is not now a science. It is an art. (**Dennett**)
- Whereas homeostasis requires nothing more than ongoing adjustment of the system's active states (M) and/or inferences about its sensory states (ϕ), in accordance with its predictive model (ψ) of the external world (Q) or vegetative body ($Q\eta$), which can be adjusted automatically on the basis of ongoing registrations of prediction error (e), quantified as free energy (F)—contextual considerations require an additional capacity to adjust the precision weighting (ω) of all relevant quantities. This capacity provides a formal (mechanistic) account of voluntary behaviour—of choice. (**Solms**)
- The unrecognized gap between the primary subjective self and the re-representational abstracted self causes much confusion. Witness the famous example of Benjamin Libet recording a delay of up to 400 ms between the physiological appearance of premotor activation and the voluntary decision to move. This is typically interpreted to mean that free will is an illusion, when in fact it shows only that reflexive re-representation of the self initiating a movement occurs somewhat later than the core self actually initiating it. (**Solms and Panksepp**)

32. Finally, language and the autobiographical self leads to all of the items of human culture.

- Autobiographical self has prompted: extended memory, reasoning, imagination, creativity, and language. Out of these came the instruments of culture: religions, justice, trade, the arts, science, and technology. (**Post 10**)

33. Bringing all of these aspects of consciousness together requires a multi-faceted framework. But it would help if this framework was organised around a single unifying concept.

- As long as one avoids confusion by being clear about one's meanings, there is great value in having a variety of concepts by which we can access and grasp consciousness in all its rich complexity. However, one should not assume that conceptual plurality implies referential divergence. Our multiple concepts of consciousness may in fact pick out varying aspects of a single unified underlying mental phenomenon. Whether and to what

extent they do so remains an open question. (**Consciousness Entry in Stanford Encyclopedia**)

- The problem of consciousness will only be solved if we reduce its psychological and physiological manifestations to a single underlying abstraction. (**Solms**)

34. Before describing my own framework and unifying concepts, a quick review of some other contenders is helpful. The Stanford Encyclopedia of Philosophy lists six separate functions of consciousness.

- How do mental processes that involve the relevant sort of consciousness differ from those that lack it? What function(s) might consciousness play? The following six notions are some of the more commonly given answers: 1) Flexible control. Though unconscious automatic processes can be extremely efficient and rapid, they typically operate in ways that are more fixed and predetermined than those which involve conscious self-awareness. 2) Social coordination. Consciousness of the meta-mental sort may well involve not only an increase in self-awareness but also an enhanced understanding of the mental states of other minded creatures, especially those of other members of one's social group. 3) Integrated representation. Conscious experience presents us not with isolated properties or features but with objects and events situated in an ongoing independent world, and it does so by embodying in its experiential organisation and dynamics the dense network of relations and interconnections that collectively constitute the meaningful structure of a world of objects. 4) Informational access. The information carried in conscious mental states is typically available for use by a diversity of mental subsystems and for application to a wide range of potential situations and actions. 5) Freedom of will. Consciousness has been thought to open a realm of possibilities, a sphere of options within which the conscious self might choose or act freely. 6) Intrinsic motivation. The attractive positive motivational aspect of a pleasure seems a part of its directly experienced phenomenal feel, as does the negative affective character of a pain. (**Consciousness Entry in Stanford Encyclopedia**)

35. A simple distinction is sometimes made between primary and higher order consciousness.

- Another theory about the function of consciousness has been proposed by Gerald Edelman called dynamic core hypothesis which puts emphasis on re-entrant connections (bi-directional connections) that reciprocally link areas of the brain in a massively parallel manner. Edelman also stresses the importance of the evolutionary emergence of higher-order consciousness in humans from the historically older trait of primary consciousness which humans share with non-human animals. (**Consciousness Wikipedia**)
- Primary consciousness is broken down into three elements: 1) Exteroceptive—Damasio's mapping of the outer world. 2) Interoceptive—signals from inside the body. 3) Affective—the experience of feeling, emotion, or mood. (**Post 11**)
- *The Ancient Origins of Consciousness* does not address higher levels of consciousness: full-blown self-awareness, meta-awareness, recognition of the self in mirrors, theory of mind, access to verbal self-reporting. (**Post 11**)

36. Another widely discussed definition divides consciousness into three forms: anoetic, noetic, and autonoetic.

- In short, the complexity of our capacity to consciously and unconsciously process fluctuating brain states and environmentally linked behavioural processes requires some

kind of multi-tiered analysis, such as Endel Tulving's well-known parsing of consciousness into three forms: *anoetic* (unthinking forms of experience, which may be affectively intense without being "known", and could be the birthright of all mammals), *noetic* (thinking forms of consciousness, linked to exteroceptive perception and cognition), and *autonoetic* (abstracted forms of perceptions and cognitions, which allow conscious "awareness" and reflection upon experience in the "mind's eye" through episodic memories and fantasies). (**Solms and Panksepp**)

37. This is similar to Antonio Damasio's three selves.

- A mind emerges from the brain when an animal is able to create images and to map the world and its body. [According to Antonio Damasio's definition,] consciousness requires the addition of self-awareness. This begins at the level of the brain stem, with "primordial feelings." The self is built up in stages starting with the proto self made up of primordial feelings, affect alone, and feeling alive. Then the core self is developed when the proto self is interacting with objects and images such that they are modified and there is a narrative sequence. Finally comes the autobiographical self, which is built from the lived past and the anticipated future. (**Post 10**)

38. Feinberg and Mallat list six adaptive advantages of consciousness organised over three different levels.

- Adaptive advantages of consciousness: 1) It efficiently organizes much sensory input into a set of diverse qualia for action choice. As it organizes them, it resolves conflicts among the diverse inputs. 2) Its unified simulation of the complex environment directs behaviour in three-dimensional space. 3) Its importance-ranking of sensed stimuli, by assigned affects, makes decisions easier. 4) It allows flexible behaviour. It allows much and flexible learning. 5) It predicts the near future, allowing error correction. 6) It deals well with new situations. (**Feinberg and Mallatt**)
- The Defining Features of Consciousness are: Level 1) General Biological Features: life, embodiment, processes, self-organising systems, emergence, teleonomy, and adaption. Level 2) Reflexes of animals with nervous systems. Level 3) Special Neurobiological Features: complex hierarchy (of networks); nested and non-nested processes, aka recursive; isomorphic representations and mental images; affective states; attention; and memory. (**Post 11**)

39. The latest hierarchy from Mike Smith on his excellent Self Aware Patterns website (which devotes a lot of time to consciousness studies) has six layers.

1. Matter: a system that is part of the environment, is affected by it, and affects it. *Panpsychism*.
2. Reflexes and fixed action patterns: automatic reactions to stimuli. If we stipulate that these must be biologically adaptive, then this layer is equivalent to *universal biopsychism*.
3. Perception: models of the environment built from distance senses, increasing the scope of what the reflexes are reacting to.
4. Volition: selection of which reflexes to allow or inhibit based on learned predictions.
5. Deliberative imagination: sensory-action scenarios, episodic memory, to enhance 4.
6. Introspection: deep recursive metacognition enabling symbolic thought.

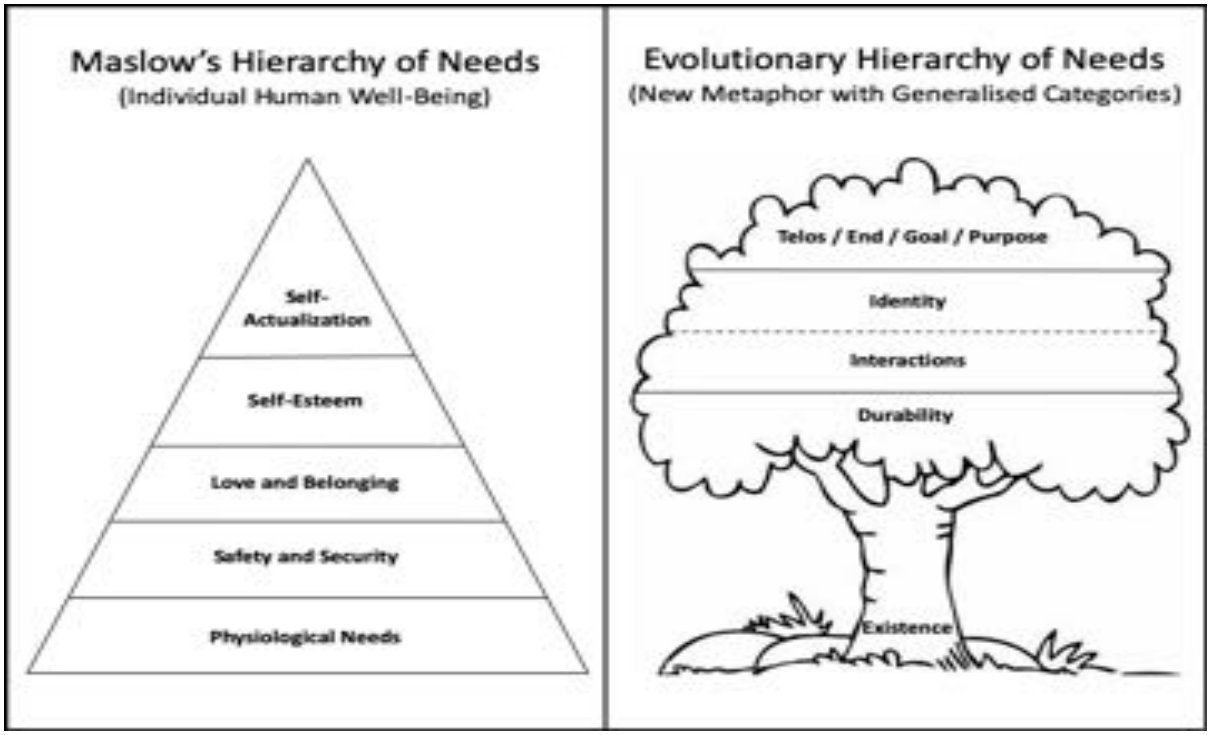
40. Lyon lists 13 functional abilities of cognition that help organisms adapt to

their environment.

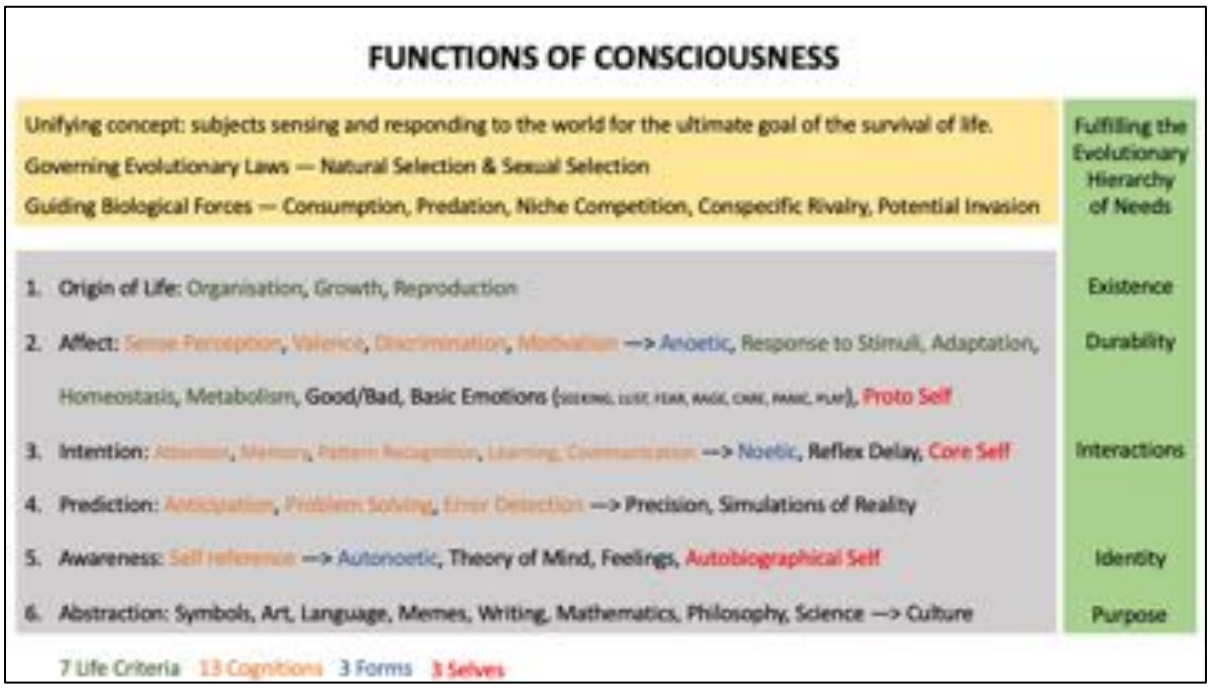
- The broadly biological conceptions of the capacities encompassed by the general concept of cognition are: (1) sense perception — ability to recognize existentially salient features of the external or internal milieu; (2) affect — valence: attraction, repulsion, neutrality / indifference (hedonic response); (3) discrimination — ability to determine that a state of affairs affords an existential opportunity or presents a challenge, requiring a change in internal state of behaviour; (4) memory — retention of information about a state of affairs for a non-zero period; (5) learning — experience-modulated behaviour change; (6) problem solving / decision making — behaviour selection in circumstances with multiple, potentially conflicting parameters and varying degrees of uncertainty; (7) communication — mechanism for initiating purposive interaction with conspecifics (or non-conspecific others) to fulfil an existentially salient goal; (8) motivation — teleonomic striving; implicit goals arising from existential conditions; (9) anticipation — behavioural change based on experience-based expectancy (i.e. if X is happening, then Y should happen), possibly evolved across generations, and which is implicit to the agent's functioning; (10) awareness — orienting response; ability to selectively attend to aspects of the external and/or internal milieu; (11) self-reference — mechanisms for distinguishing "self" or "like self" from "non-self" or "not like self"; (12) normativity — error detection, behavioural correction, value assignment based on motivational state; (13) intentionality — directedness towards an object. (**Lyon**)

41. These adaptations help meet the evolutionary hierarchy of needs of all life.

- The evolutionary perspective of our diverse and ever-changing web of life transforms Maslow's hierarchy. Starting at the bottom of the pyramid—or tree now—we see that the "physiological" needs of the human are merely the brute ingredients necessary for "existence" that any form of life might have. In order for that existence to survive through time, the second level needs for "safety and security" can be understood as promoting "durability" in living things. The third tier requirements for "love and belonging" are necessary outcomes from the unavoidable "interactions" that take place in our deeply interconnected biome of Earth. The "self-esteem" needs of individuals could be seen merely as ways for organisms to carve out a useful "identity" within the chaos of competition and cooperation that characterizes the struggle for survival. And finally, the "self-actualization" that Maslow struggled to define (and which Kenrick and Andrews discarded or subsumed elsewhere), could be seen as the end, goal, or purpose that an individual takes on so that they may (consciously or unconsciously) have an ultimate arbiter for the choices that have to be made during their lifetime. This is something Aristotle called "*telos*." Putting this all together, we may then change Maslow's hierarchical pyramid of human needs into the following multi-layered tree for any individual life. (**Gibney**)



42. Summarising all of this research, here is my proposal for a hierarchy of the functions of consciousness. They are unified under a single concept, governed by the evolutionary laws of selection, and guided by biological forces in order to meet the needs of life.



In my proposal, understanding consciousness begins with the unifying concept of "subjects sensing and responding to the world for the ultimate goal of the survival of life." This is in line with how I define consciousness (derived from the Latin for "knowing together"). The functions that evolve are governed by the laws of natural selection and (later) sexual selection. They are guided in this evolution by the biological forces that exerted on living beings: consumption, predation, niche competition, conspecific rivalry, and potential invasion. And

they help organisms meet their evolutionary hierarchy of needs.

Stepping through the hierarchy, we therefore start with the **origin of life**. Once the first three criteria from the general definition for life are happening—organisation, growth, and reproduction—subjects come into existence.

As soon as life emerges, the function of **affect** begins to take hold. Any changes to these living organisms cause chemical forces to be exerted on individual subjects. Changes that lead towards persistence are objectively defined as good. The opposite changes are bad. Stability is homeostasis. As these changes are selected for, the earliest forms of life become more complex, eventually meeting the rest of the criteria for the definition of life—response to stimuli, adaptation, homeostasis, and metabolism. These forms of life respond reflexively, using chemical emotional responses alone that develop (according to Panksepp) into seven basic emotions: foraging for resources (SEEKING), reproductive drive (LUST), protection of the body (FEAR and RAGE), maternal devotion (CARE), separation distress (PANIC), and vigorous positive engagement with conspecifics (PLAY). The first four of these have premammalian origins. In humans, the final three only date back to early primates. Among the 13 cognitive capacities that Lyon notes, 4 are required during this stage of affect: sense perception, valence, discrimination, and motivation. These can be said to produce the anoetic, proto self.

Over time, adaptations from affective reflexes alone lead to capacities for cognition that are able to interrupt these reflexes. From Lyon's list, the five capacities of attention, memory, pattern recognition, learning, and communication lead to this noetic, core self where organisms can be said to be acting with **intention**. Choices are made and to an outside observer there is a narrative sequence to life.

Once intentions exist (either one's own or the intentions of others), they can be taken into account. To do so is to use **prediction** to think through what the result will be from any intentions. This requires three more cognitive capacities from Lyon's list: anticipation, problem solving, and error detection. With these abilities, organisms can simulate reality and be led by emotions of precision to hone these simulations towards greater accuracy.

As predictions and perceptions improve, organisms eventually make the connection that there is a self which has its own mind. **Awareness** is achieved. This development is covered by the final cognitive capacity from Lyon's list: self-reference. Such conscious cognition allows memories and thoughts built from the lived past and the anticipated future to create the autonoetic, autobiographical self.

Finally, through the development of ideas about the self and other minds, brains began to imagine something that had no immediate impact on their senses. This opens up the doors for much further **abstraction**. Slowly, the evolution of symbols, art, and language took place, enabling certain abilities that perhaps only humans possess at this time. Memes, writing, mathematics, philosophy, and science make up and enable all of eventual the products of human culture.

So, there you have it. As I noted in my **brief history of the definitions of consciousness**, many, many attempts have been made at this. Maybe this is just another one. But I believe it is the kind of comprehensive definition that would allow others to draw circles around the items from their definitions and say, "that's what *I* think consciousness is." If I'm lucky, maybe they'll even switch to say, "that's what I thought consciousness *was*."

The hard problem of consciousness is often phrased as wondering how inert matter can ever evolve into the subjective experience that we humans undoubtedly feel. I think this short-changes matter. Far from being inert, matter responds to the forces exerted on it all the time. Panpsychism says mind (*psyche*) is everywhere. But to me there can be no mind without a stable subject. In my current conception, the forces that minds feel and are shaped by are merely the chemical and physical forces that shape all matter. Until something else is found, what else could there be? So, mind is not everywhere, but forces are. The Greek for force is *dynami*, so rather than panpsychism, I would say the universe has pandynamism. The psyche only originates and evolves along with life.

What about other forms of non-biological life? As I said in **post 17**, "Could artificial life also respond to these forces and be declared conscious? I think yes, although the "feeling of what it is like" to be such life would be very different from current biological life forms that are built from organic chemistry. We already believe the feeling of what is like to be a bat is likely very different from that of a cuttlefish, so the difference would be even greater for artificial life given the much larger change in underlying mechanisms. Yet both could be considered conscious in my definition."

And with that, I have my answer to the 1st of Tinbergen's 4 questions about the biological aspects of consciousness. Now that we have a clear list of the functions that consciousness enables, I'll try to match them up against the mechanisms that cause all of this. Stay tuned for that as the end of this series comes into view.

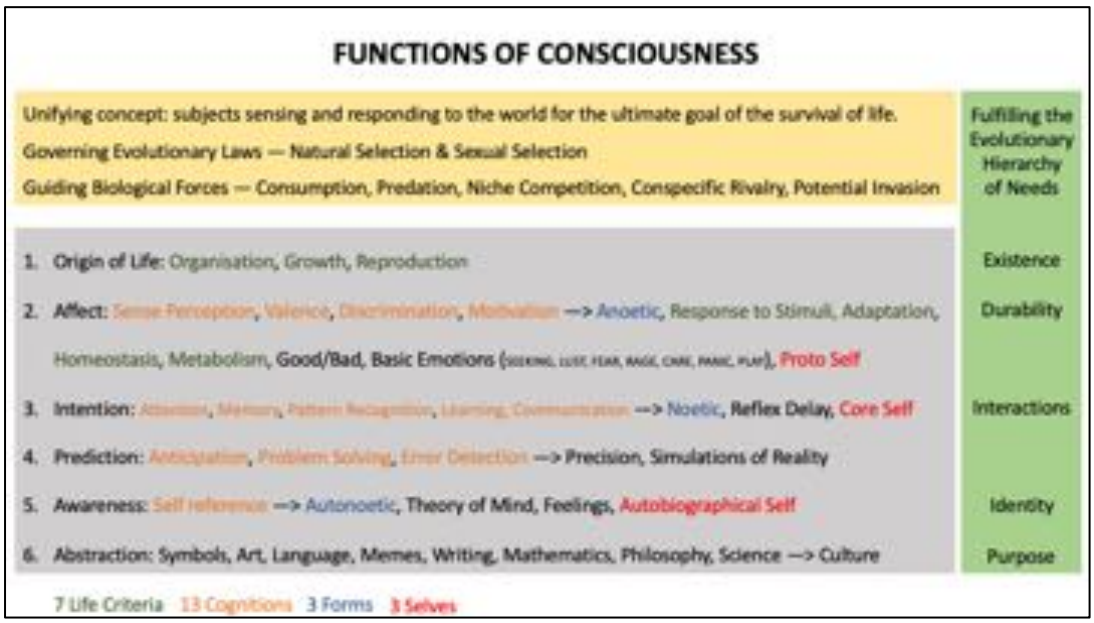
20 — The Mechanisms of Consciousness

23 September 2020

On to the next of **Tinbergen's four questions**!

As a quick reminder (since I am really dragging this series out now), I have previously defined consciousness very broadly. I posit that this still amorphous concept can best be understood as the set of processes where living organisms (governed by the various laws of selection) sense and respond to biological forces. This is currently only achieved by carbon-based life, but there's no reason that artificial life couldn't conceivably fulfil these criteria too.

I first described this definition in my **brief history of everything that has ever existed**. To fully grasp any biological phenomenon (which consciousness surely is in a natural view of the universe), Nikolaas Tinbergen developed a 2x2 matrix of things to consider, which has since become the standard in evolutionary studies. His four items are: 1) function (adaptation), 2) mechanism (causation), 3) ontogeny (development), and 4) phylogeny (evolution). In my last post, I covered the first of these items--**the functions of consciousness**—which led to the following hierarchical table:



Now that we are clear about all of the biological functions we are talking about in the multi-faceted and complex concept of consciousness, we can try to answer Tinbergen's next three questions by looking for the current mechanisms and the personal and evolutionary histories that are associated with these functions. Let's do the mechanisms first and see how they fit within the hierarchy shown above.

Just like last time, there are lot of intricate details for this large and unwieldy topic, so I'll write simple numbered statements followed by their justifications so you can quickly read the statements to get the gist of the argument, or you can dip into any of the details you might want for further information (with links there to even more). Unlike last time, however, we now have a structure to work within, so I'll try to abide by that. I should also note that it is impossible to cover the details of all of the mechanisms of all of the facets of consciousness for all of the living creatures that have ever experienced it. This post topped out at nearly 13,000 words when I included all of the details I had gathered to help me understand the mechanisms, but I have shortened it considerably (around 8,000 words now) to merely focus on the general gist of the types of mechanisms that are out there. I'm confident you can always find more from there on any specific mechanisms that interest you. Okay. Here we go!

1.0 Origin of Life (The first three criteria for life are: organisation, growth, and reproduction.)

1.01 Our current best guess for the origin of life involves lipid vesicles containing polymers that can grow and divide. The chemical structures describe the organisation. Osmotic pressure and new bonds caused growth. Mechanical forces split these growing vesicles, which led to reproduction and thus evolution.

- Let's review: Monomers diffuse into a fatty acid vesicle. Monomers spontaneously polymerize and copy any template. Heat separates strands and increases membrane permeability to monomers. Polymer backbones attract ions, increasing osmotic pressure. Pressure on the membrane drives its growth at the expense of nearby vesicles containing less polymer. Vesicles grow into tubular structures. Mechanical forces cause vesicles to divide. Daughter vesicles inherit polymers from the parent vesicle. Polymer sequences that replicate faster will dominate the population. Thus beginning evolution! (**Post 17**)

1.02 These are the specific forms of early chemical life, but what defines them can be generalised into abstract terms. The border between any self and non-self has thus been defined as a *Markov blanket*. Markov blankets have three characteristics—a physical boundary, sensory systems on the boundary, and internal mechanisms that enable the self to exist and persist.

- For a system to resist entropy, three conditions must be met: (i) There must be a boundary which separates the internal and external states of the system, and thereby insulates the system from the world. Let's call the former states "the system" and the latter states "the not-system." (ii) There must be a mechanism which registers the influence of dissipative external forces—i.e. the free energy. Let's call this mechanism the "sensory states" of the system. (iii) There must be a mechanism which counteracts these dissipative forces—i.e. which binds the free energy. Let's call this mechanism the "active states" of the system, such as motor and autonomic reflexes. … According to Friston (2013), these functional conditions—which enable self-organizing systems to exist and persist over time—emerge

naturally (indeed necessarily) within any ergodic random dynamical system that possesses a *Markov blanket*. This blanket establishes the boundary conditions above. (**Solms**)

2.0 Affect (The first four cognitive abilities—response to stimuli, adaptation, homeostasis, and metabolism—enable the fulfilment of the final four criteria for life: sense perception, valence, discrimination, and motivation.)

2.01 For the earliest forms of biological life, changes to their molecular structures exert forces on those molecules. Being surrounded by stronger or weaker osmotic forces defines the presence of others. Identical osmotic forces indicate groups of the same entities. Growth is good for survival. Loss is bad for it. This is how the first subjects begin to sense change in their environments, assign valence to these changes, and discriminate between selves and not-selves.

- Molecules are held together by covalent bonds, which involve the sharing of electron pairs between atoms. … Intermolecular forces are the forces which mediate interactions between molecules and other types of neighbouring particles such as atoms or ions. … The four key intermolecular forces are: 1) Ionic bonds; 2) Hydrogen bonding; 3) Van der Waals dipole-dipole interactions; and 4) Van der Waals dispersion forces. (**Post 17**)
- Any changes to biological molecules would generate forces. These forces are exerted on singularly identifiable objects. (**Post 19**)
- I propose that these chemical forces, once in service of biological needs, are the defined starting point for turning objects into subjects. (**Post 19**)

2.02 Changes in the organisational structure of living molecules are due to the action potentials of ion flows across cell membranes. A few kinds of ions have evolved to be especially important in life's current biochemistry for controlling basic stimulus-response mechanisms. These may be able to be mimicked in artificial life, but it would take much more diversity than is often considered.

- Membrane potential is the difference in electric potential between the interior and the exterior of a biological cell. All animal cells are surrounded by a membrane composed of a lipid bilayer with proteins embedded in it. The membrane serves as both an insulator and a diffusion barrier to the movement of ions. Transmembrane proteins, also known as ion transporter or ion pump proteins, actively push ions across the membrane and establish concentration gradients across the membrane, and ion channels allow ions to move across the membrane down those concentration gradients. Ion pumps and ion channels are electrically equivalent to a set of batteries and resistors inserted in the membrane, and therefore create a voltage between the two sides of the membrane. The membrane potential has two basic functions. First, it allows a cell to function as a battery, providing power to operate a variety of "molecular devices" embedded in the membrane. Second, in electrically excitable cells such as neurons and muscle cells, it is used for transmitting signals between different parts of a cell. (**Membrane potential**)
- In physiology, an action potential occurs when the membrane potential of a specific cell location rapidly rises and falls: this depolarization then causes adjacent locations to similarly depolarize. Action potentials occur in some plant cells and in several types of animal cells, called excitable cells, which include neurons, muscle cells, endocrine cells, and glomus cells. (**Action potential**)
- Voltage-gated-calcium-channels (VGCCs) are normally closed. They are activated (i.e., opened) at depolarized membrane potentials. The concentration of calcium (Ca2+ ions)

is normally several thousand times higher outside the cell than inside. Activation of particular VGCCs allows a Ca2+ influx into the cell, which, depending on the cell type, results in activation of calcium-sensitive potassium channels, muscular contraction, excitation of neurons, up-regulation of gene expression, or release of hormones or neurotransmitters. (**Voltage-gated Calcium Channel**)

- Voltage-gated sodium channels play an important role in action potentials. If enough channels open when there is a change in the cell's membrane potential, a small but significant number of Na+ ions will move into the cell down their electrochemical gradient, further depolarizing the cell. Thus, the more Na+ channels localized in a region of a cell's membrane the faster the action potential will propagate and the more excitable that area of the cell will be. The ability of these channels to assume a closed-inactivated state causes the refractory period and is critical for the propagation of action potentials down an axon. (**Sodium Channel**)
- A refractory period is a period of time during which an organ or cell is incapable of repeating a particular action, or (more precisely) the amount of time it takes for an excitable membrane to be ready for a second stimulus once it returns to its resting state following an excitation. It most commonly refers to electrically excitable muscle cells or neurons. (**Refractory period**)
- There are about 100 billion neurons in the human brain, each of which forms synapses with many other neurons. A synapse is the gap between two neurons (known as the presynaptic and postsynaptic neurons). The presynaptic neuron releases neurotransmitters, such as glutamate and GABA, which bind to receptors on the postsynaptic cell membrane, activating ion channels. Opening and closing those channels changes the cell's electrical potential. If the potential changes dramatically enough, the cell fires an electrical impulse called an action potential. (**Mimicking the Brain in Silicon**)
- MIT researchers designed a computer chip so that the transistors could mimic the activity of different ion channels. While most chips operate in a binary, on/off mode, current flows through the transistors on the new brain chip in analogue, not digital, fashion. A gradient of electrical potential drives current to flow through the transistors just as ions flow through ion channels in a cell. (**Mimicking the Brain in Silicon**)
- This synapse diversity could have implications for the prospect of creating artificial consciousness. (Not intelligence, but consciousness.) No computer has synapse diversity. I have met numerous people who are experts in building computers based on neural principles. As interesting as I find their presentations, they typically have built them on principles that are several decades out of date. There's no concept of the molecular organisation. It's based on a few electrophysiological parameters that are known about neurons. which comes from the era of cells and electrophysiology. Those are sort of pre-1990's stuff. (**Seth Grant**)

2.03 Chemical building blocks provide the ability to process information, which enables the repeatable decisions (cognition) necessary to remain alive.

- A candidate mechanism that may serve as the biological basis of the continuum of cognitive function [is] the chemistry of protein networks, whose potential information-processing power and similarity to neural networks in single cells was first described by Cambridge zoologist Dennis Bray, who noticed that "many proteins in living cells appear to have as their primary function the transfer and processing of information, rather than the chemical transformation of metabolic intermediates or the building of cellular structure." (**Lyon**)

- Protein signal transduction networks should be considered the basis of cognitive function. Neurons and the electrical properties of neurons come first to mind in discussions of the brain, but chemical protein networks are also widely used in that organ because, energetically, they are the cheapest way of sending and receiving information. (**Lyon**)
- In biological systems, information is transmitted "whenever a source's change in state registers as a change in state at a receiver." The change may be in environmental pH; the availability of nutrients; chemical indicators of conspecifics, predators, or prey; build-up of reactive oxygen species; osmolarity; diffusion potential, and so on. (**Lyon**)

2.04 In abstract terminology, these particular chemical responses to physical stimuli are the processing of information within Markov Blankets. They are systematically selected for by the natural processes of evolution. Logically, the systems that survive are those that enable survival. The resulting stability is called homeostasis, and the chemical processes that sustain life are its metabolism.

- For self-organizing systems—including all living things, like us—to exist, they must *resist entropy*. That is, self-organizing systems can only persist over time by occupying "preferred" states—as opposed to being dispersed over all possible states, and thereby dissipating. (**Solms**)
- A Markov blanket can only "know" states of the not-system *vicariously*. In other words, external states can only be "inferred" by the system on the basis of "sensory impressions" upon the Markov blanket. (**Solms**)
- In summary, homeostasis is explained by the causal dynamics mandated by the very existence of Markov blankets; in terms of which self-organizing systems generate a type of work that binds free energy and maintains the system in its typically occupied ("preferred" or "valued") states. (**Solms**)
- Metabolism is the set of life-sustaining chemical reactions in organisms. The three main purposes of metabolism are: the conversion of food to energy to run cellular processes; the conversion of food/fuel to building blocks for proteins, lipids, nucleic acids, and some carbohydrates; and the elimination of metabolic wastes. These enzyme-catalyzed reactions allow organisms to grow and reproduce, maintain their structures, and respond to their environments. (**Metabolism**)

2.05 Deviations from homeostasis cause internal reactions that are selected to bring systems back to their preferred state. These various reactions are the affective core of consciousness. These reactions diverged over the course of evolution into distinct facets that are recognisable as the seven basic emotions.

- [There] are various instinctual motivational circuits. These are also known as the circuits for "basic emotion". There are several classifications of these emotions. The best-known examples are those that generate (1) appetitive foraging, (2) consummatory reward, (3) freezing and flight, (4) aggressive attack, (5) nurturant care, (6) separation distress, and (7) rough-and-tumble play. It is important to note that each of the instinctual circuits generates not only stereotyped behaviours but also diverse feeling states, such as expectant interest, orgasmic delight, trepidatious fear, destructive rage, loving affection, sorrowful grief, and exuberant joy. The circuits for these basic emotions are conserved across the mammalian series, and they admit of considerable chemical specificity. They are no less innate than the vital evolutionary survival and sexual needs which gave rise to them. They are unconditioned "tools for living". (**Solms and Panksepp**)

2.06 Note that the conscious awareness and processing of affective reactions comes later in the hierarchy of consciousness. Such cognition only rides on the affective core and cannot exist in biological systems on its own.

- The removal of the neocortex has long been known to spare emotionality. Indeed, not only are the rewarding effects of subcortical brain stimulations demonstrably preserved in decorticated creatures, these animals are actually more emotional than normal. The most strikingly concordant human evidence to emerge in recent years, relevant to this broader question, concerns a condition called hydranencephaly, in which the cerebral cortex as a whole is destroyed in utero. However, the subcortical networks are functional; thus, the children are markedly emotionally functional human beings. "They express pleasure by smiling and laughter, and aversion by 'fussing' arching of the back and crying (in many gradations), their faces being animated by these emotional states. A familiar adult can employ this responsiveness to build up play sequences predictably progressing from smiling, through giggling, to laughter and great excitement on the part of the child." They also show basic emotional learning. Although there is in these children significant degradation of the types of consciousness that are normally associated with external perception, there can be no doubt that they are conscious, both quantitatively and qualitatively. (**Solms and Panksepp**)
- Let us consider a third problem with the cortico-centric approach. The third problem is that there is a brain structure which *does* pass the critical test just mentioned. This structure is located not in the cortex but the brainstem. Consciousness is obliterated by focal lesions of the brainstem core—in a region conventionally described as the extended reticulothalamic activating system (ERTAS). Recent findings indicate that the smallest lesions within the brainstem which cause total loss of consciousness (i.e., coma) are located in or near the parabrachial nuclei of the pons. (**Solms**)
- Although many cognitive scientists still must be weaned of the view that the cerebral cortex is the seat of consciousness, the weight of evidence for the alternative view that the arousal processes generated in the upper brainstem and limbic system feel like something in and of themselves, is now overwhelming. (**Solms**)
- This conclusion is further supported by the fact that drugs acting on the neuromodulators sourced in the ERTAS nuclei (serotonin, dopamine, noradrenaline, acetylcholine) have powerful effects on mood and anxiety, etc.—which is why they represent the mainstay of psychopharmacology today. (**Solms**)

3.0 Intention (Five more cognitive abilities—attention, memory, pattern recognition, learning, and communication—enable intentional actions of the core self, eventually including the delay of reflexes.)

3.01 Core affect cannot be responded to by internal reactions alone—externally observable behavioural responses must occur as well. Once they do, it can be said that these organisms act with intention. They become driven.

- The dynamics of a Markov blanket generate two fundamental properties—namely (elemental forms of) *selfhood* and *intentionality*. [T]hese dynamics also generate elemental properties of bodies—namely an *insulating membrane* and *adaptive behaviour*. (**Solms**)
- Body-monitoring nuclei…can only go so far in terms of meeting endogenous needs through internal (autonomic) adjustments. Beyond that limit, *external* action is called for. At that point, autonomic reflexes become *drives*. [For example], interoceptive "need

detectors" trigger not only autonomic reflexes but also feelings of hunger, thirst, etc. These drives typically trigger "foraging" behaviours ("SEEKING"). (**Solms**)

3.02 Further evolution along this path produces further mechanisms for producing intentions. Once multicellular forms of life evolve sufficient complexity, hormones act for intercellular communication, which allows whole organisms to respond with intent towards naturally selected goals.

- At the most basic level, the function of the nervous system is to send signals from one cell to others, or from one part of the body to others. There are multiple ways that a cell can send signals to other cells. One is by releasing chemicals called hormones into the internal circulation, so that they can diffuse to distant sites. (**Nervous Systems**)
- A hormone is any member of a class of signalling molecules, produced by glands in multicellular organisms, that are transported by the circulatory system to target distant organs to regulate physiology and behaviour. The glands that secrete hormones comprise the endocrine signalling system. Hormones affect distant cells by binding to specific receptor proteins in the target cell, resulting in a change in cell function. (**Hormones**)
- Hormones serve to communicate between organs and tissues for physiological regulation and behavioural activities such as digestion, metabolism, respiration, tissue function, sensory perception, sleep, excretion, lactation, stress induction, growth and development, movement, reproduction, and mood manipulation. (**Hormones**)

3.03 This evolution continues on to produce neurons in some forms of life. These are incredibly diverse and provide an enormous amount of additional abilities, including internally generated actions.

- In contrast to the "broadcast" mode of hormone signalling, the nervous system provides "point-to-point" signals—neurons project their axons to specific target areas and make synaptic connections with specific target cells. Thus, neural signalling is capable of a much higher level of specificity than hormonal signalling. It is also much faster: the fastest nerve signals travel at speeds that exceed 100 meters per second. (**Nervous Systems**)
- A cell that receives a synaptic signal from a neuron may be excited, inhibited, or otherwise modulated. (**Nervous Systems**)
- Even in the nervous system of a single species such as humans, hundreds of different types of neurons exist, with a wide variety of morphologies and functions. These include sensory neurons that transmute physical stimuli such as light and sound into neural signals, and motor neurons that transmute neural signals into activation of muscles or glands. (**Nervous Systems**)
- There are literally hundreds of different types of synapses. In fact, there are over a hundred known neurotransmitters, and many of them have multiple types of receptors. Molecular neuroscientists generally divide receptors into two broad groups: chemically gated ion channels and second messenger systems. When a second messenger system is activated, it starts a cascade of molecular interactions inside the target cell, which may ultimately produce a wide variety of complex effects, such as increasing or decreasing the sensitivity of the cell to stimuli, or even altering gene transcription. (**Nervous Systems**)
- Over the 1990's, there was a general model of synapses which was that they contained very few proteins and those few proteins that were known, which you could count on one hand more or less, were sufficient to produce synaptic transmission and synaptic plasticity. People thought that could account for learning, but it wasn't like that at all. What we found was that on the post-synaptic side of synapses (the side of the synapse where information first comes into a nerve cell), we identified ten times the number of proteins

that were previously known, which suddenly revealed a very much unexpected complexity. Quite a lot of people at the time thought that it was some sort of artifact, but it wasn't. It was actually only 1/10th as complex as what it turned out to be! Over the subsequent years, we found another tenfold more proteins. Many labs have confirmed this now. Essentially, inside a synapse, on the post-synaptic side, you can have more than a thousand types of proteins in the synapse. This really changed the way of thinking, from the synapse being just a "connector" in the nervous system. That's not a very nice way to talk about a synapse because in fact it's a super-sophisticated molecular computer. (**Seth Grant**)

- The human brain has a vast number of synapses—somewhere on the order of a million billion of them. So, we have a vast number of synapses, and we have a large number of proteins in the synapses. We have also uncovered evidence that these synapse proteins are not distributed the same across all synapses. In some parts of the brain, some proteins are found. In other parts of the brain, other proteins are found. And that was giving us this clue that there was this synapse diversity at a level that we hadn't really thought of before. (**Seth Grant**)
- In 2018, we published a paper that had been the first brain-wide survey of synapse diversity using a variety of methods. We now call this diversity the "synaptome." In the way the "genome" is all of the genes that an animal has, the synaptome is all of the synapses that an animal has. Just as genes have an architecture, synapse diversity has an architecture. We find that different types of synapses are found in different parts of the brain and they have certain proteins. Interestingly, those parts of the brain that are involved in higher cognitive functions—in the cortex and the hippocampus—are the parts where you find the most diversity of the types of synapses. This is telling us something important. Having all of those types of synapses is probably giving very sophisticated computation to the brain circuits for the types of behaviour like language, speech, and memory processing. (**Seth Grant**)
- If you look at the potential of the different combinations of these proteins, it's very easy to imagine that every single synapse could be different. A mouse has about 10 to the power of 11 synapses. Even with a small number of proteins, say 10, it's possible to have every one of those synapses be different. But as I've already said, there are more than a thousand in there and they also have other post-translation modifications. So, it would be very simple to have a mouse brain where every synapse is actually different. It could also easily be the case that in the human brain, which is much bigger, every synapse could be different. We don't think that's going to be the case. We think they are going to be organised and there are going to be abundant classes and non-abundant classes, and some might even have redundant functions meaning they might have different molecular makeups but function the same way. But there is plenty of scope for what you might call species differences in the synapse composition. I wouldn't be at all surprised if there were some types of synapses that are unique to mice, and some that are unique to humans. There will also be some that are conserved between the two species. It's going to be crucial to look at that. (**Seth Grant**)
- A neuron is called "identified" if it has properties that distinguish it from every other neuron in the same animal and if every individual organism belonging to the same species has one and only one neuron with the same set of properties. In vertebrate nervous systems, very few neurons are "identified" in this sense—in humans, there are believed to be none—but in simpler nervous systems, some or all neurons may be thus unique. In the roundworm C. elegans, whose nervous system is the most thoroughly described of any animal's, every neuron in the body is uniquely identifiable. One notable consequence of this fact is that the form of the C. elegans nervous system is completely specified by the genome, with no experience-dependent plasticity. (**Nervous Systems**)

- One very important subset of synapses is capable of forming memory traces by means of long-lasting activity-dependent changes in synaptic strength. The best-known form of neural memory is a process called long-term potentiation (LTP), which operates at synapses that use the neurotransmitter glutamate acting on a special type of receptor known as the NMDA receptor. Since the discovery of LTP in 1973, many other types of synaptic memory traces have been found, involving increases or decreases in synaptic strength that are induced by varying conditions, and last for variable periods of time. (**Nervous Systems**)
- Because of the variety of voltage-sensitive ion channels that can be embedded in the membrane of a neuron, many types of neurons are capable, even in isolation, of generating rhythmic sequences of action potentials, or rhythmic alternations between high-rate bursting and quiescence. When neurons that are intrinsically rhythmic are connected to each other by excitatory or inhibitory synapses, the resulting networks are capable of a wide variety of dynamical behaviors. (**Nervous Systems**)

3.04 Neurons proliferate and form nervous systems in animals that enable further reactions to the environment.

- The connections between neurons can form neural pathways, neural circuits, and larger networks that generate an organism's perception of the world and determine its behaviour. (**Nervous Systems**)
- The basic neuronal function of sending signals to other cells includes a capability for neurons to exchange signals with each other. Networks formed by interconnected groups of neurons are capable of a wide variety of functions, including feature detection, pattern generation, and timing, and there are seen to be countless types of information processing possible. (**Nervous Systems**)
- The nervous system is a highly complex part of an animal that coordinates its actions and sensory information by transmitting signals to and from different parts of its body. In vertebrates it consists of two main parts, the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS consists of the brain and spinal cord. The PNS consists mainly of nerves, which are enclosed bundles of the long fibers or axons, that connect the CNS to every other part of the body. The PNS is divided into three separate subsystems, the somatic, autonomic, and enteric nervous systems. Somatic nerves mediate voluntary movement. The autonomic nervous system is further subdivided into the sympathetic and the parasympathetic nervous systems. The sympathetic nervous system is activated in cases of emergencies to mobilize energy, while the parasympathetic nervous system is activated when organisms are in a relaxed state. The enteric nervous system functions to control the gastrointestinal system. Both autonomic and enteric nervous systems function involuntarily. (**Nervous Systems**)

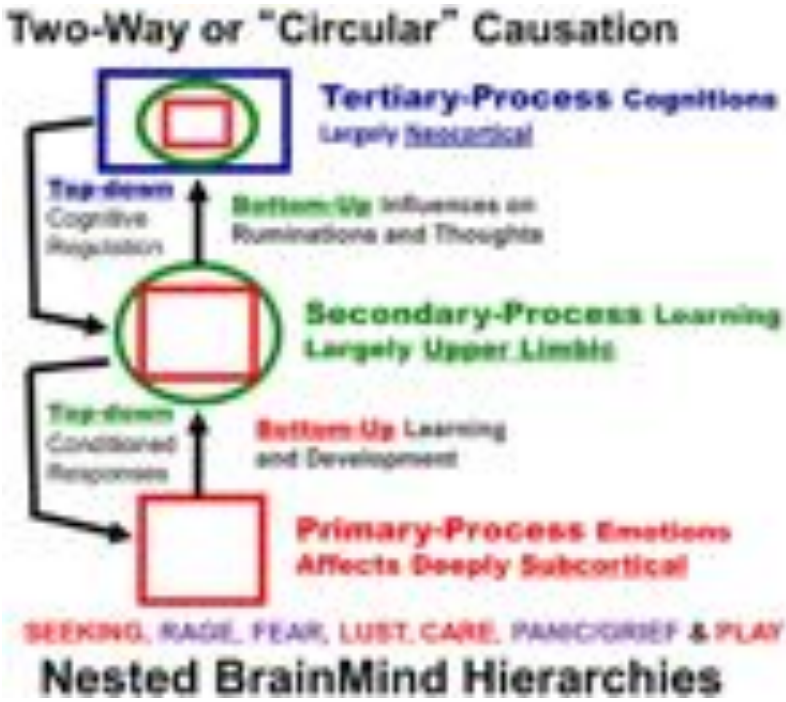
3.05 Nervous systems come together into nodes that evolve into more and more sophisticated brains. These are used to coordinate multiple streams of sensory information.

- In many species, the great majority of neurons participate in the formation of centralized structures (the brain and ganglia) and they receive all of their input from other neurons and send their output to other neurons. (**Nervous Systems**)
- Nervous systems are found in most multicellular animals but vary greatly in complexity. The only multicellular animals that have no nervous system at all are sponges, placozoans, and mesozoans, which have very simple body plans. The nervous systems of the radially symmetric organisms ctenophores (comb jellies) and cnidarians (which include anemones,

hydras, corals and jellyfish) consist of a diffuse nerve net. All other animal species, with the exception of a few types of worm, have a nervous system containing a brain, a central cord (or two cords running in parallel), and nerves radiating from the brain and central cord. The size of the nervous system ranges from a few hundred cells in the simplest worms, to around 300 billion cells in African elephants. (**Nervous Systems**)

3.06 Evolutionary pressures on organisms led them to develop brain modules, systems, and networks, which work in nested hierarchies to enable more and more complexity and effectiveness in understanding and responding to the world.

- The cerebral cortex is the largest site of neural integration in the central nervous system. It plays a key role in attention, perception, awareness, thought, memory, language, and consciousness. (**Cerebral Cortex**)
- The Global Neuronal Workspace Theory (from Dehaene) is an intellectual descendent of the Global Workspace Theory (Baars). These theories identify brain modules for: balance and coordination; memory; emotion; language; writing; attention, planning, organisation, reasoning; emotional affect, adaptability; motor / sensory; listening and decoding; reading and interpretation; visual-spatial, visual recognition. (Jerry Fodor called these modules informationally encapsulated, meaning somewhat private within each module.) … For some functions, there may be specific pathways through these modules, e.g. dorsal visual stream. … For general connections between multiple modules there may be a global workspace. This coordinates inputs from evaluative systems (value), attentional systems (focusing), long-term memory (past), and perceptual systems (present), into motor control outputs (future). Information in the global workspace is available from all modules and can be seen by each module. (**Introduction to Brain Consciousness**)
- The default mode network (DMN) is active when we're internally focused, thinking about ourselves and using our memory and imagination. The dorsal attention network (DAT), on the other hand, is activated when we're aware of and paying attention to the environment around us. (**Ramirez**)
- A team from the University of Michigan described their finding that the default mode network (DMN) and the dorsal attention network (DAT) are anti-correlated, meaning that when one is active, the other is suppressed. The team also found that neither network was highly active in people who were unconscious. These findings suggest that the interplay of the DMN and the DAT support consciousness by allowing us to interact with our surroundings then to quickly internalize those interactions, essentially turning our experiences into thoughts and memories. (**Ramirez**)
- Figure 1: (**Solms and Panksepp**)



3.07 In summary, all of these chemical, neuronal, and brain developments have evolved to produce a variety of mechanisms (too numerous to list here) for driving intentional behaviour. This extends from the simplest forms of cognition in plants up to and including the most sophisticated varieties that we are aware of in humans. These mechanisms produce the expanding abilities of functions in the 3rd level of my hierarchy of consciousness (attention, memory, pattern recognition, learning, and communication).

- Plant cognition is a field of research directed at experimentally testing the cognitive abilities of plants, including perception, learning processes, memory, and consciousness. Although they lack a brain and the function of a conscious working nervous system, plants are still somehow capable of being able to adapt to their environment and change the integration pathway that would ultimately lead to how a plant "decides" to take response to a presented stimulus. (**Plant Cognition**)
- A plant known as the *Mimosa pudica* was tested for the ability to adapt to closing its leaves upon repeated drops with no apparent harm appointed to the plant. The results showed that with repeated drops, the *Mimosa pudica* eventually stopped closing its leaves or opened its leaves quicker. This behaviour exhibited a trait in which the plant has adapted to not closing, or showing minimal closing, when repeated exposure to a non-harming situation is coupled with its own defence behaviour. (**Plant Cognition**)
- We declare the following: "The absence of a neocortex does not appear to preclude an organism from experiencing affective states. Convergent evidence indicates that non-human animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviours. Consequently, the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Non-human animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates." (**Cambridge Declaration of Consciousness**)
- Neural circuits supporting behavioural / electrophysiological states of attentiveness, sleep, and decision making appear to have arisen in evolution as early as the invertebrate

radiation, being evident in insects and cephalopod molluscs (e.g., octopus). (**Cambridge Declaration of Consciousness**)

- Present-tense emotionality is what communication by animals is mostly about. (**Sapolsky**)
- Joint attention refers to when two people look at and attend to the same thing; parents often use the act of pointing to prompt infants to engage in joint attention. The inclination to spontaneously reference an object in the world as of interest, via pointing, and to likewise appreciate the directed attention of another, may be the underlying motive behind all human communication. (**Theory of Mind Wikipedia**)

3.08 Note that all of these responses to stimuli are initially observable as reflexes. Later, brain modules and networks evolved that could sense and respond to other parts of the brain, as well as the internal affective moods that influence the entire organism. Such "meta" modules and networks are able to gain control over simpler systems by interrupting their pathways to action. This allows delays to what would otherwise be viewed as automatic responses. In effect, these act as internal (and therefore invisible) reflexes where logical if-then statements control local behaviour based on the wider context that is reported from other senses or memories.

- The simplest type of neural circuit is a reflex arc, which begins with a sensory input and ends with a motor output, passing through a sequence of neurons connected in series. This can be shown in the "withdrawal reflex" causing a hand to jerk back after a hot stove is touched. (**Nervous Systems**)
- Although the simplest reflexes may be mediated by circuits lying entirely within the spinal cord, more complex responses rely on signal processing in the brain. For example, when an object in the periphery of the visual field moves, and a person looks toward it, many stages of signal processing are initiated. The initial sensory response, in the retina of the eye, and the final motor response, in the oculomotor nuclei of the brain stem, are not all that different from those in a simple reflex, but the intermediate stages are completely different. Instead of a one- or two-step chain of processing, the visual signals pass through perhaps a dozen stages of integration, involving the thalamus, cerebral cortex, basal ganglia, superior colliculus, cerebellum, and several brainstem nuclei. These areas perform signal-processing functions that include feature detection, perceptual analysis, memory recall, decision-making, and motor planning. (**Nervous Systems**)

4.0 Prediction (This level in the hierarchy of consciousness is enabled by mechanisms for the cognitive abilities of anticipation, problem solving, and error detection.)

4.01 Once the mechanisms for memory, pattern recognition, and learning are set in place by the 3rd level of consciousness, animals with brains can begin to analyse current situations in order to predict what will happen next. This ability provides an enormous advantage to those who can master it.

- We can only perceive one signal at a time. And there is a 1/3 second time lag. Error prediction makes up for this. (**Post 9**)

4.02 There are various internal mechanisms that enable prediction. These fall under the general banner of predictive coding and include anticipation,

problem solving, error detection, and feelings of precision.

- Predictive coding is a theory of brain function in which the brain is constantly generating and updating a mental model of the environment. The model is used to generate predictions of sensory input that are compared to actual sensory input. This comparison results in prediction errors that are then used to update and revise the mental model. (**Predictive Coding**)
- The understanding of perception as the interaction between sensory stimuli (bottom-up) and conceptual knowledge (top-down) continued to be established by Jerome Bruner who, starting in the 1940s, studied the ways in which needs, motivations, and expectations influence perception, which came to be known as 'New Look' psychology. (**Predictive Coding**)
- The brain solves the seemingly intractable problem of modelling distal causes of sensory input through a version of Bayesian inference. It does this by modelling predictions of lower-level sensory inputs via backward connections from relatively higher levels in a cortical hierarchy. Constrained by the statistical regularities of the outside world (and certain evolutionarily prepared predictions), the brain encodes top-down generative models at various temporal and spatial scales in order to predict and effectively suppress sensory inputs rising up from lower levels. A comparison between predictions and sensory input yields a difference measure which, if it is sufficiently large beyond the levels of expected statistical noise, will cause the generative model to update so that it better predicts sensory input in the future. (**Predictive Coding**)
- The neural evidence for this is still in its infancy. The empirical evidence for predictive coding is most robust for perceptual processing. (**Predictive Coding**)
- The anterior cingulate is heavily involved in "error detection," noting discrepancies between what is anticipated and what occurs. (**Sapolsky**)
- Physiologically, precision is usually associated with the *postsynaptic gain* of cortical neurons reporting prediction errors. This is precisely the function of ERTAS modulatory neurons. (**Solms**)

4.03 Predictive coding changes living beings. They are no longer simply responders in the present tense to internal drives and external stimuli. Predictive beings are internally active thinkers trying to peer further and further into the future.

- Predictive coding inverts the conventional view of perception as a mostly bottom-up process, suggesting that it is largely constrained by prior predictions, where signals from the external world only shape perception to the extent that they are propagated up the cortical hierarchy in the form of prediction error. (**Predictive Coding**)
- Given that the world we live in is loaded with statistical noise, precision expectations must be represented as part of the brain's generative models, and they should be able to flexibly adapt to changing contexts. For instance, the expected precision of visual prediction errors likely varies between dawn and dusk, such that greater conditional confidence is assigned to errors in broad daylight than errors in prediction at nightfall. (**Predictive Coding**)

5.0 Awareness (This level in the hierarchy of consciousness is enabled by mechanisms for the cognitive abilities of self-reference.)

5.01 Through a mixture of abilities some brain networks gain general

awareness of some parts of the self.

- [At first,] the external body is not a subject but an object, and it is perceived in the same register as other objects. Something has to be added to simple perception before one's own body is differentiated from others. This level of representation (a.k.a. higher-order thought) enables the subject of consciousness to separate itself as an object from other objects. We envisage the process involving three levels of experience: (a) the subjective or phenomenal level of the *anoetic* self as affect, a.k.a. first-person perspective; (b) the perceptual or representational level of the *noetic* self as an object, no different from other objects, a.k.a. second-person perspective; (c) the conceptual or re-representational level of the *autonoetic* self in relation to other objects, i.e., perceived from an external perspective, a.k.a. third-person perspective. The self of everyday cognition is therefore largely an abstraction. That is why the self is so effortlessly able to think about itself in relation to objects, in such everyday situations as "I am currently experiencing myself looking at an object." (**Solms and Panksepp**)

5.02 The mechanisms underlying this awareness may be specific to each species. This is a keen area of research, but they appear to just be more and more neuroanatomy.

- Another approach to studying consciousness applies specifically to the study of self-awareness, that is, the ability to distinguish oneself from others. The classic example of testing this is known as the mirror test, which involves placing a spot of coloring on the skin or fur near an individual's forehead and seeing if they attempt to remove it or at least touch the spot. Humans (older than 18 months) and other great apes, bottlenose dolphins, killer whales, pigeons, European magpies, and elephants have all been observed to pass this test. (**Consciousness**)
- Birds appear to offer, in their behavior, neurophysiology, and neuroanatomy a striking case of parallel evolution of consciousness. Certain species of birds have been found to exhibit neural sleep patterns similar to those of mammals, including REM sleep and, as was demonstrated in zebra finches, neurophysiological patterns, previously thought to require a mammalian neocortex. Magpies in particular have been shown to exhibit striking similarities to humans, great apes, dolphins, and elephants in studies of mirror self-recognition. Evidence of near human-like levels of consciousness has been most dramatically observed in African grey parrots. (**Cambridge Declaration of Consciousness**)
- When stimuli are presented to patients, but masked so they can't detect it consciously, the visual cortex and amygdala are activated and that's it. When the stimulus is not masked, you get activation in the visual cortex, the amygdala, and the prefrontal cortex as well. ... In order to be conscious of an apple, it not only needs to be represented in your visual cortex, it needs to be re-represented, which involves the prefrontal cortex. … So, the prefrontal cortex is emerging as an important area in the consolidation of our conscious experiences into what they are. (**Post 12**)
- When conscious access occurs, brain activity is strongly activated when a threshold of awareness is crossed. At that point the signal spreads to many brain areas. There are four highly reproducible signals associated with this. Signature 1: activation in parietal and prefrontal circuits. Signature 2: a slow wave called P3 that pairs late, approximately 1/3 second after stimulus (i.e. consciousness lags behind the world). Signature 3: deep brain electrodes detect late and sudden bursts of high frequency oscillations. Signature 4: information exchange across distant brain areas. (**Post 9**)

6.0 Abstraction (This level in the hierarchy of consciousness is enabled by mechanisms for understanding and creating symbols, art, language, memes, writing, mathematics, philosophy, and science, which all act to expand culture.)

6.01 Once the abstraction of self-awareness is established, further abstractions can also be made. Abstract thinking requires representing something in the brain that does not exist in front of it. Therefore, studying the mechanisms that underlie these abstract representations requires the self-report that is currently only available in humans. This is another area of ongoing research, but it appears that many brain areas are associated with distinct forms of abstract thinking.

- The gold standard for whether a response is conscious or not is whether you can talk about it. This doesn't mean language and consciousness are identical, just that you have access to the experience to think about it (and we use language to discuss that access with one another). In non-human animal research, that doesn't exist. (**Post 12**)
- Brodmann areas have been discussed, debated, refined, and renamed exhaustively for nearly a century and remain the most widely known and frequently cited cytoarchitectural organization of the human cortex. (**Brodmann area**)
- Many of the brain areas defined by Brodmann have their own complex internal structures. In a number of cases, brain areas are organized into topographic maps, where adjoining bits of the cortex correspond to adjoining parts of the body, or of some more abstract entity. (**Brodmann area**)
- The Broca area is a region in the frontal lobe of the dominant hemisphere of the brain (usually the left) with functions linked to speech production. (**Broca's area**)
- Wernicke's area is one of the two parts of the cerebral cortex that are linked to speech, the other being Broca's area. It is involved in the comprehension of written and spoken language, in contrast to Broca's area, which is involved in the production of language. It is traditionally thought to reside in Brodmann area 22. (**Wernicke's area**)
- Brodmann area 47 has been implicated in the processing of syntax in oral and sign languages, musical syntax, and semantic aspects of language. (**Brodmann area**)
- Higher order functions of the associated cortical areas are consistently localized to the same Brodmann areas by neurophysiological, functional imaging, and other methods. However, functional imaging can only identify the approximate localization of brain activations in terms of Brodmann areas since their actual boundaries in any individual brain requires its histological [post-mortem] examination. (**Brodmann area**)

6.02 These abstract abilities are useful for grasping all sorts of knowledge about the world. But they also gave humans greater moral capacities for responding to the world. For example, it would be very difficult for evolution to suddenly produce new emotional affects for the feelings of empathy and moral disgust, but human brains have learned to apply old reactions to new circumstances according to abstract rules. This greatly extended the flexibility of human consciousness.

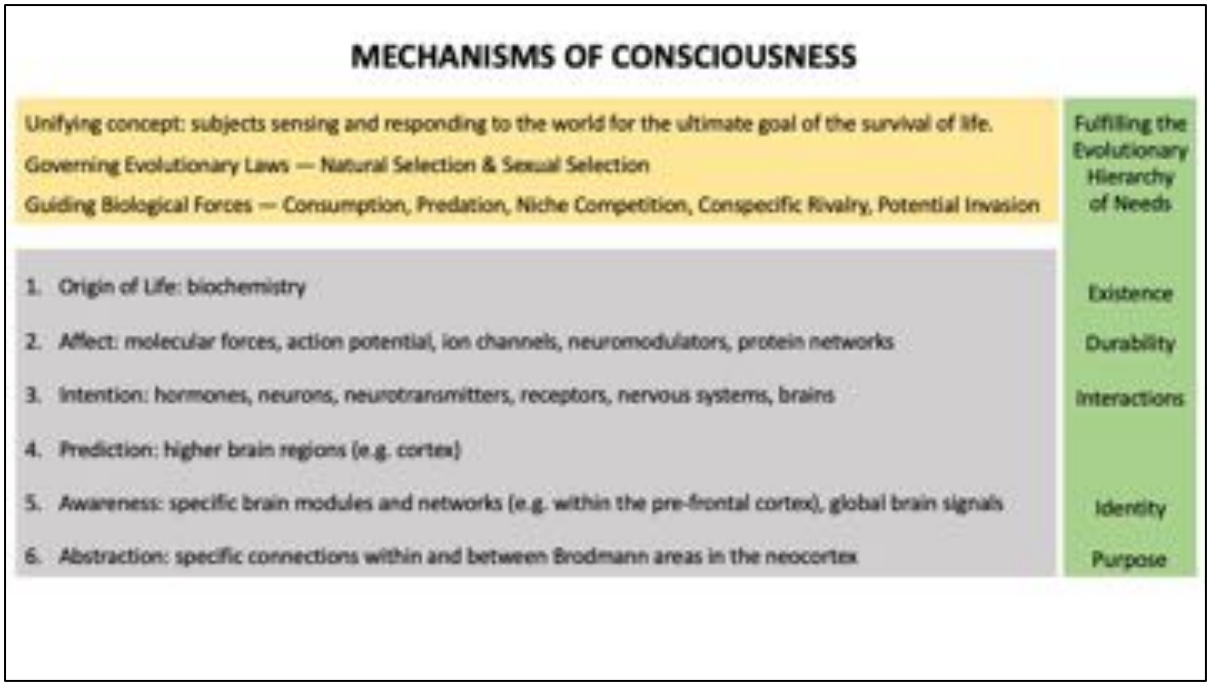
- There are still ways that humans appear to stand alone. One of those is hugely important: the human capacity to think symbolically. Metaphors, similes, parables, figures of speech —they exert enormous power over us. (**Sapolsky**)
- As our hominid ancestors kept getting better at [abstract thinking], great individual and social advantages accrued. We became capable of representing emotions in the past and

possible emotions in the future, as well as things that have nothing to do with emotion. We evolved a uniquely dramatic means of separating message from meaning and intent: lying. And we invented aesthetic symbolism; after all, those 30,000-year-old paintings of horses in Chauvet cave are not really horses. (**Sapolsky**)

- In recent years, scientists have made remarkable insights into the neurobiology of symbols. A major finding from their work is that the brain is not very good at distinguishing between the metaphorical and literal. In fact, symbols and metaphors, and the morality they engender, are the product of clunky processes in our brains. (**Sapolsky**)
- The best way to shine a light on this unwieldy process is through metaphors for two feelings critical to survival: pain and disgust. … There are fancier more recently evolved parts of the brain in the frontal cortex that assess the meaning of pain. Maybe [a pain signal is] bad news, or maybe it's good news. Much of this assessing occurs in a frontal cortical region called the anterior cingulate. This structure is heavily involved in "error detection," noting discrepancies between what is anticipated and what occurs. … In experimental settings, you're playing with two [people] and suddenly they start ignoring you and only toss the ball between them. Junior high all over again. And the brain scanner shows that the neurons in your anterior cingulate activate. In other words, rejection hurts. … Both abstract social and literal pain impact the same cingulate neurons. (**Sapolsky**)
- While in a brain scanner, you're administered a mild shock, delivered through electrodes on your fingers. All the usual brain regions activate, including the anterior cingulate. Now you watch your beloved get shocked in the same way. The brain regions that ask, "Is it my finger or toe that hurts?" remain silent. It's not their problem. But your anterior cingulate activates, and as far as it's concerned, "feeling someone's pain" isn't just a figure of speech. You seem to feel the pain too. As evolution continued to tinker, it did something remarkable with humans. It duct-taped the anterior cingulate's role in giving context to pain into a profound capacity for empathy. (**Sapolsky**)
- Studies show the human anterior cingulate is more complex than in other species, with more connections to abstract, associational parts of the cortex, regions that can call your attention to the pains of the world, rather than the pain in your big toe. (**Sapolsky**)
- Our brains' shaky management of symbols adds tremendous power to a unique human quality: morality. You're in a brain scanner and because of the scientist's weirdly persuasive request, you bite into some rotten food. Something rancid and fetid and skanky. This activates another part of the frontal cortex, the insula, which, among other functions, processes gustatory and olfactory disgust. It sends neuronal signals to face muscles that reflexively spit out that bite, and to your stomach muscles that make you puke. All mammals have an insula that processes gustatory disgust. But we are the only animal where that process serves something more abstract. … Think about something awful you once did, something deeply shameful. The insula activates. It has been co-opted into processing that human invention: moral disgust. (**Sapolsky**)

Brief comments to close

Once again, let's summarise all of the above research into a hierarchical chart. This one mimics the format of the one I produced for the functions of consciousness, but now we have its companion for the mechanisms as well.



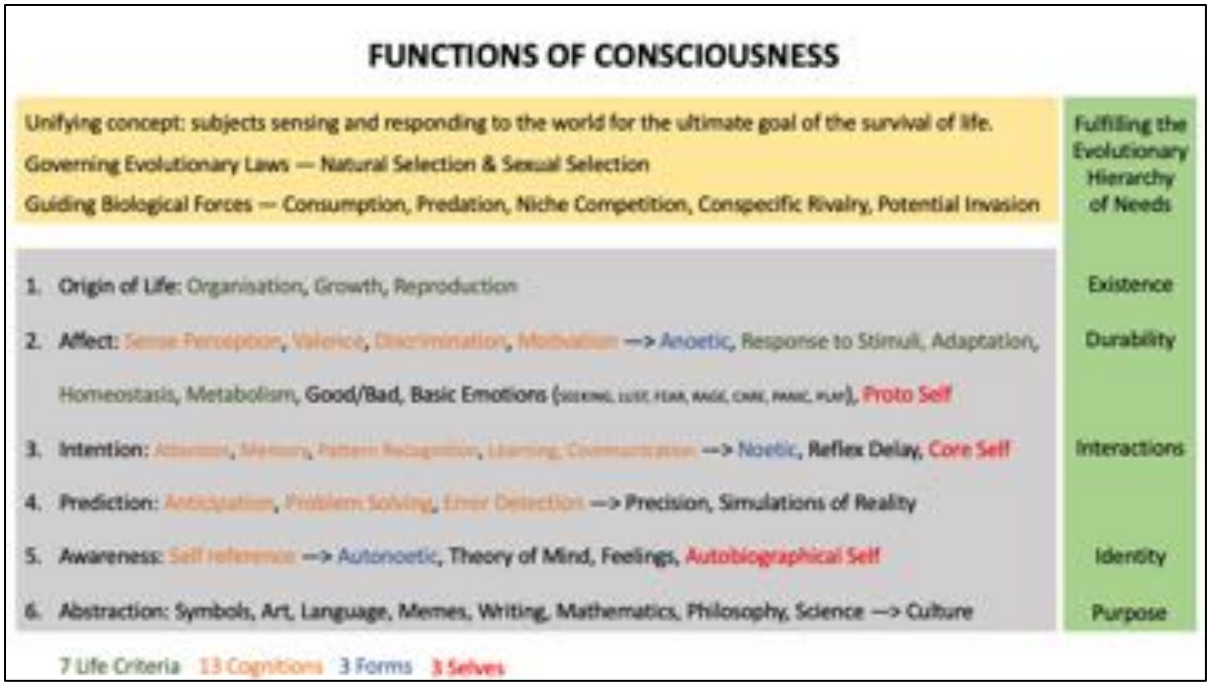
Please note once again that these are the mechanisms of consciousness as they now exist in biotic life. As I have said before, there is nothing stopping artificial life from experiencing its own unique feelings of consciousness via new mechanisms. But I maintain that these are still the hierarchies that must be observed if artificial consciousness is to become one that we will recognise.

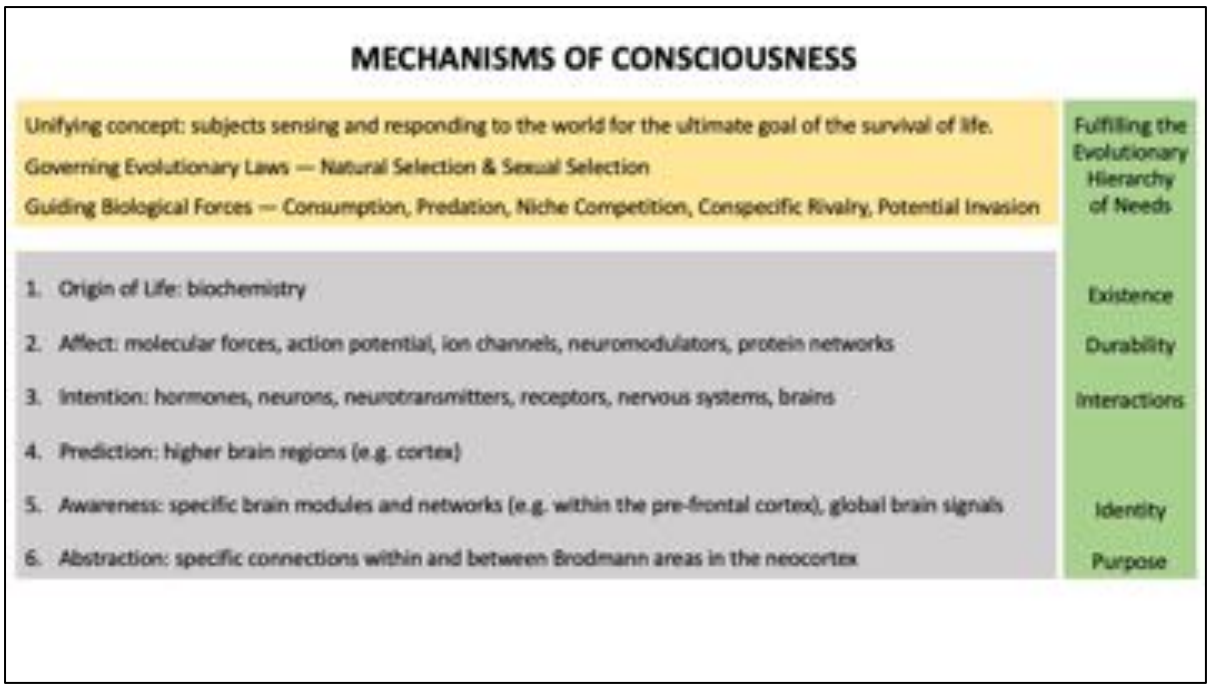
Next up, I'll tackle the ontogeny of consciousness in a human being. Hopefully that will fit just as well into this hierarchy and my use of Tinbergen's 4 questions will continue to provide a comprehensive understanding of this complex phenomenon. I'll bet you can't wait!

21 — Development Over a Lifetime (Ontogeny)

20 November 2020

Time now to tackle the third of **Tinbergen's four questions**. His framework for analysing all biological phenomena is definitely helping me shed light on the multifaceted concept of consciousness. So far, my hierarchical definition was rounded into shape by looking at **the functions of consciousness**, and then that hierarchy held up well as I went through a physicalist account of **the mechanisms of consciousness**. These two tables summarise my findings so far:





These first two questions represent the two static elements of Tinbergen's model. They are "**contemporary**" accounts that explain the current form of a behaviour in its present condition. The next two questions move on to the dynamic elements, or the "**chronicle**" accounts that explain the biological phenomenon in terms of the sequence that got it there. For example, you wouldn't have a complete understanding of a frog without exploring both its tadpole phase as well as the evolutionary history of amphibians. The species history is known as the tale of phylogeny, which I'll leave for last. In this post, I'll look at the local / proximate tale of an individual, which is called ontogeny. And in particular, I'll confine myself to the ontogeny of human consciousness, since that's the only form of consciousness that we have first-hand experience of. Before I get to the details of that, let me give a few general notes.

General notes about ontogeny

As Dan Dennett **has noted**, an evolutionary view of history requires one to drop any essentialised views. There just aren't any eternal essences in nature that suddenly turn on or off. Likewise, everything in biology has changed slowly over billions of years in tiny increments. Viewing consciousness through this lens helps you see it as a slowly growing phenomenon over the life of an individual. (We'll get to how it has slowly grown over the development of species in the next post.) This is precisely why a hierarchical definition of consciousness is required—to mimic the evolutionary change and growth of nature.

In **an earlier post**, I said this reminded me of "the parable of the immune system" that the evolutionary scientist David Sloan Wilson likes to use. On **an episode of The Psychology Podcast**, Wilson said: *"The human immune system is immensely modular. We inherit it, and it does not change during our lifetime. It is something that evolved by genetic evolution, but it is triggered by environmental circumstances. [This] adaptive component of the immune system is highly evolutionary. That's the ability of antibodies to vary and for the successful antigens to be ramped up. So that's an evolutionary process that takes place during the lifetime of the organism. The whole thing is densely modular but also amazingly open-ended. Why can't we say the same thing about the human behavioural system?"* Well, it seems obvious (to me anyway) that we *can* say the same thing about our behaviour—that it adapts during our lifetimes to successful and unsuccessful interactions with the environment. And the same thing applies to our growing powers of consciousness—they give life more and more degrees of freedom as they help an organism make more and more sense of its environment.

In **a recent Brain Science Podcast**, the neuroscientist Seth Grant discussed how we are beginning to see neurological evidence for this kind of development of consciousness over a lifetime. He noted,

"We all know that humans and every other animal go through a stereotypical trajectory of lifespan behavioural changes. When a human baby is born, they have a very limited set of behavioural responses. But they very rapidly, over the course of months and years, develop an increasingly complex behavioural repertoire. They also go through phases, famously described by Piaget and others. … These lifespan trajectories have been well documented before, but the question is, why do we have them? Why do they come about? In this paper that we've just published, we found that there was a remarkable set of changes [in synapses] across the lifespan. The synaptome map of the brain and its architecture changes at every age throughout life. In other words, our brain is always changing. Over the first few months of the mouse, which is the equivalent of the first few decades of a human, there was a remarkable explosion of synaptic diversity. Just as the behavioural repertoire was very limited in the new-born animal, so was the synaptic diversity. This fits well with our earlier work linking the two."

This neurological evidence is just beginning to come in, so I can't say much more than this about the ontology of mechanisms for consciousness, but as Grant noted, the psychologist **Jean Piaget** spent a lifetime studying the behavioural development of children. He argued that intelligence develops in a series of stages that are related to age and are progressive because one stage must be accomplished before the next can occur. By the end of the 20th century, Piaget was second only to B. F. Skinner as the most cited psychologist of that era. Although his ideas were subjected to massive scrutiny, Piaget's original model "has proved to be remarkably robust." I don't want to pretend that there haven't been "innumerable improvements and qualifications of his work, coming from a plethora of neo-Piagetian and post-Piagetian variants," but for the purposes of sketching out the ontology of consciousness, Piaget proves invaluable and I'll rely on him heavily.

Just as before, during the examination of the first two Tinbergen questions, there are lot of intricate details to consider. So, I'll continue to write simple numbered statements followed by their justifications so you can quickly read the statements to get the gist of my arguments. You can dip into any of the details for each statement if you want further information. Or click on the links there for even more. I'll also continue to work within the structure of my hierarchy since it is proving to be an effective guide to consciousness. Here goes!

1.0 Origin of Life. The first three criteria for life are: organisation, growth, and reproduction.

1.1 The genetic makeup for each new human is determined at conception. Fertilisation kicks off biological processes that, when successful, eventually lead to a life. The first three criteria for life are gradually met during the early stages of prenatal development.

- The first sperm cell to successfully penetrate the egg cell donates its genetic material (DNA) to combine with the DNA of the egg cell resulting in a new organism called the zygote. The term "conception" refers variably to either fertilization or to formation of the conceptus after its implantation in the uterus, and this terminology is controversial. (**Prenatal Development**)
- The first two weeks from fertilization is referred to as the germinal stage or pre-embryonic stage. The zygote spends the next few days traveling down the fallopian tube dividing several times to form a ball of cells called a morula. Further cellular division is accompanied by the formation of a small cavity between the cells. This stage is called a blastocyst. Up to this point there is no growth in the overall size of the embryo, as it is confined within a glycoprotein shell, known as the zona pellucida. Instead, each division produces successively smaller cells. (**Prenatal Development**)
- The blastocyst reaches the uterus at roughly the fifth day after fertilization. It is here that disintegration of the zona pellucida occurs. This process is analogous to hatching. This allows cells of the blastocyst to come into contact with, and adhere to, cells of the uterus. In most successful pregnancies, the embryo implants 8 to 10 days after ovulation. Rapid growth occurs and the embryo's main features begin to take form. This process is called differentiation, which produces the varied cell types, such as blood cells, kidney cells, and nerve cells. (**Prenatal Development**)

2.0 Affect. The first four cognitive abilities—response to stimuli, adaptation, homeostasis, and metabolism—enable the fulfilment of the final four criteria for life: sense perception, valence, discrimination, and motivation.

2.1 The final four criteria for life are met in the next stages of prenatal development. A viable fetus will, on its own, display the first four cognitive abilities in order to remain alive and begin to adapt to its environment.

- Following fertilization, the embryonic stage of development continues until the end of the 10th week. By the end of the tenth week of gestational age the embryo has acquired its basic form and is referred to as a fetus. The next period is that of fetal development where many organs become fully developed. Development continues throughout the life of the fetus and through into life after birth. Significant changes occur to many systems in the period after birth as they adapt to life outside the uterus. (**Prenatal Development**)
- The perinatal period is "around the time of birth." In developed countries and at facilities where expert neonatal care is available, it is considered from 22 completed weeks of gestation (the time when birth weight is normally 500 g) to 7 completed days after birth. In many of the developing countries the starting point of this period is considered 28 completed weeks of gestation (or weight more than 1000 g). (**Prenatal Development**)
- While there is no sharp limit of development, gestational age, or weight at which a human fetus automatically becomes viable, a 2013 study found that "While only a small proportion of births occur before 24 completed weeks of gestation (about 1 per 1000), survival is rare and most of them are either fetal deaths or live births followed by a neonatal death." According to studies between 2003 and 2005, 20 to 35 percent of babies born at 24 weeks of gestation survived, while 50 to 70 percent of babies born at 25 weeks, and more than 90 percent born at 26 to 27 weeks, survived. (**Fetal Viability**)
- One 2018 study showed that there was a significant difference between countries in what was considered to be the "grey zone": the "grey zone" was considered to be 22.0 - 22.6/23 weeks in Sweden, 23.0 – 23.6/24 weeks in the UK, and 24.0-25.6/26 weeks in Netherlands. (**Fetal Viability**)
- Viability, as the word has been used in United States constitutional law since Roe v. Wade, is the potential of the fetus to survive outside the uterus after birth, natural or induced, when supported by up-to-date medicine. Fetal viability depends largely on the fetal organ maturity, and environmental conditions. (**Fetal Viability**)
- The United States Supreme Court stated in Roe v. Wade (1973) that viability "is usually placed at about seven months (28 weeks) but may occur earlier, even at 24 weeks." The 28-week definition became part of the "trimester framework" marking the point at which the "compelling state interest" (under the doctrine of strict scrutiny) in preserving potential life became possibly controlling, permitting states to freely regulate and even ban abortion after the 28th week. (**Fetal Viability**)

2.2 In addition to the basic physical stimuli that a fetus responds to, language also induces changes in brain states at this early stage of life.

- There is evidence that the acquisition of language begins in the prenatal stage. After 26 weeks of gestation, the peripheral auditory system is already fully formed. Also, most low-frequency sounds (less than 300 Hz) can reach the fetal inner ear in the womb of mammals. Those low-frequency sounds include pitch, rhythm, and phonetic information related to language. Studies have indicated that fetuses react to and recognize differences between sounds. Such ideas are further reinforced by the fact that newborns present a preference for their mother's voice, present behavioural recognition of stories only heard during gestation, and (in monolingual mothers) present preference for their native language. A more recent study with EEG demonstrated different brain activation in newborns hearing their native language compared to when they were presented with a

different language, further supporting the idea that language learning starts while in gestation. (**Prenatal Development**)

2.3 Some abilities associated with the final criteria for life are an innate part of humans, hardwired by our genetic evolution. Much behaviour is plastic, however, and must be learned through experience.

- Fortunately, living organisms are not required to learn everything about the world from scratch. Each phenotype is endowed with innate predictions concerning biologically significant situations it is certain to encounter. Fear behaviors (freezing and fleeing), for example, are innate predictions; but each individual has to learn *what* to fear and *what else* might be done in response. What vertebrates do to meet their needs always consists in a combination of innate and learned behaviors. (**Solms**)
- As far as we know, all cortical functional specializations are developmental/epigenetic. The columns of cortex are initially almost identical in neuronal architecture, and the famous differences in Brodmann's areas probably arise from use-dependent plasticity. (**Solms and Panksepp**)
- Much of what we have traditionally thought to be unconditioned about exteroceptive consciousness is actually learned. This has been well demonstrated by the research of Mriganka Sur, which shows that total removal of "visual" cortex in fetal mice (in utero) does not impair their adult vision at all, and redirecting visual input from occipital cortex to auditory cortex in ferrets leads to reorganization of the latter tissue to support completely competent vision. Clearly, from a corticocentric viewpoint, this either means that sensory perception is completely learned, or that perceptual functionality is completely controlled by subcortical structures, with subtle developmental extensions of affective experience perhaps being the foremost vehicle. In short, one of the great mistakes of modern cognitive neuroscience may be the assumption that cortical consciousness is built on intrinsic "hard-wired" cognitive computational principles. The resolution of conscious experiences in the neocortex may be largely learned developmental/epigenetic functions of the brain. (**Solms and Panksepp**)

2.4 The first mechanism for learning is experiencing which actions or situations lead directly to states of mind that are rewarding or punishing. This innate experience of "good" or "bad" affect drives behaviour towards survival. This is the base underpinning all consciousness.

- Interoceptive consciousness is phenomenal; it "feels like" something. Above all, the phenomenal states of the body-as-subject are experienced affectively. Affects, rather than representing discrete external events, are experienced as positively and negatively valenced states. Their valence is determined by how changing internal conditions relate to the probability of survival and reproductive success. The empirical evidence for the feeling component are simply based on the highly replicable fact that wherever in the brain one can artificially evoke coherent emotional response patterns with deep brain stimulation, those shifting states uniformly are accompanied by "rewarding" and "punishing" states of mind. By attributing valence to experience—determining whether something is "good" or "bad" for the subject, within a biological system of values—affective consciousness (and the behaviours it gives rise to) intrinsically promotes survival and reproductive success. This is what consciousness is for. (**Solms and Panksepp**)

2.5 The first stages of childhood development explore the environment with

very rudimentary behaviours and reflexes.

- Stage 1.1 of Piaget's Four Stages (0–1 months): Reflex schema stage—Babies learn how the body can move and work. Vision is blurred and attention spans remain short through infancy. They are not particularly aware of objects to know they have disappeared from sight. However, babies as young as seven minutes old prefer to look at faces. The three primary achievements of this stage are: sucking, visual tracking, and hand closure. (**Object Permanence**)

3.0 Intention. Five more cognitive abilities—attention, memory, pattern recognition, learning, and communication—enable intentional actions of the core self, eventually including the delay of reflexes.

3.1 As babies interact with the environment and further develop their cognitive capabilities, they begin to act with intentions rather than mere reflexes. They pay attention, build memories, and learn, but the classic A-not-B error shows they have not yet built prediction models in their consciousness.

- To start out, infants only engaged in primarily reflex actions such as sucking, but not long after, they would pick up objects and put them in their mouths. When they do this, they modify their reflex response to accommodate the external objects into reflex actions. Because the two are often in conflict, they provide the impetus for intellectual development. The constant need to balance the two triggers intellectual growth. (**Piaget**)
- Stage 1.2 of Piaget's Four Stages (1–4 months): Primary circular reactions—Babies notice objects and start following their movements. They continue to look where an object was, but for only a few moments. They "discover" their eyes, arms, hands, and feet in the course of acting on objects. This stage is marked by responses to familiar images and sounds (including parents' faces) and anticipatory responses to familiar events (such as opening the mouth for a spoon). The infant's actions become less reflexive and intentionality emerges. (**Object Permanence**)
- Stage 1.3 of Piaget's Four Stages (4–8 months): Secondary circular reactions—Babies will reach for an object that is partially hidden, indicating knowledge that the whole object is still there. If an object is completely hidden, however, the baby makes no attempt to retrieve it. The infant learns to coordinate vision and comprehension. Actions are intentional, but the child tends to repeat similar actions on the same object. Novel behaviours are not yet imitated. (**Object Permanence**)
- Stage 1.4 of Piaget's Four Stages (8–12 months): Coordination of secondary circular reactions—This is deemed the most important for the cognitive development of the child. At this stage the child understands causality and is goal-directed. The very earliest understanding of object permanence emerges, as the child is now able to retrieve an object when its concealment is observed. This stage is associated with the classic **A-not-B error**. After successfully retrieving a hidden object at one location (A), the child fails to retrieve it at a second location (B). (**Object Permanence**)

3.2 Actions gradually become more complex through the integration of more information, including the formation of memories. It is our neuroplasticity that enables each individual to learn from their particular lived experience.

- At first glance, it may seem that the reason we don't remember being babies is because infants and toddlers don't have a fully developed memory. But babies as young as six

months can form both short-term memories that last for minutes, and long-term memories that last weeks, if not months. (**Shinskey**)

- Neural plasticity is the ability of the brain to change continuously throughout an individual's life, e.g., brain activity associated with a given function can be transferred to a different location, the proportion of grey matter can change, and synapses may strengthen or weaken over time. Neuroplasticity can be observed at multiple scales, from microscopic changes in individual neurons to larger-scale changes such as cortical remapping in response to injury. Behavior, environmental stimuli, thought, and emotions may also cause neuroplastic change through activity-dependent plasticity, which has significant implications for healthy development, learning, memory, and recovery from brain damage. (**Neuroplasticity**)
- Hebbian theory is a neuroscientific theory claiming that an increase in synaptic efficacy arises from a presynaptic cell's repeated and persistent stimulation of a postsynaptic cell. It is an attempt to explain synaptic plasticity, the adaptation of brain neurons during the learning process. The theory is often summarized as "Cells that fire together wire together." (**Neuroplasticity**)

3.3 As babies begin to understand more and more about their environment, early forms of communication with them becomes possible.

- Joint attention refers to when two people look at and attend to the same thing; parents often use the act of pointing to prompt infants to engage in joint attention. The inclination to spontaneously reference an object in the world as of interest, via pointing, and to likewise appreciate the directed attention of another, may be the underlying motive behind all human communication. (**Theory of Mind Wikipedia**)

4.0 Prediction. This level in the hierarchy of consciousness is enabled by mechanisms for the cognitive abilities of anticipation, problem solving, and error detection.

4.1 Once intentions exist (either one's own or the intentions of others), they can be taken into account. To do so is to use prediction to think through what the result will be from any intentions. This requires three more cognitive capacities from Lyon's list: anticipation, problem solving, and error detection.

- Infants' understanding of attention in others acts as a "critical precursor" to the development of theory of mind. Understanding attention involves understanding that seeing can be directed selectively as attention, that the looker assesses the seen object as "of interest", and that seeing can induce beliefs. (**Theory of Mind**)
- In this process, the organism must stay "ahead of the wave" of the biological consequences of its choices (to use the analogy that gave Andy Clark's (2016) book its wonderful title: *Surfing Uncertainty*): "To deal rapidly and fluently with an uncertain and noisy world, brains like ours have become masters of prediction—surfing the waves of noisy and ambiguous sensory stimulation by, in effect, trying to stay just ahead of the place where the wave is breaking (p. xiv)." (**Solms**)
- The Bayesian brain is [optimized] through the encoding of better models of the world leading to better predictions. It is important to note that in this model, prediction error (mediated by the sensory affect of surprise), is a "bad" thing, biologically speaking. The more veridical the brain's generative model of the world, the less surprise (the less salience, the less consciousness, the more automaticity), the better. Freud called this the "Nirvana principle". [In simpler terms,] the goal of all learning is automatised mental

processes, increased predictability, and reduced uncertainty or surprise. (**Solms and Panksepp**)

4.2 The ability of brains to predict the world is signalled by the full grasping of object permanence. Human babies generally reach this level of development around the age of two. From then on, they operate according to schemata, which are continually refined throughout life.

- Object permanence is the understanding that objects continue to exist even when they cannot be seen, heard, touched, smelled, or sensed in any way. It is one of an infant's most important accomplishments, as, without this concept, objects would have no separate, permanent existence. In Piaget's theory of cognitive development, infants develop this understanding by the end of the "sensorimotor stage", which lasts from birth to about two years of age. (**Object Permanence**)
- Stage 1.5 of Piaget's Four Stages (12–18 months): Tertiary circular reaction—The child gains means-end knowledge and is able to solve new problems. The child is now able to retrieve an object when it is hidden several times within their view but cannot locate it when it is outside their perceptual field. (**Object Permanence**) During this stage infants explore new possibilities of objects; they try different things to get different results. (**Piaget**)
- Stage 1.6 of Piaget's Four Stages (18–24 months): Invention of new means through mental combination—The child fully understands object permanence. They will not fall for A-not-B errors. Also, a baby is able to understand the concept of items that are hidden in containers. If a toy is hidden in a matchbox then the matchbox put under a pillow and then, without the child seeing, the toy is slipped out of the matchbox and the matchbox then given to the child, the child will look under the pillow upon discovery that it is not in the matchbox. The child is able to develop a mental image, hold it in mind, and manipulate it to solve problems, including object permanence problems that are not based solely on perception. The child can now reason about where the object may be when invisible displacement occurs. (**Object Permanence**) [In other words, this stage involves] internalization of schemata. (**Piaget**)
- A Schema is a structured cluster of concepts, it can be used to represent objects, scenarios, or sequences of events or relations. The original idea was proposed by philosopher Immanuel Kant as innate structures used to help us perceive the world. A schema (pl. schemata) is the mental framework that is created as children interact with their physical and social environments. (**Piaget**)
- While much research has been done on infants, theory of mind develops continuously throughout childhood and into late adolescence as the synapses (neuronal connections) in the prefrontal cortex develop. Children seem to develop theory of mind skills sequentially. The first skill to develop is the ability to recognize that others have diverse desires. Children are able to recognize that others have diverse beliefs soon after. The next skill to develop is recognizing that others have access to different knowledge bases. Finally, children are able to understand that others may have false beliefs and that others are capable of hiding emotions. (**Theory of Mind**)

5.0 Awareness. This level in the hierarchy of consciousness is enabled by mechanisms for the cognitive abilities of self-reference.

5.1 By making cognitive connections between intentions, predictions, and internal affective feelings, the development of self-awareness slowly arises.

- [At first,] the external body is not a subject but an object, and it is perceived in the same register as other objects. Something has to be added to simple perception before one's own body is differentiated from others. This level of representation (a.k.a. higher-order thought) enables the subject of consciousness to separate itself as an object from other objects. We envisage the process involving three levels of experience: (a) the subjective or phenomenal level of the *anoetic* self as affect, a.k.a. first-person perspective; (b) the perceptual or representational level of the *noetic* self as an object, no different from other objects, a.k.a. second-person perspective; (c) the conceptual or re-representational level of the *autonoetic* self in relation to other objects, i.e., perceived from an external perspective, a.k.a. third-person perspective. The self of everyday cognition is therefore largely an abstraction. That is why the self is so effortlessly able to think about itself in relation to objects, in such everyday situations as "I am currently experiencing myself looking at an object." (**Solms and Panksepp**)
- As predictions and perceptions improve, organisms eventually make the connection that there is a self which has its own mind. Awareness is achieved. This development is covered by the final cognitive capacity from Lyon's list: self-reference. Such conscious cognition allows memories and thoughts built from the lived past and the anticipated future to create the autonoetic, autobiographical self. (**Post 19**)

5.2 Mirror Self-Recognition tests currently act as our best marker for the attainment of this level of consciousness. Passing this test is correlated with object permanence, and similarly occurs in humans around the age of two, although there are differences in this timeline that appear to be related to rearing behaviours.

- The Mirror Self-Recognition test is the traditional method for attempting to measure self-awareness. However, agreement has been reached that animals can be self-aware in ways not measured by the mirror test, such as distinguishing between their own and others' songs and scents. … Very few species have passed the MSR test. Species that have include the great apes (including humans), a single Asiatic elephant, dolphins, orcas, the Eurasian magpie, and the cleaner wrasse. A wide range of species have been reported to fail the test, including several species of monkeys, giant pandas, and sea lions. … A strong correlation between self-concept and object permanence has been demonstrated. (**Mirror Test**)
- From the ages of 6 to 12 months, the child typically sees a "sociable playmate" in the mirror's reflection. Self-admiring and embarrassment usually begin at 12 months, and at 14 to 20 months, most children demonstrate avoidance behaviors. Finally, at 18 months, half of children recognize the reflection in the mirror as their own and by 20 to 24 months, self-recognition climbs to 65%. (**Mirror Test**)
- A 2010 cross-cultural study observed variations in the presence of self-oriented behaviors exhibited by children (ranging from 18 to 55 months old) from non-Western rural communities and Western urban and rural communities when each was given the mark test. They found that children from Western communities showed earlier signs of self-oriented behaviors toward the mark when given the mirror mark test, whereas an absence of this behavior was seen in children from non-Western communities. Such results do not suggest a delayed development in cognition in the latter group, but rather the potential of how differences in parenting styles (as influenced by culture) impact the way children express self-concept. … For example, a Cameroonian Nso sample of infants 18 to 20 months of age had an extremely low amount of self-recognition outcomes at 3.2%. The study also found two strong predictors of self-recognition: object stimulation (maternal

effort of attracting the attention of the infant to an object either person touched) and mutual eye contact. (**Mirror Test**)

6.0 Abstraction. This level in the hierarchy of consciousness is enabled by mechanisms for understanding and creating symbols, art, language, memes, writing, mathematics, philosophy, and science, which all act to expand culture.

6.1 Passing through the first five levels of consciousness in this hierarchy creates a very aware living being that is rare in the animal kingdom. And yet, we don't remember any of the steps it took to get us there.

- Most of us don't have any memories from the first three to four years of our lives—in fact, we tend to remember very little of life before the age of seven. The phenomenon, known as "childhood amnesia", has been puzzling psychologists for more than a century—and we still don't fully understand it. (**Shinskey**)
- Virtually nobody has memories from very early childhood. It's clear that young children do remember facts in the moment such as who their parents are, or that one must say "please" before mom will give you candy. This is called "semantic memory." Until sometime between the ages two and four, however, children lack "episodic memory"—memory regarding the details of a specific event. (**Shouse**)
- The typical boundary for the offset of childhood amnesia—three and a half years—shifts with age. Children and teenagers have earlier memories than adults do. This suggests that the problem may be less with forming memories than with maintaining them. (**Shinskey**)
- [There is a] theory that we can't remember our first years simply because our brains hadn't developed the necessary equipment. The explanation emerges from the most famous man in the history of neuroscience, known simply as patient HM. After a botched operation to cure his epilepsy damaged his hippocampus, HM was unable to recall any new events. Intriguingly, however, he was still able to learn other kinds of information—just like babies. When scientists asked him to copy a drawing of a five-pointed star by looking at it in a mirror (harder than it sounds), he improved with each round of practise—despite the fact the experience itself felt completely new to him. Perhaps, when we're very young, the hippocampus simply isn't developed enough to build a rich memory of an event. Baby rats, monkeys and humans all continue to add new neurons to the hippocampus for the first few years of life and we are all unable to form lasting memories as infants—and it seems that the moment we stop creating new neurons, we're suddenly able to form long-term memories. (**Gorvett**)
- While the neurological explanation does account for blanks in very young children's memories, it does not give a full explanation for childhood amnesia because it fails to account for the years after the age of four. It also fails to address the issue that children themselves do not show childhood amnesia. Children around the age of two to three have been found to remember things that occurred when they were only one to two years old. This discovery that three-year-olds can retrieve memories from earlier in their life implies that all necessary neurological structures are in place to recall episodic information over the short-term, but evidently not over the long-term into adulthood. (**Childhood Amnesia**)

6.2 It may be that this lack of memory is precisely because one must pass through the first five levels of consciousness in order to begin recording an autobiographical self.

- The development of a cognitive self is also thought by some to have a strong effect on encoding and storing early memories. As toddlers grow, a developing sense of the self begins to emerge as they realize that they are a person with unique and defining characteristics and have individual thoughts and feelings separate from others. As they gain a sense of the self, they can begin to organize autobiographical experiences and retain memories of past events. This is also known as the development of a theory of mind which refers to a child's acceptance that they have beliefs, knowledge, and thoughts that no one else has access to. The developmental explanation asserts that young children have a good concept of semantic information but lack the retrieval processes necessary to link past and present episodic events to create an autobiographical self. Young children do not seem to have a sense of a continuous self over time until they develop awareness for themselves as an individual human being. (**Childhood Amnesia**)

6.3 The development of language is key to this ability to think about the world, to recall past events, or to imagine future possibilities. Language is an abstract representation of these things, which allows them to be repeatedly brought into the present tense of a mind, which is how memories are formed and reformed.

- Another factor that we know plays a role is language. From the ages of one to six, children progress from the one-word stage of speaking to becoming fluent in their native language(s), so there are major changes in their verbal ability that overlap with the childhood amnesia period. This includes using the past tense, memory-related words such as "remember" and "forget", and personal pronouns, a favourite being "mine". A child's ability to verbalise about an event at the time that it happened predicts how well they remember it months or years later. One lab group conducted this work by interviewing toddlers brought to accident and emergency departments for common childhood injuries. Toddlers over 26 months, who could verbalise about the event at the time, recalled it up to five years later, whereas those under 26 months, who could not talk about it, recalled little or nothing. (**Shinskey**)
- On the "self-awareness being tied to language" note, I found this quote from Helen Keller interesting: "Before my teacher came to me, I did not know that I am. I lived in a world that was a no-world. I cannot hope to describe adequately that unconscious, yet conscious time of nothingness. (…) Since I had no power of thought, I did not compare one mental state with another." (Hellen Keller, 1908: quoted by Daniel Dennett, 1991, *Consciousness Explained*. p 227) (**Hiskey**)
- If I ask you to picture a rope and climbing up it, you can do it. I specifically chose those objects and actions because it is exactly what a chimp in a zoo is familiar with. If I asked a chimp to do the same thing, could it? We don't know, but I suspect not, because you can't do it wordlessly. You need to be able to interact using language. Without language, I don't think you have the cognitive systems for self-simulation and self-probing that we have. … Language allows us to be conscious of things we otherwise wouldn't be able to be conscious of. (**Dennett**)

6.4 Language also vastly enlarges our abilities to recognise patterns in the world. This is vital for understanding and predicting events.

- Differences in knowledge yield striking differences in the capacity to pick up patterns. Expert chess players can instantly perceive (and subsequently recall with high accuracy) the total board position in a real game but are much worse at recall if the same chess pieces are randomly placed on the board, even though to a novice both boards are equally hard to recall. This should not surprise anyone who considers that an expert speaker of

English would have much less difficulty perceiving and recalling: "The frightened cat struggled to get loose" than "Te serioghehnde t srugfcalde go tgtt ohle" which contains the same pieces, now somewhat disordered. Expert chess players, unlike novices, not only know how to *Play* chess; they know how to *read* chess—how to see the patterns at a glance. (**Dennett**)

6.5 The development of language occurs gradually over the last three of Piaget's stages. This drastically expands the use of symbols and logic, which are the hallmarks of this sixth level of abstract consciousness.

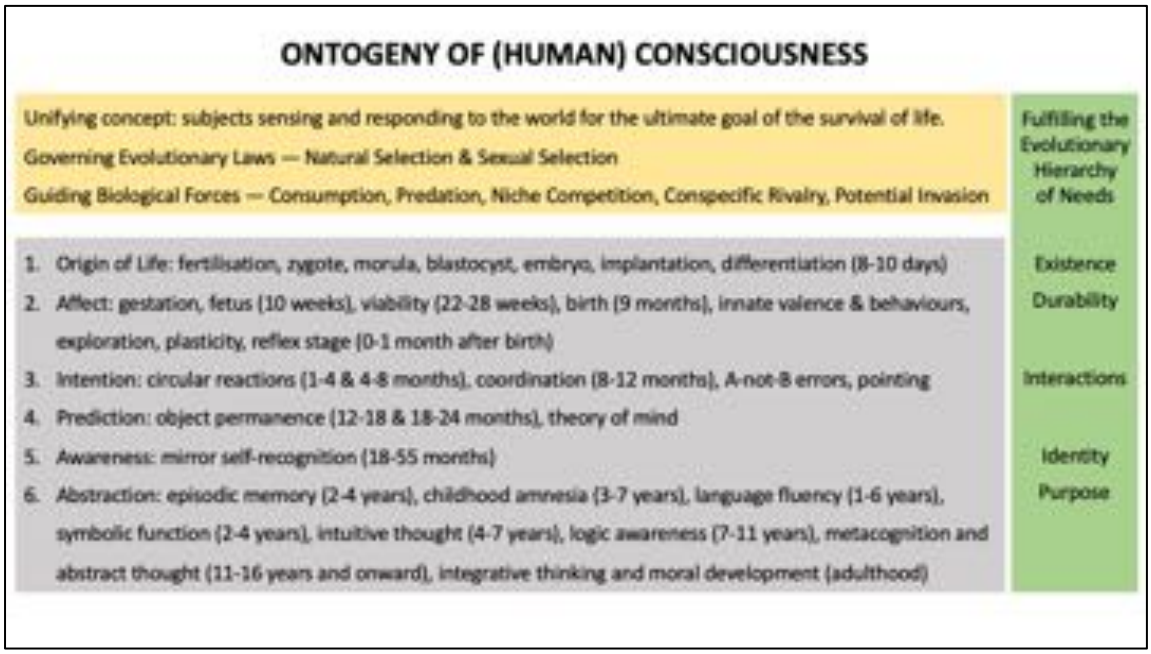
- Stage 2 of Piaget's Four Stages—Preoperational stage: starts when the child begins to learn to speak at age two and lasts up until the age of seven. During the pre-operational stage of cognitive development, Piaget noted that children do not yet understand concrete logic and cannot mentally manipulate information. Children's increase in playing and pretending takes place in this stage. However, the child still has trouble seeing things from different points of view. The Pre-operational Stage is split into two substages: the symbolic function substage, and the intuitive thought substage. (**Piaget**)
- Stage 2.1 of Piaget's Four Stages—Symbolic Function Substage: from two to four years of age children find themselves using symbols to represent physical models of the world around them. Children are able to understand, represent, remember, and picture objects in their mind without having the object in front of them. Play is demonstrated by the idea of checkers being snacks, pieces of paper being plates, and a box being a table. (**Piaget**)
- Stage 2.2 of Piaget's Four Stages—Intuitive Thought Substage: between about the ages of four and seven, children tend to become very curious and ask many questions, beginning the use of primitive reasoning. Children tend to propose the questions of "why?" and "how come?" This stage is when children want the knowledge of knowing everything. Piaget called it the "intuitive substage" because children realize they have a vast amount of knowledge, but they are unaware of how they acquired it. (**Piaget**)
- Stage 3 of Piaget's Four Stages—Concrete operational stage: from ages seven to eleven. Children can now conserve and think logically (they understand reversibility) but are limited to what they can physically manipulate. They are no longer egocentric. During this stage, children become more aware of logic and conservation, topics previously foreign to them. Children also improve drastically with their classification skills. (**Piaget**)
- Stage 4 of Piaget's Four Stages—Formal operational stage: from age eleven to sixteen and onwards. Children develop abstract thought and can easily conserve and think logically in their mind. Abstract thought is newly present during this stage of development. Children are now able to utilize metacognition. Along with this, the children in the formal operational stage display more skills oriented towards problem solving, often in multiple steps. The child is able to identify the properties of objects by the way different kinds of actions affect them. This is the process of "empirical abstraction." By repeating this process across a wide range of objects and actions, the child establishes a new level of knowledge and insight. Once the child has constructed these new kinds of knowledge, he or she starts to use them to create still more complex objects and to carry out still more complex actions. As a result, the child starts to recognize still more complex patterns and to construct still more complex objects. (**Piaget**)

6.6 Beyond Piaget's examination of childhood, human consciousness can continue to expand by integrating more and more information. An objectively good criterion for this expansion of consciousness would evaluate whether it created better and better models for surviving and thriving.

- Piaget's theory stops at the formal operational stage, but other researchers have observed the thinking of adults is more nuanced than formal operational thought. This fifth stage has been named post formal thought or operation. There are many theorists, however, who have criticized "post formal thinking," because the concept lacks both theoretical and empirical verification. The term "integrative thinking" has been suggested for use instead. (**Piaget's theory of cognitive development**)
- Lawrence Kohlberg's stages of moral development constitute an adaptation of a psychological theory originally conceived by the Swiss psychologist Jean Piaget. The theory holds that moral reasoning, a necessary (but not sufficient) condition for ethical behavior, has six developmental stages, each more adequate at responding to moral dilemmas than its predecessor. Kohlberg followed the development of moral judgment far beyond the ages studied earlier by Piaget, who also claimed that logic and morality develop through constructive stages. The six stages of moral development occur in three levels. Level 1, Pre-Conventional, consists of: Stage 1—obedience and punishment orientation (How can I avoid punishment?); and Stage 2—self-interest orientation (What's in it for me?). Level 2, Conventional, consists of Stage 3—interpersonal accord and conformity (The good boy/girl attitude); and Stage 4—authority and social-order maintaining orientation (Law and order morality). Finally, level 3, Post-Conventional, consists of Stage 5—social contract orientation; and Stage 6—universal ethical principles. (**Lawrence Kohlberg's stages of moral development**)

Brief comments to close

I didn't know how this post was going to go, but it is extremely exciting to see the biological and psychological research conform perfectly to my hierarchy. It's a great example of consilience where multiple streams of evidence are all pointing to the same thing. So, riding high, I'll close now with my third summary chart and look forward to completing my examination of consciousness next time with the fourth of Tinbergen's four questions. I can hardly wait for that.



22 — Our Shared History (Phylogeny)

24 December 2020

Almost done now…

Finally, we come to the fourth of **Tinbergen's four questions**. His framework for analysing all biological phenomena has proven to be extremely useful. The first two questions represented the two static elements of Tinbergen's model. They tell the "**contemporary**" accounts of the current form of a behaviour in its present condition. For consciousness, that helped me develop my hierarchy by looking at **the functions of consciousness**, and then I saw that this hierarchy held up well as I went through a physicalist account of **the mechanisms of consciousness**. The next two questions move on to the dynamic elements, or the "**chronicle**" accounts that explain a biological phenomenon in terms of the sequence that got it there. In the last post, I looked at ontogeny, or **the development of consciousness over a (human) lifetime**. Now, we can look at the ultimate tale of consciousness—its phylogenetic history.

I noted at the end of my last post about ontogeny that I wasn't sure how well my hierarchy would continue to hold up, but it was extremely exciting to see that the biological and psychological research into human development conformed perfectly well with my hierarchy. This was a great example of consilience where multiple streams of evidence are all pointing to the same thing. Now, it's time for one final check. But before I can go through the details of the hierarchical development of consciousness, I have to lay the groundwork with a general overview of the phylogeny of life.

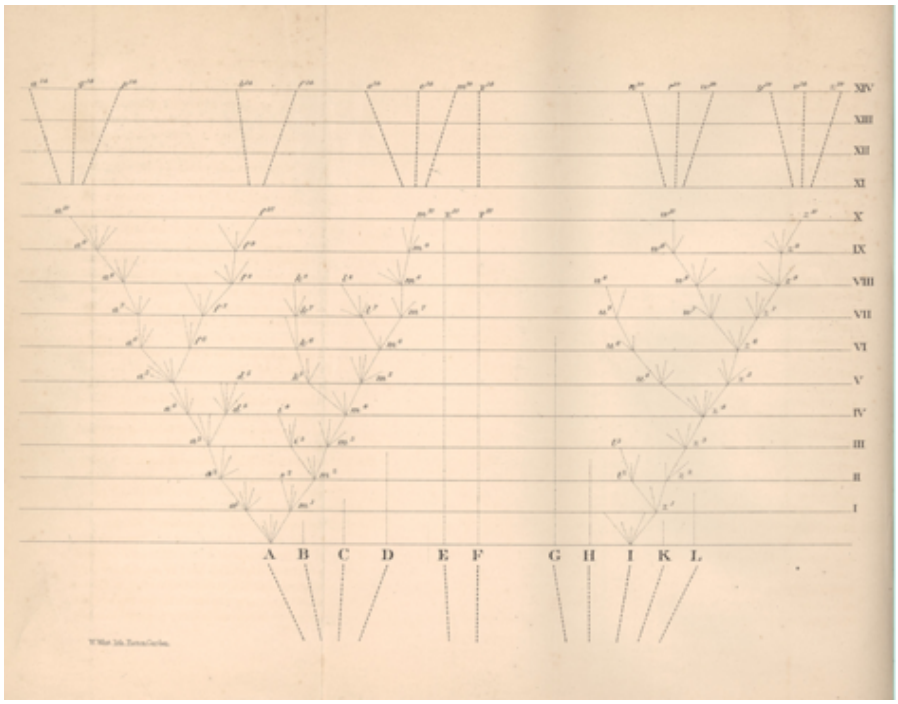
General notes about phylogeny

David Christian has perhaps developed the most well-known sweeping story of existence with his groundbreaking **Big History** project. He divides the history of the universe into eight fundamental thresholds.

1. The Big Bang kicked off the origins of all cosmology about 13.7 billion years ago.
2. The first stars and galaxies appeared perhaps 100 million years later.
3. Chemical elements were created inside dying stars just 1 or 2 million years after that.
4. The Earth and the Solar System were created about 4.5 billion years ago.
5. The first evidence of life on Earth comes about 3.8 billion years ago.
6. The creation of our own species, Homo sapiens, happened about 300,000 years ago.
7. The emergence of agriculture occurred about 11,000 years ago.
8. Finally, the Modern era of human history covers the last three or four centuries.

That's a nice and simple outline. Bill Gates liked it so much **he invested $10 million into the Big History Project in 2011** to help try it out in actual classrooms. But for understanding the growth of consciousness during the evolutionary history of life, Big History is pretty weak. A panpsychist might like the fact that 4 of the 8 phases occur before life, but as **Dan Dennett** said, "Electrons can't accrue memories. They do not change over billions of years. They do not *participate* in the arrow of time, so there is no way for them to be said to have intentions, feelings, purposes, or goals." I agree. To me, you need a subject to have a subjective experience. Therefore, we need a lot more detail about the 5th threshold covering the 3.6 billion years of life before Homo sapiens.

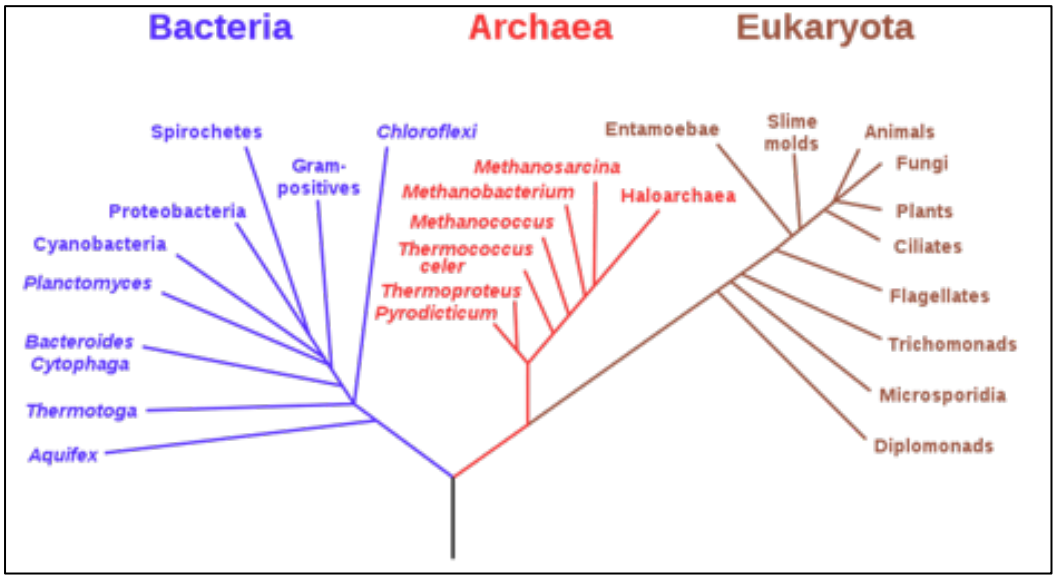
The most common way to do this is with a **tree of life**. There are literally hundreds of them out there, all of which were inspired by the only illustration that appeared in Darwin's *On the Origin of Species* in 1859.



Many of us will know the branches of these trees by a **mnemonic** such as "King Phillip came over for great spaghetti." (Or, if you are a fan of **Community**, "Kevin, please come over for gay sex.") But after 1990, three domains were identified on top of all this, which necessitated a new phrase like "Do kindly place candy out for good students." This, of course, stands for:

Domain → Kingdom → Phylum → Class → Order → Family → Genus → Species

Here's a simple phylogenetic tree showing the current three-domain system where all smaller branches can be considered kingdoms.



Note how plants, animals, and fungi are just three twigs on the far upper right of this tree. An even more humbling depiction was produced by **David Mark Hillis** in 2008 based on completely sequenced genomes. See his popular "Hillis Plot" depiction below, where Homo sapiens are just one twig placed two ticks before midnight. Take a moment to google a few of our closest relatives. It's only 10 steps to brewer's yeast!

In 2015, the first draft of the **Open Tree of Life** was published, in which information from nearly 500 previously published trees was combined into a single free online database. The first draft included 2.3 million species, mainly composed of bacteria.

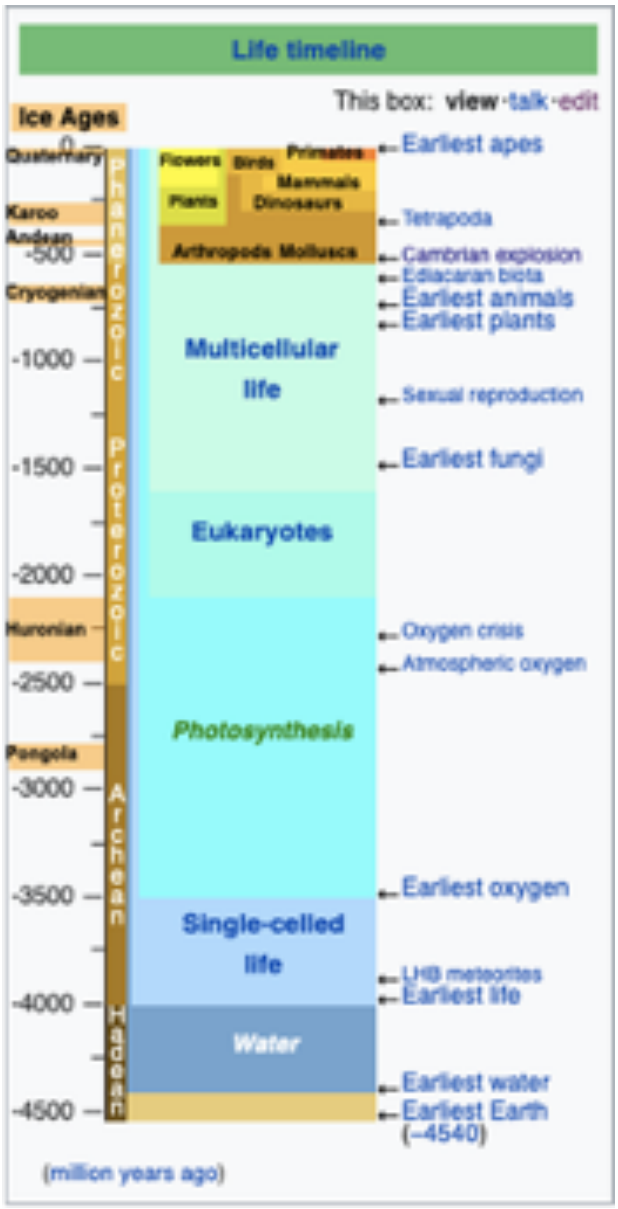
We now know that **horizontal gene transfer** may mean that these trees are more of a **tangled thicket** than neatly branching pyramids, but the most important points still stand. All of life is related, and we all share the same building blocks. And this is important to the understanding of consciousness because of everything these building blocks do.

In neuroscientist Peter Sterling's new book **What Is Health? Allostasis and the Evolution of Human Design**, Sterling details the evolutionary continuity from earlier life forms to humans. He breaks our evolutionary past into four epochs — single cell life, multicellular organisms, mammals, and humans. What's important for him as a neuroscientist, though, is just how much went on during that first epoch. Single cell life evolved the core metabolic processes such as using ATP for energy as well as the genetic code that enables all life to share the same proteins and enzymes. A full 75% of our proteins are homologues (similar in position, structure, and origin but not necessarily in function) to those seen in prokaryotes, the first life forms.

Another neuroscientist, **Seth Grant**, notes that "all of the proteins that are in our synapses evolved before the brain and before multi-cellular organisms. They evolved in unicellular

organisms that have lived on the planet for several billion years before the first multi-cellular animal. [T]hose proteins are controlling the behaviour of unicellular organisms. They control how they adapt and respond to their environment. They are involved in how they learn responses to their environment. And this tells us then that the fundamental molecular machinery of the behaviour of the human brain is actually the fundamental molecular machinery of behaviour in unicellular organisms, and some of those molecules go all the way back to the last universal common ancestor, which is about 3.5 billion years ago."

So, through a variety of methods, we have an incredibly detailed picture of the interrelated lines of descent for all of life. When did these changes happen? Well, you can look into any one species to know more about its heritage, but here is **a general timeline**:



Okay. So where is consciousness??

As **Dan Dennett** said, "The search for the simplest form of consciousness is a snipe hunt. Starfish have some elements of consciousness, so do trees, and bacteria. (But not electrons.) We can argue about motor proteins. The question of 'where do you draw the line?' is an ill-

motivated question. Where do you draw the line between night and day?"

Dennett's pithy quote echoes comments made in the Wikipedia entry for **Consciousness**.

"Opinions are divided as to where in biological evolution consciousness emerged and about whether or not consciousness has any survival value. Some argue that consciousness is a by-product of evolution. It has been argued that consciousness emerged (i) exclusively with the first humans, (ii) exclusively with the first mammals, (iii) independently in mammals and birds, or (iv) with the first reptiles. Other authors date the origins of consciousness to the first animals with nervous systems or early vertebrates in the Cambrian over 500 million years ago. Donald Griffin suggests in his book Animal Minds *a gradual evolution of consciousness."*

Basically, and unsurprisingly, seeing the emergence of consciousness depends on your definition of consciousness. I myself believe that the amorphous concept of consciousness can best be understood as the collection of processes that enable living organisms (governed by the various laws of selection) to sense and respond to biological forces. This leads us to another of "**Darwin's strange inversions**" where, rather than looking for a single set of essential criteria that tells us whether consciousness is on or off, we are much better served by considering all of the tiny incremental additions of capabilities that has led to everything we now call consciousness. In this way, I think my theory brings **other definitions of consciousness** together under one umbrella and it matches the biological history of life. Having a dynamic picture, in the form of a growing hierarchy, mimics the growing tree of life. And this makes thinking about consciousness much clearer. (At least it does for me.)

So, let's look at how my hierarchy lays on top of all of this. Just as before, during the examination of the first three Tinbergen questions, there are lot of intricate details to consider. Therefore, I'll once again write simple numbered statements in bold, followed by their justifications in bullet point format, so you can quickly read the statements to get the gist of my arguments. This allows you to dip into any of the details for each statement if you want further information. Or you can click on the links provided for even more. Since my hierarchy has proven to work well so far, I'll continue to work within it and hope that it proves to be an effective guide to consciousness this one last time. Here goes!

1.0 Origin of Life. The first three criteria for life are: organisation, growth, and reproduction.

1.1 Our best hypothesis for the initial creation of life is that it emerged from basic chemical processes alone. Our best estimate is that this occurred in the form of microbes somewhere between 3.8 and 4.4 billion years ago.

- Just like RNA, early nucleotides could both store information and function as enzymes. Early polymer enzymes would: enhance replication, use high energy molecules in the environment (near thermal vents) to recharge monomers, synthesize lipids from other molecules in the environment, and modify your lipids so they don't leave your membrane. And that's it. A simple 2-component system that spontaneously forms in the pre-biotic environment can eat, grow, contain information, replicate, and evolve, simply through thermodynamic, mechanical, and electrical forces. No ridiculous improbability, no supernatural forces, no lightening striking a mud puddle. Just chemistry. (**Abiogenesis**)
- The earliest known life forms on Earth are putative fossilized microorganisms found in hydrothermal vent precipitates. The earliest time that life forms first appeared on Earth is at least 3.77 billion years ago, possibly as early as 4.28 billion years, or even 4.5 billion

years — not long after the oceans formed 4.41 billion years ago, and after the formation of the Earth 4.54 billion years ago. (**Earliest Known Life Forms**)

2.0 Affect. The first four cognitive abilities—response to stimuli, adaptation, homeostasis, and metabolism—enable the fulfilment of the final four criteria for life: sense perception, valence, discrimination, and motivation.

2.1 As soon as life begins (by organising, growing, and reproducing), a subject emerges. Since physical forces act on all physical matter, these subjects will be affected. Since change is inevitable in a dynamic universe, they will face selection pressures. These forces and pressures are the initial core of subjectivity that affect life right from the very beginning.

- How far back in evolution does the ability to detect and respond to danger go? Other nonhuman animals do this. Even bees. But it's much older still. Protozoa like paramecia or amoeba do it. Even bacteria do. In fact, it goes all the way back to the beginning of life. (**Post 12**)
- Higher cortical regions add much to consciousness. Of course they do. But the evolutionary "roots" of consciousness are to be found elsewhere, and they are probably affective. (**Solms and Panksepp**)

2.2 Once life is faced with selection pressures, tiny lifespans with even minimal variability will produce incredibly vast numbers of trials and errors over millions and billions of years. This evolutionary process will produce all sorts of solutions that will, logically, result in longer and/or more stable forms of survival.

- It's not just detecting danger either — incorporating nutrients, balancing fluids and ions, thermoregulation, reproduction for the species to survive — all of these behaviours exist in animals, but also in single-cell microbes. Value / valence / affect has been present since the beginning of life. (**Post 12**)
- Organisms evolve digestive vs. respiratory vs. thermoregulatory vs. immune systems. Each such specialized system is governed by a homeostatic imperative of its own. Metabolic energy balance, oxygenation, hydration, and thermoregulation (for example) are not the same things, although each of them contributes to the overall imperative of organism-wide [optimization]. (**Solms**)

2.3 This affective core of consciousness, driven by chemistry alone, is the simplest version of the objective chemical changes we call emotions. Note that these are embodied moods, which are separate from the subjective mental feelings that some individuals later evolve to have.

- I like Damasio's distinctions between emotions, feelings, and valences. This fits very well with my own system for mapping cognitive appraisals (i.e. judging if something is good, bad, or unknown, aka valanced) onto different events in the past, present, or future, in order to generate the things we typically call emotions (but which Damasio would distinguish as feelings). I can certainly get behind his distinction here. I could also adopt his labelling. I think he's got "the strange order of things" right by saying the chemical emotional responses would have come first before the feelings in our self became able to identify them. (**Post 10**)

2.4 In humans, we know that the earliest emotional responses differentiated into four basic drives over 300 million years ago. They keep individuals alive. Later, between 55 and 85 million years ago, they differentiated further into three more basic drives that benefit social groups.

- Evolutionary neuroscientist Jaak Panksepp of Bowling Green State University has identified seven emotional systems in humans that originated deeper in our evolutionary past than the Pleistocene era (over 2.5 million years ago). The emotional systems that Panskepp terms CARE (tenderness for others), PANIC (from loneliness), and PLAY (social joy) date back to early primate evolutionary history (55-85 million years ago), whereas the systems of FEAR, RAGE, SEEKING, and LUST, which govern survival instincts for the individual, have even earlier, premammalian origins (older than 300 million years ago). (**Gibney**)(**Panksepp**)

2.5 The full history of the development of affect in living organisms is of course too long and varied to give in detail. But here are some interesting highlights from the rough timeline along the way as the final four criteria for life have developed — sense perception, valence, discrimination, and motivation.

- The emergence of nervous systems has been linked to the evolution of voltage-gated sodium (Nav) channels. The Nav channels allow for communication between cells over long distances through the propagation of action potentials, whereas voltage-gated calcium (Cav) channels allow for unmodulated intercellular signaling. It has been hypothesized that Nav channels differentiated from Cav channels either at the emergence of nervous systems or before the emergence of multicellular organisms, although the origin of Nav channels in history remains unknown. (**Nervous System**)
- A voltage-gated sodium channel is present in members of the choanoflagellates, thought to be the closest living, unicellular relative of animals. This suggests that an ancestral form of the animal channel was among the many proteins that play central roles in animal life, but which are thought to have evolved before multicellularity. (**Sodium Channel**)
- Multicellularity has evolved independently at least 25 times in eukaryotes, and also in some prokaryotes. The first evidence of multicellularity is from cyanobacteria-like organisms that lived 3–3.5 billion years ago. (**Multicellular organism**)
- Sponges were first to branch off the evolutionary tree from the common ancestor of all animals (roughly 580 to 750 million years ago), making them the sister group of all other animals. Sponges have no cells connected to each other by synaptic junctions, that is, no neurons, and therefore no nervous system. Unlike other animals, they lack true tissues and organs. Sponges do not have nervous, digestive, or circulatory systems. Instead, most rely on maintaining a constant water flow through their bodies to obtain food and oxygen and to remove wastes. Sponge cells have the ability to communicate with each other via calcium signalling or by other means. Sponge larvae differentiate sensory cells which respond to stimuli including light, gravity, and water movement, all of which increase the fitness of the organism. (**Sponge**)
- The nerve net is the simplest form of a nervous system found in multicellular organisms. Unlike central nervous systems, where neurons are typically grouped together, neurons found in nerve nets are spread apart. This nervous system allows cnidarians to respond to physical contact. They can detect food and other chemicals in a rudimentary way. While the nerve net allows the organism to respond to its environment, it does not serve as a means by which the organism can detect the source of the stimulus. For this reason,

simple animals with nerve nets, such as Hydra, will typically produce the same motor output in response to contact with a stimulus regardless of the point of contact. (**Nervous System**)

- Nerve nets are found in species in the phyla Cnidaria (e.g. scyphozoa, box jellyfish, and sea anemones), Ctenophora, and Echinodermata. Cnidaria and Ctenophora both exhibit radial symmetry and are collectively known as coelenterates. Coelenterates diverged 570 million years ago, prior to the Cambrian explosion, and they are the first two phyla to possess nervous systems which differentiate during development and communicate by synaptic conduction. The nervous systems of coelenterates allow for sensation, contraction, locomotion, and hunting/feeding behaviors. (**Nervous System**)
- The vast majority of existing animals are bilaterians, meaning animals with left and right sides that are approximate mirror images of each other. All bilateria are thought to have descended from a common wormlike ancestor that appeared in the Ediacaran period, 550–600 million years ago. The fundamental bilaterian body form is a tube with a hollow gut cavity running from mouth to anus, and a nerve cord with an enlargement (a "ganglion") for each body segment, with an especially large ganglion at the front, called the "brain". (**Nervous System**)
- Neo-Piagetian stages have been applied to the maximum stage attained by various animals. For example, spiders (an order of arthropods) attain the circular sensory motor stage, coordinating actions and perceptions. (**Piaget**)
- The evolutionary ancestry of arthropods dates back to the Cambrian period. Small arthropods with bivalve-like shells have been found in Early Cambrian fossil beds dating 541 to 539 million years ago. (**Arthropod**)
- Vertebrates originated about 525 million years ago during the Cambrian explosion, which saw a rise in organism diversity. (**Vertebrate**)
- When we look at all of the proteins in mammals or vertebrate species, we find that there are a lot more of them than we find in invertebrate species. So, how could that be? It turned out that the genomes of some animal that is the ancestor of all vertebrates underwent an entire genome duplication event that was the biggest mutation of them all. It inherited an extra copy of its entire genome. And one of its descendants after that did the same thing all over again so that this animal had four copies more than the invertebrate ancestor. And it is that organism that gave rise to all of the vertebrate species on the planet. And that's why vertebrates have much more complex genomes because they've had these genome duplication events. They have more genes in all their families. And as a result of that, you have more synapse proteins, which give the animals a more complex behavioural repertoire. (**Seth Grant**)

3.0 Intention. Five more cognitive abilities—attention, memory, pattern recognition, learning, and communication—enable intentional actions of the core self, eventually including the delay of reflexes.

3.1 The five cognitive abilities that drive intention do not leave fossil records. But we know from the types of organisms that exhibit them now that they will have emerged and grown since at least the rise of complex multicellularity starting possibly 1.6 billion years ago.

- It's not just detecting danger either — incorporating nutrients, balancing fluids and ions, thermoregulation, reproduction for the species to survive — all of these behaviours exist in animals, but also in single-cell microbes. So, behaviour and even learning and memory do not require nervous systems. (**Post 12**)

- Chemical building blocks provide the ability to process information, which enables the repeatable decisions (cognition) necessary to remain alive. (**Post 20**)
- A candidate mechanism that may serve as the biological basis of the continuum of cognitive function [is] the chemistry of protein networks, whose potential information-processing power and similarity to neural networks in single cells was first described by Cambridge zoologist Dennis Bray, who noticed that "many proteins in living cells appear to have as their primary function the transfer and processing of information, rather than the chemical transformation of metabolic intermediates or the building of cellular structure." (**Lyon**)
- Plant cognition is a field of research directed at experimentally testing the cognitive abilities of plants, including perception, learning processes, memory, and consciousness. Although they lack a brain and the function of a conscious working nervous system, plants are still somehow capable of being able to adapt to their environment and change the integration pathway that would ultimately lead to how a plant "decides" to take response to a presented stimulus. (**Plant Cognition**)
- A plant known as the *Mimosa pudica* was tested for the ability to adapt to closing its leaves upon repeated drops with no apparent harm appointed to the plant. The results showed that with repeated drops, the *Mimosa pudica* eventually stopped closing its leaves or opened its leaves quicker. This behaviour exhibited a trait in which the plant has adapted to not closing, or showing minimal closing, when repeated exposure to a non-harming situation is coupled with its own defence behaviour. (**Plant Cognition**)
- Complex multicellular organisms evolved only in six eukaryotic groups: red algae, green algae, fungi, animals, land plants, and brown algae. (**Multicellular organism**)
- Red algae appeared perhaps 1.6 billion years ago. (**Red algae**)
- Green algae appeared between 1.6 and 1 billion years ago. (**Green algae**)
- Fungi appeared perhaps 1 billion years ago. (**Fungi**)
- Animals appeared between 1 billion and 600 million years ago. (**Animal**)
- Land plants appeared perhaps 500 million years ago. (**Land plants**)
- Brown algae appeared between 200 and 150 million years ago. (**Brown algae**)
- Attention comes in very early in evolution, and over time it becomes more and more complex. There's central attention, sensory attention, more cognitive kinds of attention, and they emerge gradually over this sweep of history from about half a billion years ago up to the present. (**Post 13**)

4.0 Prediction. This level in the hierarchy of consciousness is enabled by mechanisms for the cognitive abilities of anticipation, problem solving, and error detection.

4.1 Once actions become intentional, they and their effects in the world can be modelled so as to improve outcomes and avoid miscues. This appears to only happen in animals with brains that have neuroplasticity and can learn from experience.

- A neuron is called "identified" if it has properties that distinguish it from every other neuron in the same animal and if every individual organism belonging to the same species has one and only one neuron with the same set of properties. In vertebrate nervous systems, very few neurons are "identified" in this sense—in humans, there are believed to be none—but in simpler nervous systems, some or all neurons may be thus unique. In the roundworm C. elegans, whose nervous system is the most thoroughly described of any animal's, every neuron in the body is uniquely identifiable. One notable consequence of

this fact is that the form of the C. elegans nervous system is completely specified by the genome, with no experience-dependent plasticity. (**Nervous Systems**)

- Animals encounter so many unpredictable challenges under natural conditions that it would be very difficult if not impossible for any combination of genetic instructions and individual experience to specify in advance the entire set of actions that are appropriate. But thinking about alternative actions and selecting one believed to be best is an efficient way to cope with unexpected dangers and opportunities. In theory such versatility might result from nonconscious information processing in the brain. But conscious thinking may well be the most efficient way for a central nervous system to weigh different possibilities and evaluate their relative advantages. (**Griffin**)
- Cues are enough to stimulate the behaviour independent of the presence of the stimuli themselves. The representation alone is enough to guide the behaviour. That capacity exists in invertebrates, and on into all vertebrates, e.g. fish and reptiles. When you get to mammals, you have a much more complex form of cognitive representation, where it begins to look deliberative, i.e. the ability to form mental models that can be predictive of things not existing. It's a much more complicated thing than having a static memory of what was there. (**Post 12**)

4.2 Before brains emerged, more sophisticated cognition came from the emergence of faster internal communication systems built using neurons. These first appeared, driven by predation, during the 'Cambrian explosion' approximately 525 million years ago.

- Trails left by the early grazers were straight and simple, but they became more circuitous in later times (550–540 million years ago), and finally showed signs of digging into the substratum by the beginning of the 'Cambrian explosion' of fossils (~540 million years ago). These trails disappeared by 525 million years ago and were replaced by animals with hard coverings shaped into a wide variety of spikes, shells, and plates. The rich array of external armor and weapons in the fossil record strongly suggests that animals started to prey upon each other. The larger size of these animals put a premium on keeping different parts of the body coordinated, and their predatory behavior favored animals capable of making quick movements to obtain food, and to avoid becoming someone else's food. Both demands favored the evolution of a fast-conducting system like neurons. The first clear indication of nervous tissue was the appearance of well-formed eyes and faint outlines of nervous systems in fossils from ~525 million years ago. (**Evolution of neurons**)
- Isomorphic maps are the cornerstone of image-based sensory consciousness. These maps evolved in early vertebrates more than 520 million years ago, and this process was the natural result of the extraordinary innovations of the camera eye, neural crest, and placodes. (**Post 11**)

4.3 Neurons soon bundled together into simple brains, which then developed more features, complexity, and cognitive abilities over the last 520 million years.

- A central, brain-like structure was present in the ancestors of the vertebrates. These primitive, fish-like creatures probably resembled the living lancelet, a jawless filter-feeder. The brain of the lancelet barely stands out from the rest of the spinal cord, but specialised regions are apparent: the hindbrain controls its swimming movements, for instance, while the forebrain is involved in vision. (**A brief history of the brain**)

- As early fish struggled to find food and mates, and dodge predators, many of the core structures still found in our brains evolved: the optic tectum, involved in tracking moving objects with the eyes; the amygdala, which helps us to respond to fearful situations; parts of the limbic system, which gives us our feelings of reward and helps to lay down memories; and the basal ganglia, which control patterns of movements. (**A brief history of the brain**)
- By 360 million years ago, our ancestors had colonised the land, eventually giving rise to the first mammals about 200 million years ago. These creatures already had a small neocortex – extra layers of neural tissue on the surface of the brain responsible for the complexity and flexibility of mammalian behaviour. (**A brief history of the brain**)
- The first big increases in brain size were in the olfactory bulb, suggesting mammals came to rely heavily on their noses to sniff out food. There were also big increases in the regions of the neocortex that map tactile sensations – probably the ruffling of hair in particular – which suggests the sense of touch was vital too. (**A brief history of the brain**)
- Why did we become social? It started when we became warm blooded. Warm blooded creatures need about 10 times more nutrition though. One way to compensate for this requirement was for mammals to develop a new structure in the brain—a cortex—which allowed them to store a tremendous amount of information in the brain and to integrate it. The cortex relied on the subcortical parts of the brain for motivations, sleep/wake patterns, etc., but the cortex allowed for a kind of predictive prowess that had not been seen on the planet before. (**Post 8**)
- Traditionally, scientists believed that the first true warm-blooded animals were mammal ancestors that appeared around 270 million years ago. … The discovery of [a] special kind of bone in *Ophiacodon* fossils is evidence that it could also grow rapidly, which in turn means that it probably had an endothermic metabolism to sustain this growth spurt. *Ophiacodon* lived 300 million years ago, during the Carboniferous period. That was at least 30 million years before the appearance of the first true known mammals, indicating that the furry creatures did not invent warm-bloodedness but rather inherited it from their more reptile-like forefathers. (**Lacerda**)
- After the dinosaurs were wiped out, about 65 million years ago, some of the mammals that survived took to the trees – the ancestors of the primates. Good eyesight helped them chase insects around trees, which led to an expansion of the visual part of the neocortex. (**A brief history of the brain**)
- Mastering the social niceties of group living requires a lot of brain power. Robin Dunbar at the University of Oxford thinks this might explain the enormous expansion of the frontal regions of the primate neocortex, particularly in the apes. … Besides increasing in size, these frontal regions also became better connected, both within themselves, and to other parts of the brain that deal with sensory input and motor control. (**A brief history of the brain**)
- Our conclusion for the moment is, this, that chimpanzees understand others in terms of a perception-goal psychology, as opposed to a full-fledged, human-like belief-desire psychology. (**Call and Tomasello**)
- All of [this brain development] equipped the later primates with an extraordinary ability to integrate and process the information reaching their bodies, and then control their actions based on this kind of deliberative reasoning. Besides increasing their overall intelligence, this eventually leads to some kind of abstract thought: the more the brain processes incoming information, the more it starts to identify and search for overarching patterns that are a step away from the concrete, physical objects in front of the eyes. Which brings us neatly to an ape that lived about 14 million years ago in Africa. It was a very smart ape but the brains of most of its descendants – orangutans, gorillas, and

chimpanzees – do not appear to have changed greatly compared with the branch of its family that led to us. (**A brief history of the brain**)

- Millions of years after early hominids became bipedal, they still had small brains. We can only speculate about why their brains began to grow bigger around 2.5 million years ago, but it is possible that serendipity played a part. In other primates, the "bite" muscle exerts a strong force across the whole of the skull, constraining its growth. In our forebears, this muscle was weakened by a single mutation, perhaps opening the way for the skull to expand. This mutation occurred around the same time as the first hominids with weaker jaws and bigger skulls and brains appeared. (**A brief history of the brain**)
- Once we got smart enough to innovate and adopt smarter lifestyles, a positive feedback effect may have kicked in, leading to further brain expansion. … The development of tools to kill and butcher animals around 2 million years ago would have been essential for the expansion of the human brain, since meat is such a rich source of nutrients. A richer diet, in turn, would have opened the door to further brain growth. Primatologist Richard Wrangham at Harvard University thinks that fire played a similar role by allowing us to get more nutrients from our food. Eating cooked food led to the shrinking of our guts, he suggests. Since gut tissue is expensive to grow and maintain, this loss would have freed up precious resources, again favouring further brain growth. (**A brief history of the brain**)
- Humans (species in the genus Homo) are the only animals that cook their food, and Wrangham argues Homo erectus emerged about two million years ago as a result of this unique trait. Cooking had profound evolutionary effect because it increased food efficiency, which allowed human ancestors to spend less time foraging, chewing, and digesting. H. erectus developed a smaller, more efficient digestive tract, which freed up energy to enable larger brain growth. Wrangham also argues that cooking and control of fire generally affected species development by providing warmth and helping to fend off predators, which helped human ancestors adapt to a ground-based lifestyle. Wrangham points out that humans are highly evolved for eating cooked food and cannot maintain reproductive fitness with raw food. (***Catching Fire: How Cooking Made Us Human***)
- The overall picture is one of a virtuous cycle involving our diet, culture, technology, social relationships, and genes. It led to the modern human brain coming into existence in Africa by about 200,000 years ago. (**A brief history of the brain**)

4.4 As you would expect from the blind trials and errors of evolution, the development of intelligent predictions is not linear. Parallel examples of the emergence of traits are numerous. An excellent example of this is found in cephalopods whose intelligence is utterly alien to us vertebrates.

- *Other Minds* is a 2016 bestseller by Peter Godfrey-Smith on the evolution and nature of consciousness. … Godfrey-Smith's premise in this book is the fact that intelligence has evolved separately in two groups of animals: in cephalopods like octopuses and cuttlefish, and in vertebrates like birds and humans. He notes that studying cephalopods is "probably the closest we will come to meeting an intelligent alien", but that "the minds of cephalopods are the most other of all." (***Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness***)
- Godfrey-Smith disagrees with an old philosophical idea that consciousness suddenly emerged from unthinking matter; it is an active relationship with the world, built up in small steps with separate capabilities for perceiving the world, taking action with muscles, remembering the simplest of events. Such capabilities, in Godfrey-Smith's view, are present in some degree even in bacteria, which detect chemicals in their environment, and

in insects such as bees, which recall the locations of food sources. (***Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness***)

- Since most of the animals' neurons are in their partly-autonomous arms, "for an octopus, its arms are partly self – they can be directed and used to manipulate things. But from the central brain's perspective they are partly non-self too, partly agents of their own. This is as alien a mind as we could hope to encounter." (***Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness***)

5.0 Awareness. This level in the hierarchy of consciousness is enabled by mechanisms for the cognitive abilities of self-reference.

5.1 Awareness is also called reflective consciousness as it involves thinking about thoughts or feelings themselves. These models of thinking about the self may simply arise by turning predictive models of others towards the sensory input of the self.

- It is important to distinguish between perceptual and reflective consciousness. The former, called "primary consciousness" by Farthing (1992), Lloyd (1989), and others, includes all sorts of awareness, whereas the latter is a subset of conscious experiences in which the content is conscious experience itself. Reflective consciousness is thinking, or experiencing feelings, about thoughts or feelings themselves, and it is often held to include self-awareness. (**Griffin**)
- [At first,] the external body is not a subject but an object, and it is perceived in the same register as other objects. Something has to be added to simple perception before one's own body is differentiated from others. This level of representation (a.k.a. higher-order thought) enables the subject of consciousness to separate itself as an object from other objects. We envisage the process involving three levels of experience: (a) the subjective or phenomenal level of the *anoetic* self as affect, a.k.a. first-person perspective; (b) the perceptual or representational level of the *noetic* self as an object, no different from other objects, a.k.a. second-person perspective; (c) the conceptual or re-representational level of the *autonoetic* self in relation to other objects, i.e., perceived from an external perspective, a.k.a. third-person perspective. (**Solms and Panksepp**)

5.2 Like all complex phenomena that evolve over time, awareness does not just "turn on" like a light bulb. There are many intermediate steps of many types of awareness that many different animals likely possess.

- The very difficulty of detecting whether animals experience reflective consciousness should make us cautious about concluding that it is impossible. Most of the suggestive evidence that will be discussed in this book points toward perceptual rather than reflective consciousness. The relation between these two general categories of consciousness can be illustrated by considering a class of intermediate cases, namely, an animal's awareness of its own body—for example, the appearance of its feet or the feeling of cold as a winter wind ruffles its fur. This tends to become an intermediate category between perceptual and reflective consciousness, for an animal might be consciously aware not only of some part of its body but also of what that structure was doing. It might not only feel its teeth crunching on food but also realize that it tastes good. Or it might not only feel the ground under its feet but also recognize that it is running in order to escape from a threatening predator. Furthermore, an animal capable of perceptual consciousness must often be aware that a particular companion is eating or fleeing. This means that it is consciously aware of both the action and of who is performing it. These would all be special cases of

perceptual consciousness. This leads to inquiring how likely it is that such an animal would be incapable of thinking that it, itself, was eating or fleeing. If we grant an animal perceptual consciousness of its own actions, the prohibition against conscious awareness of who is eating or fleeing becomes a somewhat strained and artificial restriction. Furthermore, a perceptually conscious animal could scarcely be unaware of its own enjoyment of eating or its fear of the predator from which it is trying desperately to escape. One could argue that perceptually conscious animals are aware of their actions but not of the thoughts and feelings that motivate them. But emotional experiences are often so vivid and intense that it seems unlikely that when an animal is conscious of its actions it could somehow be unaware of its emotions. (**Griffin**)

5.3 Trace conditioning appears to be a type of learning that requires conscious awareness. We know this from the self-report of humans who do or do not learn during trace conditioning trials. But we see that some animals are also capable of trace conditioning. This is a strong indicator of awareness in non-human animals.

- Robert Clark and Larry Squire published the results of a classical Pavlovian conditioning experiment in humans. Two different test conditions were employed, both using the eye-blink response to an air puff applied to the eye, but with different temporal intervals between the air puff and a preceding, predictive stimulus (a tone). In one condition, the tone remained on until the air puff was presented and both coterminated (delay conditioning). In the other, a delay (500 or 1000 ms) was used between the offset of the tone and the onset of the air puff (trace conditioning). In both conditions, experimental subjects were watching a silent movie while the stimuli were applied, and questions regarding the contents of the silent movie and test conditions were asked after test completion. In the delay conditioning task, subjects acquired a conditioned response over 6 blocks of 20 trials: as soon as the tone appeared they showed the eye-blink response before the air puff arrived. This is a classical Pavlovian response in which a shift is noted from reaction to action, also known as specific anticipatory behaviour. This shift occurred whether subjects had knowledge of the temporal relationship between tone and air puff or not: both subjects who were aware of the temporal relationship—as judged by their answers to questions regarding this relationship after test completion—and subjects who were unaware of the relationship learned this experimental task. One could say that this type of conditioning occurs automatically, reflex-like, or implicitly. In contrast, the trace conditioning task required that the subjects explicitly knew or realized the temporal relationship between the tone and air puff. Only those subjects knowing this relationship explicitly—as judged by their answers to questions regarding this relationship—succeeded in performing the task; those that were not, failed. In other words, subjects had to be explicitly aware or have conscious knowledge of the task at hand in order to bring the shift about, that is, to respond after the tone and before the air puff. This is called explicit or declarative knowledge. Interestingly, amnesia patients could perform the delay conditioning task, but not the trace conditioning task. These patients suffer from damage to the hippocampal formation or medial temporal lobe, suggesting that such an intact structure is a necessary condition for trace but not for delay conditioning to occur. Now what do animals do in this task? Interestingly, the same difference in task procedure and effects of hippocampal lesion is found in, for instance, rabbits: intact rabbits acquire both tasks, hippocampal lesioned rabbits only the delay conditioning task (Clark & Squire, 1998; Wallenstein et al., 1998). So, this would suggest that rabbits—like humans—are aware of the temporal relationship between the stimuli or have conscious knowledge of this temporal relationship and act on this. In other words, it would seem that a classical

Pavlovian task might reveal aspects of awareness or consciousness in animals and "raise[s] the intriguing possibility that delay and trace conditioning could be used to study aspects of awareness in nonhuman animals." (**van den Bos**)

5.4 The most well-known test for conscious awareness is the mirror self-recognition test. Several non-human animal species appear to pass this test, including mammals, birds, and fish. This would indicate that awareness may have arisen as long ago as early vertebrates (which, as noted above, first appeared approximately 525 million years ago during the Cambrian explosion).

- The Mirror Self-Recognition test is the traditional method for attempting to measure self-awareness. However, agreement has been reached that animals can be self-aware in ways not measured by the mirror test, such as distinguishing between their own and others' songs and scents. … Very few species have passed the MSR test. Species that have include the great apes (including humans), a single Asiatic elephant, dolphins, orcas, the Eurasian magpie, and the cleaner wrasse. A wide range of species has been reported to fail the test, including several species of monkeys, giant pandas, and sea lions. (**Mirror Test**)
- Until the 2008 study on magpies, self-recognition was thought to reside in the neocortex area of the brain. However, this brain region is absent in nonmammals. Self-recognition may be a case of convergent evolution, where similar evolutionary pressures result in similar behaviors or traits, although species arrive at them by different routes, and the underlying mechanism may be different. (**Mirror Test**)

5.5 Awareness may have evolved independently in cephalopods too, which would mean it may be very widespread in the animal kingdom.

- Godfrey-Smith follows the neuroscientist Stanislas Dehaene in suggesting that "there's a particular style of processing—one that we use to deal especially with time, sequences, and novelty—that brings with it conscious awareness, while a lot of other quite complex activities do not." The ability of octopuses to learn new skills, of the kind that may demand consciousness, indicates the possibility of "an awareness that in some ways resembles our own." (***Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness***)

6.0 Abstraction. This level in the hierarchy of consciousness is enabled by mechanisms for understanding and creating symbols, art, language, memes, writing, mathematics, philosophy, and science, which all act to expand culture.

6.1 Language is the vital element for abstract consciousness. It is the ability to evoke something that isn't present in the senses by the use of another sound or movement. Its evolutionary origins are unknown and said to be one of the hardest problems in science. It may have originated in humans somewhere between 2.3 and 6 million years ago.

- "I cannot doubt that language owes its origin to the imitation and modification, aided by signs and gestures, of various natural sounds, the voices of other animals, and man's own

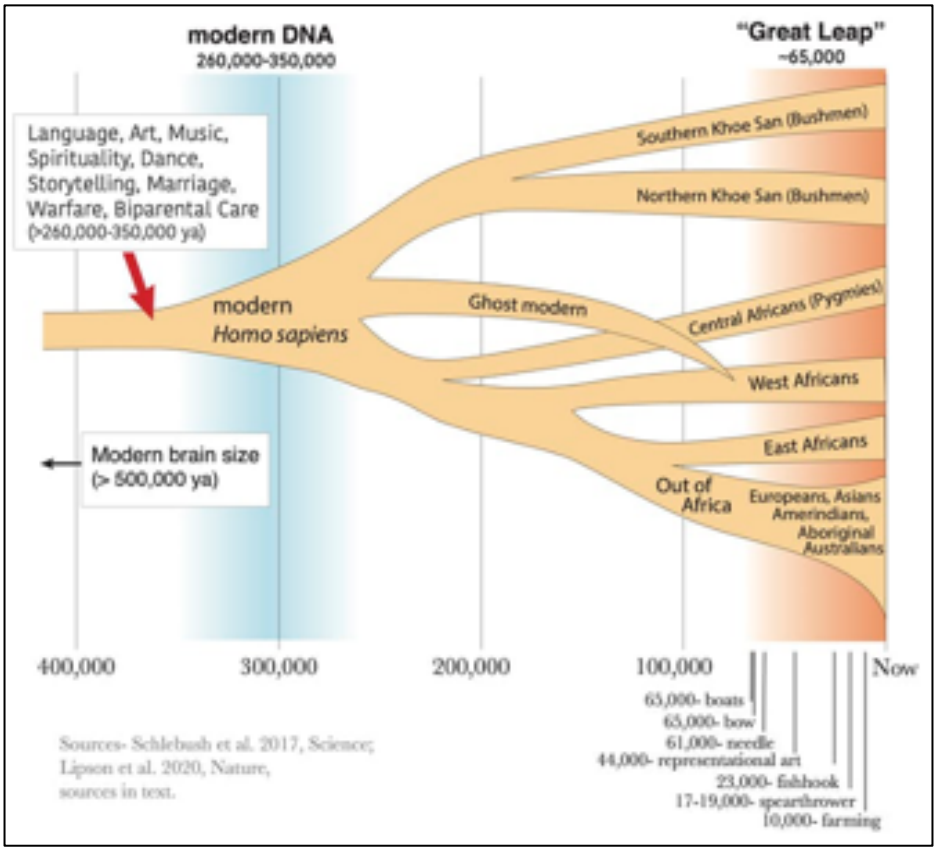
instinctive cries." — *Charles Darwin, 1871. The Descent of Man, and Selection in Relation to Sex* (**Origin of Language**)

- Today, there are various hypotheses about how, why, when, and where language might have emerged. Despite this, there is scarcely more agreement today than a hundred years ago, when Charles Darwin's theory of evolution by natural selection provoked a rash of armchair speculation on the topic. Since the early 1990s, however, a number of linguists, archaeologists, psychologists, anthropologists, and others have attempted to address with new methods what some consider one of the hardest problems in science. (**Origin of Language**)
- How [do we] get from blind genetic evolution to Bach? The first step is synanthropic words. Synanthropic means things that thrive along with humans (e.g. seagulls, cockroaches, etc.). Nobody owned the first words; they were just habits that developed. [E.g. screeching for certain predators or specific dangers.] Next are domesticated words. Domesticated means the reproduction is controlled. For words, this means conscious choosing of one over the other. This leads to differential replication. Meanings or pronunciations can change over time, but the best ones survive, usually without even noticing why. The next step are coined words, deliberately designed, although their survival is still down to selection. Then there are technical terms, which are very carefully designed, and curated under strong group pressure. E.g. phenotype vs. genotype. These are hyper-domesticated words. (**Post 6**)
- The time range for the evolution of language and /or its anatomical prerequisites extends, at least in principle, from the phylogenetic divergence of Homo (2.3 to 2.4 million years ago) from Pan (5 to 6 million years ago) to the emergence of full behavioral modernity some 50,000–150,000 years ago. Few dispute that Australopithecus probably lacked vocal communication significantly more sophisticated than that of great apes in general, but scholarly opinions vary as to the developments since the appearance of Homo some 2.5 million years ago. Some scholars assume the development of primitive language-like systems (proto-language) as early as Homo habilis, while others place the development of symbolic communication only with Homo erectus (1.8 million years ago) or with Homo heidelbergensis (0.6 million years ago) and the development of language proper with Homo sapiens, currently estimated at less than 200,000 years ago. (**Origin of Language**)
- Once early humans started speaking, there would be strong selection for mutations that improved this ability, such as the famous *FOXP2* gene, which enables the basal ganglia and the cerebellum to lay down the complex motor memories necessary for complex speech. (**A brief history of the brain**)

6.2 A host of abstractions emerged as Homo sapiens developed. Language, art, and storytelling all appeared from 260,000 to 350,000 years ago. More advanced abstractions and technologies have emerged steadily since then.

- Bones of primitive Homo sapiens first appear 300,000 years ago in Africa, with brains as large or larger than ours. They're followed by anatomically modern Homo sapiens at least 200,000 years ago, and brain shape became essentially modern by at least 100,000 years ago. (**Longrich**)
- Starting about 65,000 to 50,000 years ago, more advanced technology started appearing: complex projectile weapons such as bows and spear-throwers, fishhooks, ceramics, sewing needles. People made representational art—cave paintings of horses, ivory goddesses, lion-headed idols, showing artistic flair and imagination. A bird-bone flute hints at music. Meanwhile, arrival of humans in Australia 65,000 years ago shows we'd mastered

seafaring. This sudden flourishing of technology is called the "great leap forward," supposedly reflecting the evolution of a fully modern human brain. (**Longrich**)



6.3 The study of language abilities in nonhuman animals is now a fast-growing field. Many examples have recently become available which show that simple forms of language are present in nonhuman animals, meaning the origins of abstract thinking may stretch back much further in time than originally thought.

- The gestural theory states that human language developed from gestures that were used for simple communication. Research has found strong support for the idea that verbal language and sign language depend on similar neural structures. Nonhuman primates can use gestures or symbols for at least primitive communication, and some of their gestures resemble those of humans, such as the "begging posture", with the hands stretched out, which humans share with chimpanzees. (**Origin of Language**)
- In the wild, the communication of vervet monkeys has been the most extensively studied. They are known to make up to ten different vocalizations. Many of these are used to warn other members of the group about approaching predators. They include a "leopard call", a "snake call", and an "eagle call". Each call triggers a different defensive strategy in the monkeys who hear the call and scientists were able to elicit predictable responses from the monkeys using loudspeakers and prerecorded sounds. (**Origin of Language**)
- In experiments on 100 study participants across age groups, cultures, and species, researchers found that indigenous Tsimane' people in Bolivia's Amazon rainforest, American adults and preschoolers, and macaque monkeys all show, to varying degrees, a knack for "recursion", a cognitive process of arranging words, phrases or symbols in a way that helps convey complex commands, sentiments, and ideas. The findings, published

today (26 June 2020) in the journal *Science Advances*, shed new light on our understanding of the evolution of language, researchers said. (**Anwar**)

- In **an online event with Eva Meijer** about her book *When Animals Speak*, Meijer discussed studies that show dolphins and bats use names for each other and chickens even create names for the people they regularly interact with.
- Jays and crows choose particular gifts they believe will appeal to their partners, and so have a "theory of mind" — they can see things from another's point of view. Prairie dogs use chattering calls to describe different intruders — not only a human, but how large he or she is, the colour of their clothes, and whether they are carrying an umbrella or a gun. Many mammals can learn human words, produce new sounds, or acquire other languages: orcas, for example, can imitate the cries of dolphins. (**Meijer**)
- Neo-Piagetian stages have been applied to the maximum stage attained by various animals. For example, … pigeons attain the sensory motor stage, forming concepts. (**Piaget**)
- Birds are descendants of the primitive avialans which first appeared about 160 million years ago in China. (**Bird**)

6.4 Spoken languages among animals are therefore a difference of degree rather than kind. Written language, however, appears to be a uniquely human phenomenon. This capability in humans first emerged less than 6,000 years ago and would appear to be the technology that is most responsible for our cultural evolution accelerating to the point that humans now dominate the planet.

- Cuneiform is an ancient writing system that was first used in around 3400 BC. Distinguished by its wedge-shaped marks on clay tablets, cuneiform script is the oldest form of writing in the world. (**Origin of Language**)
- The MacCready Explosion: 10,000 years ago, human population plus livestock and pets were approximately 0.1% of terrestrial vertebrate biomass. Today, it is 98%. This is probably the biggest, fastest, biological change on the planet ever. Genes don't explain it. Technology does. (**Post 6**)

6.5 This completes the final Evolutionary Epistemological Mechanisms (EEMs) from Donald Campbell, which have been slowly accruing during this evolutionary history.

- Campbell settled on a 10-step outline that showed the broad categories of mechanisms that biological life has used to gain knowledge. This starts with the earliest origins of life where problems were solved over generations through mere genetic variance alone, without any aids from motion or the formation of memories. This earliest slow accrual of genetic knowledge eventually led, according to Campbell, to the other mechanisms: movement, habit, instinct, visually-supported decisions, memory-supported decisions, observational learning from social interactions, language, cultural transmissions, and finally, scientific accumulations of knowledge. (**Gibney**)

Brief comments to close

This may have been the hardest of the 4 Tinbergen questions to answer. Our scientific explorations into this realm have left us with vast ranges for when different cognitive abilities may have emerged and slowly grown. But summarising the findings above, and focusing simply on the emergence of each level, we find this final chart for my hierarchies of consciousness:

So, that's just about it in this long series on consciousness. Once again, we have excellent consilience where multiple streams of evidence are all pointing to the same thing. All that is left for me now is to put together one final summary and see how my new definition and mapping of the features of consciousness might help us answer the most stubborn problems the field has previously acknowledged. I can hardly wait.

23 — Summary of My Evolutionary Theory

Time to look over everything.

22 February 2021

I started this series on the 15th of March 2020. Looking back, I thought I had an ambitious plan then, but it turned out it wasn't nearly enough. After thousands of written words (and probably thousands of pages read), I'm finally ready to present a summary of my current evolutionary theory of consciousness. I'll start with the background of my metaphysical hypotheses. Then I'll recount the theories that I like best for the two biggest mysteries for this topic — the hard problem of consciousness and the emergence of life. I'll note how new forces emerge once life enters the universe. Then, this will put me in a position to state an expansive definition of consciousness that fits with all of those pieces of the puzzle. And finally, I'll finish with a bit more depth on the details and definitions in the theory so that it's as clear as I can make it in one essay.

I thought I'd be wrapping up this series with this summary, but my research shows that I really ought to have at least one more post after this to go over the traditional objections to a materialist account of consciousness. If my stated theory can answer all of those, then I'll at least have a case for a coherent (if not correct) theory. On to the summary!

Preface — Epistemological and Metaphysical Background

According to an evolutionary worldview, the universe is always changing, we cannot see what the future will bring, and one can never get outside of it all to gather objective facts about the true state of the world. Therefore, **knowledge cannot be justified, true, belief**, as Plato thought. Knowledge can only ever be justified, beliefs, that are currently surviving our best rational tests. Knowledge is always provisional, and there is no bedrock upon which certainty can rest. (For now. Even that fact isn't known for sure.)

To live, we must act. To decide upon actions from this state of fundamental ignorance, we must start with hypotheses and then test them. In evolutionary philosophy, my first tenet is the

first hypothesis that is necessary to get us off the ground and running.

1. We live in a rational, knowable, physical universe. Effects have natural causes. No supernatural events have ever been unquestionably documented.

Through the eons of the entire age of life, and over all the instances of individual organisms acting within the universe, the ability of life to predict its environment and continue to survive in it has required that the universe must be singular, objective, and knowable. If it were otherwise, life could not make sense of things and survive here. We may never know if that is absolutely true, but so far that knowledge has *survived*. The objective, physical, natural, material existence of the universe may indeed be an assumption, but as a starting point, it seems to be the strongest knowledge we have. All previously uncovered mysteries have not changed this fact, so rationality dictates that we ought not to abandon it without good cause.

Before life emerged, all evidence points to a universe made of matter that interacted according to the fundamental laws of physics and then chemistry. All objects were affected by forces, but there were no subjects, minds, intentions, or consciousness. So, how did we get **from physics to chemistry to biology**? And how might consciousness suddenly appear during that time?

Hypotheses on the Mysteries of Abiogenesis and The Hard Problem

First, let's look at the emergence of life. We may never know for sure how it actually happened. There may never be a way to find conclusive evidence or rule out all but one possibility. But the hypothesis that is a leading contender and makes the most sense to me right now is known as the "RNA World" hypothesis. Nobel Prize winner Jack Szostak's work on this is explained very simply in a short video called **The Origin of Life**, and there is a much longer **video series** on the topic as well. I've covered this in more depth in **a previous post**, but a quick overview is that polymer chains and membranes form spontaneously in the environment and it's very plausible to see how a simple 2-component system might form that can eat, grow, contain information, replicate, and evolve, simply through thermodynamic, mechanical, and electrical forces. That would kickstart evolution, which means the development of life would be off and running towards the present day. No ridiculous improbabilities are needed, no supernatural forces, and no lightning striking a mud puddle. Just chemistry and mechanical activity.

This leads us to a simple explanation for life, which is defined as something that preserves, furthers, or reinforces its existence in the given environment. There are **seven traits** currently considered to enable this kind of self-prolonged existence: organisation, growth, reproduction, response to stimuli, adaptation, homeostasis, and metabolism. We can now see how physics plus chemistry plus natural selection might have led to all of these traits. But what about consciousness?

This is basically **the hard problem** as coined by David Chalmers. We have subjective experience. Evolutionary studies have shown us that there is an unbroken line in the history of life. But water and rocks don't appear to have anything like consciousness. So, how can inert matter ever evolve into the subjective experience that we humans undoubtedly feel?

Chalmers has proposed that subjective experience may be a fundamental property of the universe, like the spin of electromagnetism. I have come to accept that as a likely hypothesis. All matter is affected by the forces of physics and chemistry. But until that matter is organised

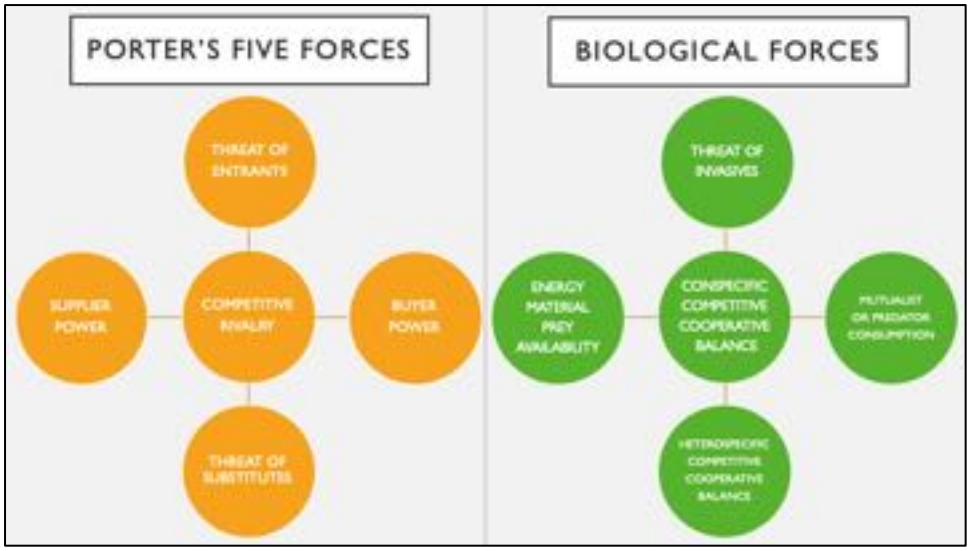
into a living subject that is capable of responding to those forces in such a way as to remain alive, it makes no sense to talk of non-living matter as 'feeling' or 'experiencing' those forces. Inert matter has no structure capable of living through subjective activities. Panpsychism claims that minds (*psyche*) are everywhere, and they don't need physics and matter to exist. But this raises innumerable difficulties, including an enormous change to one's metaphysics that supposedly cannot be detected by science. What I hypothesise instead is that the forces of physics are everywhere, and it is a fundamental property of the universe that these forces are felt subjectively when subjects emerge. Since the Greek for force is *dynami*, I would say the universe has *pandynamism* rather than *panpsychism*. The psyche only originates and evolves along with life. This psyche expands as the living structures expand their capabilities of sensing and responding to these forces. And the 'flavour' of experiences within this psyche are utterly dependent upon the underlying mechanisms of which particles of matter are being subject to which particular forces.

(For example, the retch of disgust from accidentally eating something harmful maps almost exactly onto the retch of moral disgust from accidentally witnessing something beyond the pale such as a mutilated dead body. These experiences come from very different sources, and they process very different bits of information, so we might expect them to feel very different, but we know from neuroscience that the brain has duct-taped the feelings of moral disgust onto the existing architecture for gustatory disgust and that is what explains the similar conscious experience. This is another striking bit of support for a materialist understanding of consciousness.)

A New Category of Forces Arise

So, the theories of the RNA World and pandynamism get us from the inert landscape of the early universe to the rich and vibrant present-day world with biology and subjective experience. Physicalism still holds as a viable hypothesis for the metaphysics of the universe if these theories are correct. And this view makes it clear that something else also emerges with the emergence of life. Given that living things are (to the best we know) merely structures of matter that have come to be organised so as to be self-sustaining and self-replicating, two new categories of things in the world appear which are related to that: 1) things that help life stay alive, and 2) things that harm life from staying alive. That division has no meaning in physics or chemistry, but they are fundamental once biology emerges. Through the trials and errors of natural selection, living systems become sensitive to these positive and negative aspects of the world and they respond accordingly. In science, something exists to the extent that it exerts causal power over other things. Gravity exists because it exerts power over mass. Electricity exists because it exerts power over charged particles. Similarly, the power these categories exert over living things implies they exist too. I call them 'biological forces'.

So, what do these biological forces look like? My conception is that they look like Porters Five Forces from the business world, which maps the competitive and cooperative forces that affect any organisation as it tries to stay profitable (aka alive) in its industrial ecosystem. I hypothesise that this framework, taught in MBA curricula around the world, actually works because it is a fractal of the competition and cooperation that all life must navigate in its own ecosystems. These forces can be depicted side by side to make the analogy clear.



Therefore, in biology, there are 1) battles for consumption of upstream inputs of energy, material, or prey (*a la* suppliers); 2) battles for consumption of downstream outputs by mutualists, micro- or macroscopic predators (*a la* buyers); 3) battles with potentially invasive species (*a la* threat of entrants); 4) battles with current niche competitors from heterospecifics in other species (*a la* substitutes); and 5) the balance between competition and cooperation among conspecifics from the same species (*a la* competitive rivalry).

In the great interrelated web of life, any individual or species can play any of these parts depending on how you define the circle around an ecosystem for analysis. We all get eaten at some point. If biological behaviour is determined, it is not by the laws of physics and chemistry, but by the unwritten laws of these biological forces. However, just as the complexity in the system makes Porter's Five Forces seemingly impossible to calculate with precision, this is even more true for anyone hoping to calculate outcomes from biological forces. Still, we can illustrate them and discuss their relative strengths to aid in analysis and understanding of life and its choices. This brings us to a place where we can now propose a new definition for consciousness.

An Evolutionary Theory of Consciousness

In my **(sorta) brief history of the definitions of consciousness**, I noted that previous attempts stretch all the way from consciousness being something as small as "the private, ineffable, special feeling that only we rational humans have when we think about our thinking," right on down to it being "a fundamental force of the universe that gives proto-feelings to an electron of what it's like to be that electron." That's why the Wikipedia entry on consciousness notes:

"The level of disagreement about the meaning of the word indicates that it either means different things to different people, or else it encompasses a variety of distinct meanings with no simple element in common."

I believe the shape of a proper answer comes from Dan Dennett's 2016 paper "**Darwin and the Overdue Demise of Essentialism**," where he said:

"We should quell our desire to draw lines. We can live with the quite unshocking and unmysterious fact that there were all these gradual changes that accumulated over many millions of years. … The demand for essences with sharp boundaries blinds thinkers to the prospect of gradualist theories of complex phenomena, such as life,

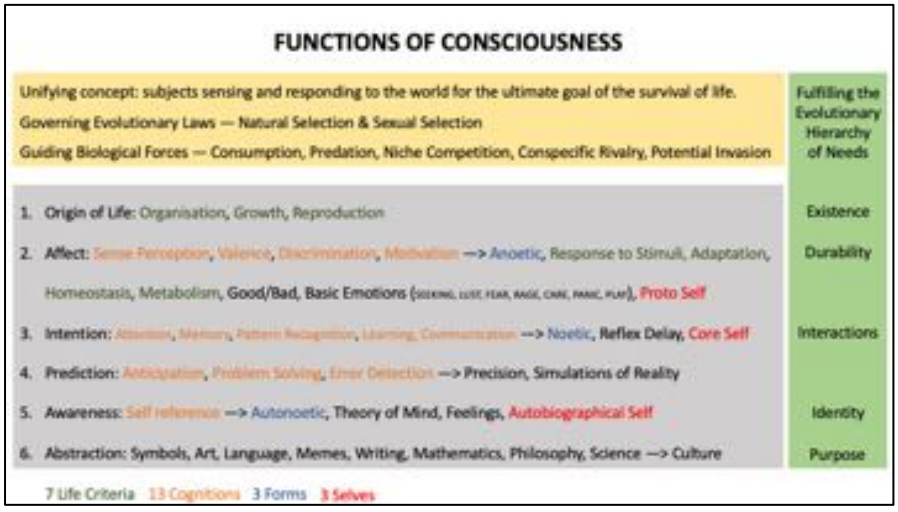
intentions, natural selection itself, moral responsibility, and consciousness."

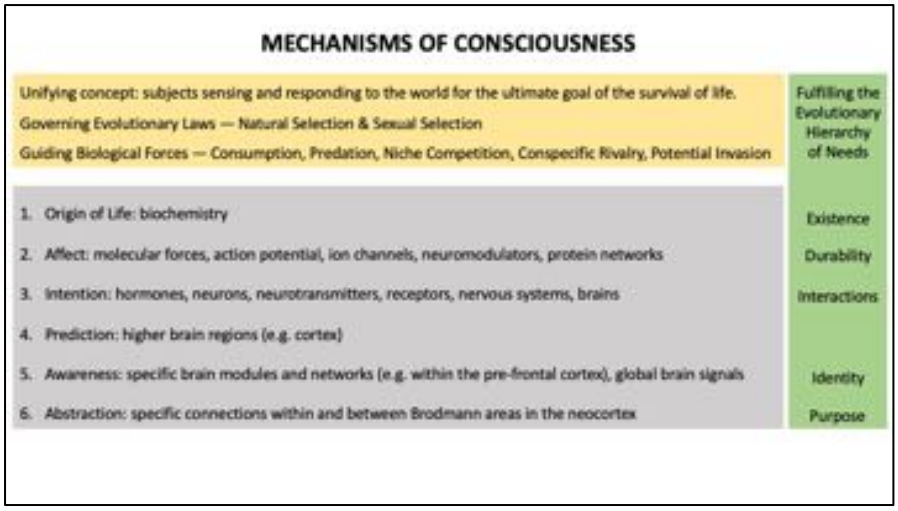
Indeed. But based on the story of abiogenesis outlined above, I think that a natural joint to carve a philosophical place for consciousness is in the biological realm. The emergence of life is sufficiently hazy and fuzzy in its origin so as to cast doubt on any overly specific claim that one particular molecular structure came together and suddenly turned consciousness on like a light switch. But no one needs to find such a **grain of sand that turned a pile into a heap**. We're looking for a gradualist theory of the complex phenomena of consciousness, and its development along with life fits that bill.

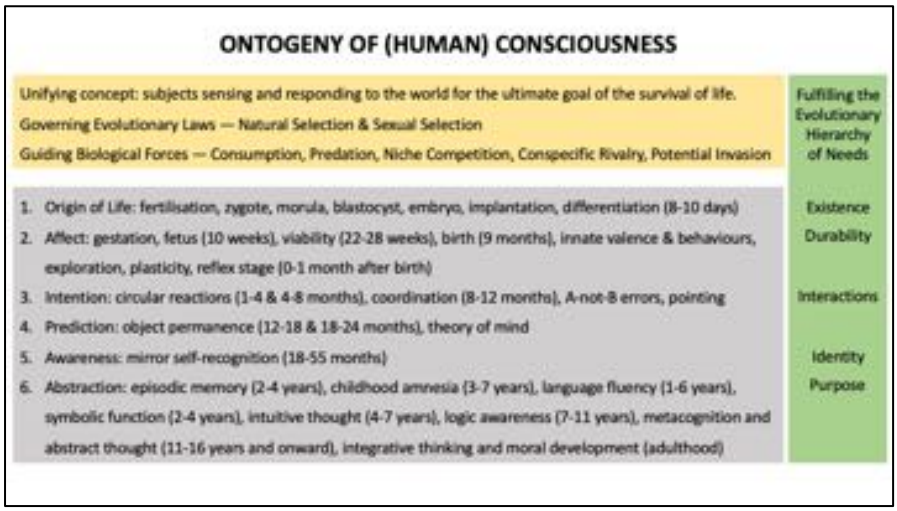
This binding of consciousness to life also fits the etymological root of the word. The English word 'conscious' originally derived from the Latin *conscius* where *con* meant 'together' and *scio* meant 'to know'. According to this literal interpretation, to be conscious would be 'to know', which requires a knower. And to 'know together', this conscious thing would need to know at least two things. Do sub-atomic particles feeling fundamental forces meet these criteria? No. Do elements from the periodic table feeling intermolecular forces meet these criteria? Also no. Do living things feeling biological forces meet these criteria? Yes. Once chemistry makes the jump to biology, the resulting proto-life forms have a defined self *and* they begin to compete for resources with other potential entrants, substitutes, or conspecifics in order to self-replicate and survive. They react to the world as if they know what they are *and* what they need. Thus:

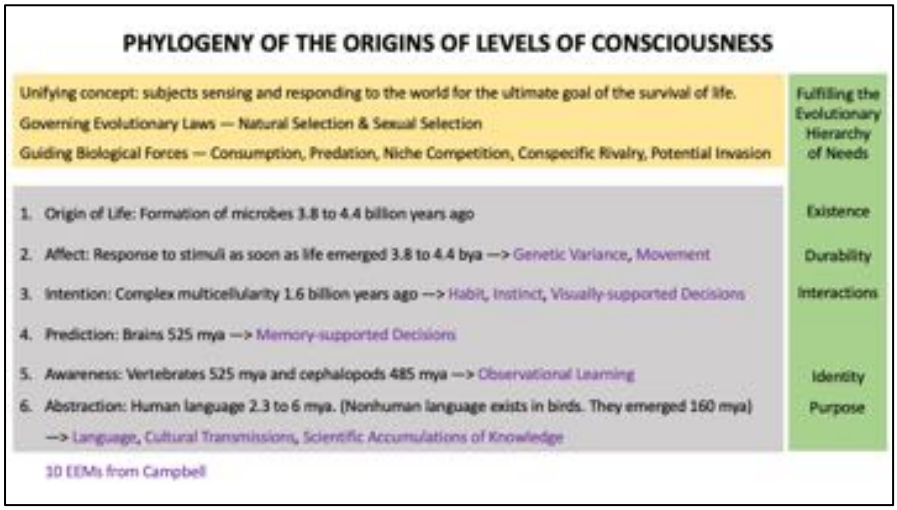
Consciousness, according to this evolutionary theory, is an infinitesimally growing ability to sense and respond to any or all biological forces in order to meet the needs of survival. These forces and needs can vary from the immediate present to infinite timelines and affect anything from the smallest individual to the broadest concerns (both real and imagined) for all of life.

Such a definition accords with our intuitions to exclude non-living matter from consciousness studies. Rocks and water just don't respond to any threats to their existence. But all living things do. And in an incredibly wide and diverse manner. In order to map the contours of such a broad definition, I spent several posts conducting a **Tinbergen analysis** of the **functions**, **mechanisms**, **ontogeny**, and **phylogeny** of consciousness, which is the standard procedure in evolutionary studies for coming to know all of the elements of any biological phenomenon. That massive review resulted in the following four charts:









For more details on each of these charts, read the individual posts where they were developed, which altogether give a full picture of the various aspects of consciousness. The first tier in this hierarchy — **1) Origin of Life** — has already been discussed above. The remaining tiers are:

2) Affect: This is the valence, tone, or mood that is capable of distinguishing differences between good stimuli as opposed to bad ones, which results in responses of graduated arousal

and intensity. **Mark Solms** calls this the primary experience and purpose of consciousness. He asks, rhetorically, how can affective arousal (i.e., the arousal of feeling) go on without any inner feel? It cannot. This accords with my theory of pandynamism, where such feelings are felt subjectively as soon as subjects appear and are affected by biological forces. At first, these affects will generate what we think of as instinctual unconscious reactions. These can involve any or all of **Jaak Panksepp's** seven basic emotions (in capital letters to denote a distinction between them and their common usage): SEEKING, FEAR, RAGE LUST, PANIC, CARE, and PLAY. Later, once many more structures have evolved, these affects can be registered, and eventually named, in conscious awareness.

3) Intention: This development in consciousness marks the ability of one reaction to interrupt or override others within an organism. From the perspective of an outside observer, choices appear to be made and there is a narrative sequence to life. Like affect, this can take place unconsciously within humans, so presumably it can in other forms of life as well, but it does empirically exist in very simple life using cognitive abilities such as memory, pattern recognition, and learning. Much later in evolutionary history, this can also be accessed and rationally considered in order to create extremely complex and far-ranging intentions.

4) Prediction: Once intentions exist (either one's own or the intentions of others), the next development in consciousness is to take them into account by predicting how intentions will interact with the world. Organisms no longer just respond to the present by building up memories of the past; they begin to guess the future too. This appears to happen only in animals with brains that have neuroplasticity and can learn from experience. It also would seem that predictions about the intentions of others are particularly vital, which would explain why neurons and brains appear to have emerged during the Cambrian explosion due to the onset of predation. The success or failure of one's predictions about their predators or prey would have been a powerful driver for change in any arms race occurring in this new dimension of consciousness. Surprise and uncertainty would be a bad emotion for any prediction, which would eventually help to hone the development of feelings of precision to extremely high levels.

5) Awareness: The next level of consciousness comes in now that structures have evolved to trigger affective emotions in the present (level 2), evaluate the past to make complex choices (level 3), and predict further and further into the future what the actions of the self and others may result in (level 4). The interaction and comparison of these three phenomena allows for the dawning of awareness of a self that is different from others. The richness of this distinction grows with the number of sensations that are able to be evaluated against one another within more and more sophisticated models of elements of the world. Studies have shown that conscious awareness is indeed necessary for some types of learning that give organisms additional plasticity to respond to new and novel stimuli in their environment, thus cementing the evolutionary advantage of gaining and retaining this ability.

6) Abstraction: The final level of consciousness in this hierarchy comes when models of reality go beyond mere direct representation and begin to use symbolic representations to evoke, communicate, and manipulate thoughts and feelings about the world. While nonhuman animals have displayed rudimentary or latent abilities for abstraction, the emergence and development of this capability in humans has been of such enormous import that it is considered the latest of **the major transitions of evolution**. Symbols, art, and language have driven the cultural evolution of memes, writing, mathematics, philosophy, and science that make up all of the powerful products of human culture. The causes for the emergence of this type of consciousness are mysteriously shrouded in the history of one

species at the moment, but there is no denying the power, for good or for ill, that this has enabled. May our fuller grasping of the biological forces that affect the consciousness of all of life motivate us to realise what good is and bring it into fruition.

Related Definitions

To wrap up this discussion, and to help avoid some confusion, here are a few definitions of common terms in consciousness studies which sometimes differ between technical and folk usages. Where such differences exist, I have chosen a definition that best fits with my concept of consciousness as outlined above.

Accessible: This adjective is used to refer to the contents of consciousness that humans are able to recall and report upon. It is contrasted against the unconscious and inaccessible contents that may still drive behaviour.

Attention: According to **Michael Graziano** and his attention schema theory, attention is the basic ability of a nervous system to focus on a few things at a time and process them deeply. Some forms of attention go back possibly all the way to the beginnings of nervous systems. Graziano thinks attention comes in very early in evolution, and over time it becomes more and more complex. There's central attention, sensory attention, more cognitive kinds of attention, and they emerge gradually over this sweep of history from about half a billion years ago up to the present. Global Workspace Theory says attention is achieved by attending to signals as they become stronger and stronger compared to other signals. At some point, the signals become so strong that they reach a state called 'ignition' when they can then influence wide networks around the brain. Once that occurs, we humans can talk about that signal, we can move toward it, and we can remember it later.

Bottom-up vs. Top-down: In neuroscience, **these terms** describe specific directions of information processing. Sensory input is typically considered bottom-up, while higher cognitive processes, which have more information from other sources, are considered top-down. A bottom-up process is characterized by an absence of higher-level direction, whereas a top-down process is driven by other cognition, such as from goals or targets. In reality, there is a multi-directional feedback loop among these systems, and any talk of top-down control should not be meant to signify a **homunculus** in the brain, a designer from on high, or a **skyhook** that acts independently from all other causes.

Cognition: **Pamela Lyon** lists 13 functional abilities of cognition that help organisms adapt to their environment. I have found that these are distributed throughout my hierarchy of consciousness and they have developed in a logical fashion that is also supported by empirical evidence from evolutionary history. These are: (1) **sense perception** — ability to recognize existentially salient features of the external or internal milieu; (2) **affect** — valence, attraction, repulsion, neutrality / indifference (hedonic response); (3) **discrimination** — ability to determine that a state of affairs affords an existential opportunity or presents a challenge, requiring a change in internal state of behaviour; (4) **motivation** — teleonomic striving; implicit goals arising from existential conditions; (5) **attention** — awareness, orienting response; ability to selectively attend to aspects of the external and/or internal milieu; (6) **memory** — retention of information about a state of affairs for a non-zero period; (7) **pattern recognition** — intentionality, directedness towards an object; (8) **learning** — experience-modulated behaviour change; (9) **communication** — mechanism for initiating purposive interaction with conspecifics (or non-conspecific others) to fulfil an existentially salient goal; (10) **anticipation** — behavioural change based on

experience-based expectancy (i.e. if X is happening, then Y should happen), possibly evolved across generations, and which is implicit to the agent's functioning; (11) **problem solving** — decision making, behaviour selection in circumstances with multiple, potentially conflicting parameters and varying degrees of uncertainty; (12) **error detection** — normativity, behavioural correction, value assignment based on motivational state; and (13) **self-reference**— mechanisms for distinguishing "self" or "like self" from "non-self" or "not like self".

Communication: This is **described as**, "an act of interchanging ideas, information, or messages, via words or signs, which are understood to both parties. Every living thing communicates in some way. Fish jump, sometimes for sheer joy. Birds sing their cadences to communicate a variety of purposes. Dogs bark, cats meow, cows moo, and horses whinny. These noises, or other interactions, communicate or transfer information of some kind. Communication is, at its core, a two-way activity, consisting of seven major elements: sender, message, encoding, channel, receiver, decoding, and feedback." Note that this is distinct from language. See the definition of language below for comparison.

Conscious vs. Unconscious: These terms are generally used more like medical descriptions, corresponding to awake and aware (conscious) vs. asleep or unresponsive (unconscious). They can both be described using various scales of physical attributes (cf. the **Glasgow Coma Scale**) and both fit within my hierarchy of consciousness as being more or less able to sense and respond to the needs of remaining alive. (Note that I generally stay away from Freud's usage of the unconscious mind as it is a mixed bag of insight and imagination that would take a lot of patience to unpack.)

Emotions vs. Feelings: In **my previous writings on emotion**, I noted that this is a complex psychophysiological experience where an individual's state of mind interacts with biochemical (internal) and environmental (external) influences. Emotions can be seen as mammalian elaborations of general vertebrate arousal patterns, in which neurochemicals (for example, dopamine, noradrenaline, and serotonin) step-up or step-down the brain's activity level, as visible in body movements, gestures, and postures. An influential theory of emotion is that of Lazarus: emotion is a disturbance that occurs in the following order: 1) cognitive appraisal—the individual assesses the event cognitively, which cues the emotion; 2) physiological changes—the cognitive reaction starts biological changes such as increased heart rate or pituitary adrenal response; 3) action—the individual feels the emotion and chooses how to react. Lazarus stressed that the quality and intensity of emotions are controlled through cognitive processes. I think these descriptions cover the broad evolutionary emergence and growth of affective feeling from the most basic valence in organisms even simpler than bacteria all the way to the sophisticated naming and therapeutic modification of human moods. **Antonio Damasio** tries to separate emotions from feelings, saying emotions are chemical reactions, while feelings are the conscious experience of emotions. This is overly confusing and unnecessary to me, and apparently Damasio is not always consistent with this usage either. If you consciously feel an emotion, you feel affect (level 2 in my hierarchy) and you have awareness (level 5 in my hierarchy). I find that easier to understand.

Evolutionary Hierarchy of Needs: When I define consciousness as the ability to sense and respond to any or all biological forces in order to meet the needs of survival, these are the needs that I am talking about. For full details, see my article about **Replacing Maslow with an Evolutionary Hierarchy of Needs**, but here are some important points to consider. There are many ways that the ultimate question of survival can be determined, and life has been slowly learning to sense and understand these over billions of years. For example, there

are so many things that can kill you, your genes, your kin, or your species, and they can all do so in the immediate, medium, or very long term. Living organisms that can sense and respond to more and more of these threats are the ones that will last and emerge over time. Such organisms will sense many, many *needs* to meet all of the threats (and exploit all of the opportunities) in its environment. Each living organism's unique genetic, environmental, and evolutionary histories are constantly leading to changes in the relative strengths of these needs, but at no point does something outside of the physical realm enter into the equation. All of these needs can be described through physical properties, even if the magnitude of their felt force cannot yet be calculated. The ever-growing list of threats and opportunities is why the needs of life are ever-growing too. The psychologist Abraham Maslow studied these for individual humans and produced his famous Hierarchy of Needs. I have generalised these and adapted them to apply to all of life, thereby producing something I call an Evolutionary Hierarchy of Needs. Starting at the bottom of Maslow's pyramid, we see that the 'physiological' needs of the human are merely the brute ingredients necessary for 'existence' that any form of life might have. In order for that existence to survive through time, the second-level needs for 'safety and security' can be understood as promoting 'durability' in living things. The third-tier requirements for 'love and belonging' are necessary outcomes from the unavoidable 'interactions' that take place in our deeply interconnected biome of Earth. The 'self-esteem' needs of individuals could be seen merely as ways for organisms to carve out a useful 'identity' within the chaos of competition and cooperation that characterises the struggle for survival. And finally, the 'self-actualisation' that Maslow struggled to define could be seen as the end, goal, or purpose that an individual takes on so that they may (consciously or unconsciously) have an ultimate arbiter for the choices that have to be made during their lifetime. This is something Aristotle called '*telos*'. Taken as a whole, these are the needs that life must evolve to become more and more conscious of if it is to survive over longer and longer spans of time.

Evolutionary Epistemology Mechanisms: As part of **Donald Campbell's** work defining the field of evolutionary epistemology, he settled on a 10-step outline that showed the broad categories of mechanisms that biological life has used to gain knowledge. I have found that these fit well within my hierarchy and in the same order along with my map of the phylogenetic history of consciousness (see chart above). These EEMs start with the earliest origins of life where problems were solved over generations through mere genetic variance alone, without any aids from motion or the formation of memories. That earliest slow accrual of genetic knowledge eventually led, according to Campbell, to the other mechanisms: movement, habit, instinct, visually-supported decisions, memory-supported decisions, observational learning from social interactions, language, cultural transmissions, and finally, scientific accumulations of knowledge.

Exteroception vs. Interoception: **Exteroception** is any form of sensation that results from stimuli located outside the body and is detected by exteroceptors, including vision, hearing, touch or pressure, heat, cold, pain, smell, and taste. **Interoception** is any form of sensation arising from stimulation of interoceptors and conveying information about the state of the internal organs and tissues, blood pressure, and the fluid, salt, and sugar levels in the blood.

Intentionality vs. Intentional Stance: **Intentionality** is a technical term in philosophy that was introduced by Franz Brentano in the last quarter of the nineteenth century. It should not be confused with the ordinary meaning of the word intention. While an intention is just an internal aim or goal, intentionality refers to mental directedness towards objects, as if the mind were a bow whose arrows could be properly aimed at different targets. It is also

sometimes referred to as aboutness. On the other hand, the intentional stance has been defined by Daniel Dennett as an understanding that others' actions are goal-directed and arise from particular beliefs or desires. It is intentionality aimed at subjects. The understanding of others' intentions is a critical precursor to understanding other minds. Since the seminal (1978) paper by primatologists David Premack and Guy Woodruff entitled "Does the chimpanzee have a theory of mind?", much empirical research has been devoted to the question of whether non-human primates can ascribe psychological states with intentionality to others. **Call and Tomasello** concluded in 2008 that chimpanzees understand others in terms of a perception-goal psychology, as opposed to a full-fledged, human-like belief-desire psychology. This is an interesting distinction in the way that minds may work.

Involuntary vs. Voluntary: In biology, **involuntary** control refers to bodily activity "not under the control of the will of an individual." These involuntary responses by muscles, glands, etc., occur automatically when required; many such responses, such as gland secretion, heartbeat, and peristalsis, are controlled by the autonomic nervous system and effected by involuntary muscle. **Voluntary** muscles, by contrast, are under our conscious control so we can move these muscles when we want to. These are the muscles we use to make all the movements needed in physical activity. Note that these two physiological terms are not concerned with the question of free will and whether 'conscious control' is ultimately under our control or not. (That is another large topic in metaphysics for another time.)

Language: This is **defined** as "a distinctly human activity that aids in the transmission of feelings and thoughts from one person to another. It is how we express what we think or feel —through sounds and/or symbols (spoken or written words), signs, posture, and gestures that convey a certain meaning. The purpose of language is making sense of complex and abstract thought. Whereas communication is an experience, language is a tool." Language allows for much greater scale and scope in cognition. It increases our ability to make sense of the world compared to working memory alone. It vastly enlarges the recognition of patterns in the world. And language enables deep and precise exploration of the self and the world around us. The power of language is perhaps best displayed by Hellen Keller who did not always have it. She said, "Before my teacher came to me, I did not know that I am. I lived in a world that was a no-world. I cannot hope to describe adequately that unconscious, yet conscious time of nothingness. (…) Since I had no power of thought, I did not compare one mental state with another."

Mind: The **mind** is "the set of faculties including cognitive aspects such as consciousness, imagination, perception, thinking, intelligence, judgement, language and memory, as well as noncognitive aspects such as emotion and instinct. Under the scientific physicalist interpretation, the mind is produced at least in part by the brain. The primary competitors to the physicalist interpretations of the mind are idealism, substance dualism, types of property dualism, eliminative materialism, and anomalous monism. There is a lengthy tradition in philosophy, religion, psychology, and cognitive science about what constitutes a mind and what are its distinguishing properties." In this series, I have done my best to describe and defend a physicalist interpretation of all of these aspects of mind.

Qualia vs. Something-it-is-like vs. Subjective Experience: The term **qualia** derives from the Latin adjective *qualis* meaning "of what sort" or "of what kind" in a specific instance, such as "what it is like to taste a specific apple, this particular apple now." There are many definitions of qualia, but one of the simpler and broader definitions is: "The 'what it is like' character of mental states. The way it feels to have mental states such as pain, seeing red, smelling a rose, etc." This 'what it is like' is also a reference to Thomas Nagel's paper **What**

is it Like to Be a Bat? in which Nagel famously asserts that "an organism has conscious mental states if and only if there is something that it is like to be that organism—something it is like for the organism." In other words, it is the experience of being a subject, hence the other term for this phenomenon as subjective experience. The supposed ineffableness of qualia, the purported inability to describe "the redness" of a rose, is completely effable within the evolutionary theory of consciousness presented above. The hard problem of why the experience happens at all is assumed just to be a fundamental property of the universe, which arises in subjects once they evolve the structure to sense and respond to stimuli. After that, the "redness" is completely described by the Tinbergen analysis which shows the adaptive functions of seeing red, the mechanisms involved in sensing wavelengths of light in the red spectrum, the general phylogenetic history of how sensing red has evolved in our species, and the specific ontogenetic history of personal experience that the individual has had with different intensities of redness during their life. What else is left to explain?

Conclusion

For thousands of years of human history, including several hundred after the scientific revolution, the existence and diversity of life was a mystery because evolution and the processes of natural selection were unknown. Once Darwin gathered the evidence to make his case for the theory of evolution, much of that mystery evaporated and any hazy fog that obscured what life is all about has been slowly evaporating with more and more scientific exploration. Within such research, consciousness has remained behind a stubborn patch of murkiness, even after several decades of dedicated **consciousness studies**. Perhaps this has remained so because of the invisibility of biological forces (like the proverbial water surrounding a fish). Or perhaps it was because consciousness as a fundamental part of the physical universe (like gravity or electromagnetism) just hasn't been accepted or explained via a hypothesis like pandynamism. Or perhaps consciousness just hasn't been properly illuminated by a comprehensive analysis using Tinbergen's framework for all biological phenomena. Now that I have gone through all three of these additions, however, perhaps the view of consciousness might finally become a bit clearer.

What do you think? Does this theory of consciousness make sense to you? What questions has it left unanswered? In my final post, I will check these ideas against the traditional objections to physicalist conceptions of consciousness, but please share your own in the comments so I might consider them as well.

IMPLICATIONS FOR THE CONCEPT OF FREE WILL

My Review of "Just Deserts" by Daniel Dennett and Gregg Caruso

23 March 2021

I was very excited to receive my pre-ordered copy of ***Just Deserts*** in early 2021. Dan Dennett is an obvious influence and inspiration to all philosophers with an evolutionary view, and I was lucky enough to meet Gregg Caruso a few years ago when he came to Newcastle to debate free will with Christian List. As I raced through *JD*, I was offered the opportunity to write a review of it, which was subsequently published at *3 Quarks Daily*. This book and review really helped me clarify my own position on this metaphysical issue, and I consider it a major accomplishment that both authors have said it was a fine review. Please check it out and let me know what you think in the comments.

"Just Deserts: Debating Free Will" By Daniel Dennett and Gregg Caruso

Just Deserts is a surprisingly slim book, only 206 pages long, which could almost be a chapter for one of its authors, let alone a full book from two. It has a whimsical title that hints it might simply be the sweet ending of a multi-course meal cooked up and eaten elsewhere. But don't be fooled! *Just Deserts* holds a titanic discussion concerning two huge cracks in the foundations of human thought. The first is the stated crack about the well-known problems of free will, moral responsibility, and social justice. The second crack is an unstated one that only reveals itself in a meta consideration of the styles of the two authors. That shows us there's a very deep question underneath it all concerning how we should even do philosophy to properly think about these topics.

I'll return to that second crack once we've explored the first one. But why do that at all? Does free will matter to anyone but a couple of bickering philosophers? Of course it does! Sam Harris noted in his recent **Final Thoughts on Free Will** that this topic "touches nearly everything we care about: morality, law, politics, religion, public policy, intimate relationships, and feelings of guilt and personal accomplishment. … In fact, the Supreme Court of the United States has worried about this and called free will a 'universal and persistent foundation for a system of law' and has said that determinism is 'inconsistent with the underlying precepts of our criminal justice system.' So, this idea of free will seems to be doing a lot of work in the world."

Indeed it does! But do we actually have it?

Guiding us through this question are Dan Dennett and Gregg Caruso. *Just Deserts*grew out of their widely read ***Aeon* article** from 2018, which Dan and Gregg have now revised and greatly expanded into 107 individual exchanges grouped into three main parts and a dozen subsections. Lucky us. These are two of the top philosophers in the world on this subject. They speak without jargon wherever possible. They display an incredible command over the academic literature in the field, yet somehow manage not to assume us readers know any of it. They don't duck or back down from direct questions. They are witty, respectful, and well acquainted with one another's work. And they write informally and at times emotionally with one another. It produces a literally page-turning experience, like an epistolary novel, where I couldn't stop myself at times from flipping ahead to see how one or the other would react to what was being said.

Sadly, their mutual understanding gets strained near the end of the book as the two sides struggle to reconcile their positions with one another. What's the big fight? To sharpen that up, we first need to know the many, many things that Gregg and Dan agree upon. Both are naturalists (*JD* p.171) who see no supernatural interference in the workings of the world. That leaves both men accepting general determinism in the universe (*JD* p.33), which simply means all events and behaviours have prior causes. Therefore, the libertarian version of free will is out. Any hope that humans can generate an uncaused action is deemed a "non-starter" by Gregg (*JD* p.41) and "panicky metaphysics" by Dan (*JD* p.53). Nonetheless, both agree that "determinism does not prevent you from making choices" (*JD* p.36), and some of those choices are hotly debated because of "the importance of morality" (*JD* p.104). Laws are written to define which choices are criminal offenses. But both acknowledge that "criminal behaviour is often the result of social determinants" (*JD* p.110) and "among human beings, many are extremely unlucky in their initial circumstances, to say nothing of the plights that befall them later in life" (*JD* p.111). Therefore "our current system of punishment is obscenely cruel and unjust" (*JD* p.113), and both share "concern for social justice and attention to the well-being of criminals" (*JD* p.131).

That's a lot to agree with! They sound like natural allies, right? This is likely why Dan and Gregg have continued to write together. It would be a monumental advance in political and metaphysical philosophy if they could hash out their differences and build a united coalition against the *status quo*. Gregg and Dan both want to halt the demonization of criminals, and the monstrous forms of retributive justice that exist because of such notions. But both have very different approaches on how to do so.

The central issue at hand is the relationship between *free will* and *moral responsibility*. These are inextricably tied to one another. The more you believe in a person's free will, the more you will hold them morally responsible for their actions. (This relationship has even been demonstrated empirically in **recent experiments**.) And the amount that you hold a person responsible is related to how much they *deserve* to be praised or blamed, rewarded or punished, which, of course, affects the entire justice system.

Dan is a *compatibilist* in this debate. He has long rejected the extreme version of *libertarian free will*, and instead defends a less radical version he calls *the free will worth wanting*. Likewise, he rejects extreme versions of moral responsibility, and instead defends *a familiar concept of desert* that keeps moral responsibility on the table, albeit only for the consequentialist "forward-looking benefits of the whole system" (*JD* p.45).

Gregg, however, is an *incompatibilist* about all of this. In addition to rejecting the extreme version of *libertarian free will*, he uses two further arguments to say we ought to be *free will skeptics* and discard the notion entirely. Then, on the flip side, Gregg cites a definition for *basic-desert moral responsibility* that shows that must go too. He concludes that many terms in this debate are too tainted to rescue, so he has built an entire replacement for them in what he calls his Public Health-Quarantine Model. (There is much more on this in his forthcoming book ***Rejecting Retributivism***.)

Now for the big question—who wins? I have my own personal choice, but the publishers of *Just Deserts* have been gathering opinions from readers in **an online survey** and the current data basically shows a dead heat. So, there isn't *an answer*that undeniably emerges here. But honestly, has any book of philosophy ever done that? Few, if any, readers will finish one and think, "right, that's sorted." But by allowing us to witness their debates, Gregg and Dan have given us a book that's more valuable than if either had just written another one alone. So, the real winner? It's us, obviously.

As for me, I read through the exchanges rooting for Dan and Gregg to come together to solve this issue for us all. But what became obvious over the course of the book was that the second major crack I mentioned at the top of this essay—the crack that lies at the heart of doing philosophy—seemed to doom Gregg and Dan to remain divided. Gregg is an extremely proficient analytical philosopher. He strives for absolute clarity in his arguments, and quite possibly achieves the Holy Grail of *internal consistency* for his project. But Dan looks at too many of his neat and tidy definitions and says, hang on, you can't look at the world that way. Dan repeatedly accuses Gregg of ***rathering*** when he says something must be *this* rather than *that*. To Dan, that's often just a false dichotomy; a binary choice in an analog world. This is an evolutionary view of things that is best exemplified by the following quote from Dan's 2016 essay "**Darwin and the Overdue Demise of Essentialism**."

> "When Darwin came along with the revolutionary discovery that the sets of living things were not eternal, hard-edged, in-or-out classes but historical populations with fuzzy boundaries … the main reactions of philosophers were to either ignore this hard-to-deny fact or treat it as a challenge: Now how should we impose our cookie-cutter set theory on this vague and meandering portion of reality?"

In other words, Gregg is like a master cookie cutter. He continually presents his tidy definitions, only to have Dan say yes-but-um-well-also-no. I liken this to the reliance of most philosophers on classical logic, which says *A* is *A*, *not-A* is *not-A*, and **the law of the excluded middle** says there is nothing else possible in between. Such rigid definitions work well in the precise worlds of mathematics and Newtonian physics, but not in the fuzzy world of biology. In that realm, the ethologist Nikolaas Tinbergen gave us his **Four Questions** which are now the generally accepted framework of analysis for all biological phenomena. To understand anything there, Tinbergen says you have to understand its function, mechanism, personal history (ontogeny), and evolutionary history (phylogeny). As a very simple example, philosophers could tie themselves in knots trying to define 'a frog' such that this or that characteristic is *A* or *not-A*, but it's just so much clearer and more informative to include the stories of tadpole development and the slow historical diversion from salamanders.

So, is free will more like a geometry proof or a frog? I haven't seen Dan refer to Tinbergen's questions, but this is basically what he was getting at when he wrote that ***Freedom Evolves***. Gregg explicitly questions this approach when he states, "What baffles me about your position…is that nothing remains fixed and agents can go from *having free will* at moment t1

to *not having free will* at t2" (*JD* p.93). A few pages later, Dan confirms this is accurate but says, "this is a feature not a bug" (*JD* p.96). By the end of the book, Gregg expresses frustration with Dan's position saying it "is like wrestling an eel—every time I have a grip on it, or think I do, it slips out of my hands" (*JD* p.198). To me, these are examples of how it may sound innocent and uncontroversial to admit we all evolved — duh! — but the **universal acid** of Darwinism "eats through just about every traditional concept, and leaves in its wake a revolutionized world-view" (*Darwin's Dangerous Idea* p.63). This is still eating its way through much of philosophy, and to me, this acid eats through some of Gregg's as well.

I mentioned earlier that Gregg uses two additional arguments to reject the concept of free will entirely. These are the *manipulation argument* and the *luck argument*. Debates about these take up over fifty pages of *Just Deserts*, but here is a very short summary. We are entirely the products of our genes and our environment, but we didn't choose our genes (*constitutive luck*) and we are never in control of our environments (*present luck*). So, according to Gregg, "luck swallows all" (*JD* p.15). Then, one can use a variety of thought experiments about neurosurgeons *manipulating* someone's actions to show us that we wouldn't blame a person or hold them morally responsible for these manipulated actions, so why should we do so when luck determined all the forces manipulating our actions now?

Dan does a fine job resisting these arguments, but adding Tinbergen's perspective gives us a few additional tricks. It isn't luck that I grew up to be a person rather than a horse. Once I was conceived, the evolutionary history (phylogeny) that led up to me put a lot of constraints on my personal development (ontogeny). Luck may explain all the *differences* between me and every other person out there, but we needn't worry about luck when describing all the things we have *in common*. There are hordes of characteristics that all humans share, but the one that is most important for this debate is our capacity to *learn*. The extreme neuroplasticity we have (a mechanism of free will) is what enables all but the most unfortunate humans to sense and respond to their environments (a function for free will) to the point where they slowly, slowly become a unique self.

Furthermore, all of the *differences* in our environments may be down to luck, but once again there are many elements here that are *the same*. This is what gives the blind justice system its ability to function across all society—our shared capacity for learning from the elements of culture that we also all share. Wherever we find examples of impairment in learning, or poor exposure to 'good' or 'bad' culture, we lower the judgments of moral responsibility for that person. This is precisely why children are judged differently than adults. It would result in a ***sorites paradox*** to try to precisely define when a child becomes an adult but we can agree to legally draw a bright line for convenience and still roughly understand the fuzziness in Tinbergen analyses that show how an evolved self eventually becomes the location of major influences on our judgeable actions. This is a key part of Dan's argument about the importance of *control* to the debate on free will. He isn't talking about control in a skyhook kind of way that comes from nowhere. He's just talking about the engineering sense of control as in *where the decision-making is located.* And for many decisions, the self *is* this location. It is also where any punishments can work to teach new habits to that self. This, then, is why the manipulation argument fails to persuade. The evil neurosurgeons cannot be taught a lesson by punishing their puppets, so their imaginary examples cannot be extended to the general case of being a human.

To me, these arguments shake the foundations of Gregg's project, but I admit he's shaken my beliefs too. What remains clear from that shakeup is the need to reform all retributive attitudes and any awful prison conditions that criminals currently face. That kind of

backward-looking *basic desert* has no justification, and there is no place for it in an enlightened society. It may feel too radical, too certain, to use a lot of classical logic to implement these reforms by throwing out so much terminology and culture that has co-evolved over centuries of civilization, but what we're left, then, is Dan's view of *free will worth wanting* and the *familiar concept of desert* that most people understand when I say this *deserves* deep consideration. We may not *have* free will, but we *are* a will with an infinite degree of freedom (subject to certain restrictions).

Is that enough? Can we be *a puppet who loves his strings* (*FD* p.86)? Do we have to patiently endure the slow evolution of the definitions involved in this debate? Or can an abrupt replacement of concepts give us the clarity we are seeking and the compassionate reforms we want. Can we find a path between moral outrage that is too hot and logical pragmatism that we fear is too cold? The more people who read *Just Deserts,* and contemplate the cracks between Dan Dennett and Gregg Caruso, the better chance we'll have of finding out.

A Few Further Thoughts on Just Deserts

29 March 2021

I got quite a lot of nice comments last week on **my review of *Just Deserts***. The authors of the book—Gregg Caruso and Dan Dennett—both told *3 Quarks Daily* that it was a good review so I consider that a real feat to have satisfied both sides in such an argumentative book.

One of the comments I saw was a wish to hear a debate between Dan Dennett and Sam Harris, who recently posted his "**Final Thoughts On Free Will**." I can't make that happen, but while I'm focused on this topic, I thought I should write a little something about Sam's position.

I'll get to that soon, but first, I just thought I'd share a quick post with a few of the paragraphs that had to get cut from my 2,500-word review. I may want to refer to these later, and they really were darlings I hated to kill. Enjoy! I'll be back soon with more on this topic.

The fear of determinism is an ancient one, stretching back to early religious questions about whether gods or the fates foresaw and controlled everything we humans do. When the Enlightenment came along, and Newton showed us the mechanical workings of the cosmos, and Darwin showed us the blind nature of natural selection, our fear of control shifted from warm and (hopefully) friendly gods to the cold and calculating inevitability of logic and mathematics. Dostoyevsky wrote a wonderful passage about this in *Notes from Underground* in 1864:

> *"You say, science itself will teach man that he never had any caprice or will of his own, and that he himself is something of the nature of a piano key or the stop of an organ, and that there are, besides, things called the laws of nature; so that everything he does is not done by his willing it, but is done of itself, by the laws of nature. … [But I] would not be in the least surprised if all of a sudden, a propos of nothing, [a man were to arise and] say to us all: 'I say, gentlemen, hadn't we better kick over the whole show and scatter rationalism to the winds, simply to send these logarithms to the devil, and to enable us to live once more at our own sweet foolish will!'"*

If you'd prefer to just form your own opinion and cast your own vote, go read *Just Deserts* now. If you want to hear what I think, here goes. But it helps to put my cards on the table first, so you know where I'm coming from. I call myself an evolutionary philosopher. I think paying close attention to the history of evolution gives us new insights into age-old philosophical questions. So, I'm obviously a huge fan of Dan Dennett. But I've also seen Gregg debate free will at a local event, and I got to have a few beers with him in the pub afterwards while he continued the debate informally. I found him extremely impressive and persuasive. (He's also just a very nice guy.) When I found out about *Just Deserts*, I couldn't wait to get my hands on it and see how the two of these guys got on with things.

In some sense, both of these are quite radical positions, and both of them are conservative as well. Gregg is simply doing standard analytic philosophy—dissecting definitions and logically analysing their properties and relationships—while driven by a commonly held desire to reform a prison system we almost all agree is not working. But his conclusions demand that we drop all longstanding usages of free will, desert, responsibility, blame, and punishment. He thinks they are all too tainted and has built an entire replacement for them that he calls a Public Health-Quarantine Model. (There is much more on this in his forthcoming book ***Rejecting Retributivism***.) Dan is never one to shy away from an unpopular opinion (c.f. "consciousness is an illusion"), but he is deeply skeptical of such radicalism here. He maintains that respect for the law "is a foundational requirement of stability in a state" (*JD* p.164). Instead, he would rather propose deep changes to philosophy "because we cannot do the job right while sequestered in our ivory towers" (*JD* p.165). He seems to think folk terminology is worth holding on to here, even if their meanings must unavoidably shift.

Ultimately, this may just be a choice in strategy between Dan's position and Gregg's. If so, that would mimic a story Dan told in his 2008 essay "**Some Observations On the Psychology of Thinking About Free Will**." Regarding Daniel Wegner's book title ***The Illusion of Conscious Will***, Dan wrote, "Our disagreement was really a matter of expository tactics, not theory. … Should one insist that free, conscious will is real without being magic, without being what people traditionally thought it was (my line)? Or should one concede that traditional free will is an illusion—but not to worry: Life still has meaning, and people can and should be responsible (Wegner's line)? The answer to this question is still not obvious." Perhaps Dan is still wrestling with this choice, although it's clear Gregg thinks his choice is the right one judging by the weight of recent books and articles he has put behind it.

Another Free Will Debate — Kaufman v. Harris (Part 1/2)

5 April 2021

On March 22nd, 3 Quarks Daily published my review of Gregg Caruso and Dan Dennett's new book ***Just Deserts: Debating Free Will***. Ten days earlier, Sam Harris released his **Final Thoughts on Free Will** on his *Making Sense* podcast. Was he trying to scoop me? I wish! Did he even mention *Just Deserts* in his podcast? Surprisingly no! Why not? Probably because he and Dan Dennett have already had several heated conversations about free will. There was **Dan's dismissive review of one of Sam's books**, **Sam's pissed-off response to that review**, and then a **2-hour podcast discussion** trying to smooth the water between them. No need to go back to all that!

So, what prompted Sam to speak out about free will now? Well, I think the real reason Sam posted his thoughts when he did was because it was fresh on the heels of a 3-hour discussion he had with Scott Barry Kaufman on ***The Psychology Podcast***. Scott published his amazing book ***Transcend*** last year, which has the sub-title "The New Science of Self-Actualization". In other words, having a self that is free to be actualized is kind of an essential part of Scott's project. But Sam is famous for denying these things in his work, including his 2012 book ***Free Will***.

Since I'm deeply immersed in the topic of free will right now, I thought I'd spend a few posts on these recent discussions. I'll get to Sam's "final thoughts" in a few posts, but first, let's take a closer look at **Part 1 of Sam's conversation with Scott**, which was posted on February 25th. Next time, I'll delve into **Part 2**, which was posted on March 4th. I won't bother transcribing all three hours of these free podcasts, so please listen to them for yourself for the full story. But here are some important bits that I'd like to comment on.

- **Sam:** When I was in college, a girlfriend broke up with me and I just became this machine that was producing unhappiness until an MDMA experience showed me that that could be interrupted with no reason attached.

Sam has become a strong proponent of psychedelic drug use after this early experience kicked off his life as a contemplative and public intellectual focusing on consciousness and free will. I haven't used such drugs myself personally, but as a 49-year-old-man now, I have to say that every time Sam talks about the important lessons he got from the experience, I think I've already learned those lessons from other experiences. (Notably, in grand nature spots, but also while studying astronomy, geology, and deep evolutionary history.) Could I have learned these lessons earlier in my life on a drug trip? Maybe. But I tend to agree with Abraham Maslow who thought such experiences were cheating to try to get to self-actualization. Better to have

reasons for your emotions and learn from those.

- [Sam] studied for years with the leading Buddhist meditation thinkers. There are dualistic vs. non-dualistic forms of awareness meditation, with different sets of instructions for each one as to what to pay attention to and why. With these exercises, you aren't meditating yourself into perfection; you are just learning to recognize something that is already there.

I've not been on any lengthy retreats yet, but I have done a fair bit of reading about meditation, and I have practiced it on my own for nearly 20 years now with the help of many guided meditation sessions along the way (including lots from Sam's **Waking Up** app). Meditating has been a good and useful experience in my life, but, a bit like using psychedelics, I think it's an artificial experience that doesn't have quite the relevance to everyday life that Sam thinks it does. I'll say more about that in the next post, but I wanted to flag that I have meditated and enjoyed it.

- **Scott:** I want to read a sentence you wrote because I have issues and questions with it " Consider what it would actually take to have free will. You would need to be aware of all the factors that determine your thoughts and actions and you would need to be in complete control of these factors." This sentence reads like you are an implicit dualist. Who is the "you" in that sentence?

This is a great observation that I also notice whenever I read or listen to Sam. He continually toggles back and forth between his cold declarations about the lack of a self or free will and then his hot instructions about what "you" need to do or notice about "your consciousness and its contents." In **Dan Dennett's review of *Free Will***, Dan pointed this out too. He noted a sentence on p9 that said "I, as the conscious witness of my experience, no more initiate events in my prefrontal cortex than I cause my heart to beat." Dan said, "If this isn't pure Cartesianism, I don't know what it is. His prefrontal cortex is part of the I in question." That's exactly right. Consciousness is **embodied** and should not be spoken of so separately as Sam is wont to do.

What does Sam say about this sentence when Scott asks him how to understand it? He shifts into **Zen Koan** mode.

- **Sam:** Yeah. It's not really understandable in that way. What you've really just landed on is the problem with the concept of free will. It's an incoherent idea. ... As you know, Dan Dennett has tried to purify the concept so as to have in his terms a "free will worth wanting". ... [But] you're not acknowledging just how many important things shift ethically once you let go of that spooky free will. Things really do change. And they change in ways that are important not just for our justice system and our concept of justice, they are important for ethical intuitions about what it means to be a good person and how we should feel in the presence of all the misadventures we have in life...and Dan Dennett's project acknowledges none of that. That's why he and I have never agreed on this topic.

This is an incredibly disingenuous reading of Dan's work and his previous exchanges with Sam. If anything, it's the other way around as Sam has not done the hard work of trying to really see what goes away when the concept of free will disappears. As one example, Dan noted in is review of *Free Will* that "entirely missing from Harris's account...is any acknowledgement of the morally important difference between...the raving psychopath and us." Perhaps this is why Dan has just moved on to debate Gregg Caruso instead, since he's

actually a serious thinker who has tried to develop **a Public Health Quarantine Model** to replace our current retributive justice system. Poking at the holes in Caruso's model took up a significant portion of *Just Desesrts*. Sam doesn't even have a model to poke at. And it's not just psychopaths he doesn't see as any different than the rest of us.

- **Sam:** The rules, ethically and psychologically, seem to change entirely for people, when you are talking about [other] people. They don't think this way about chimpanzees. They don't think this way about people with certain kinds of brain damage. ...The problem is that it doesn't make any sense. ... It's very difficult to make sense of this in terms of the streams of causality that I'm not aware of, in terms of gene transcription, and neurotransmitter behaviour, and all of the causes reaching back to the Big Bang that I didn't author.

I'll point this out again in the next post, but the way Sam speaks about humans is literally dehumanising. In case it's not obvious how dangerous that is, David Livingstone Smith has done excellent work on the subject. (See **this book review** by Smith for some examples of Nazi dehumanisation.) I get that Sam is merely recognising here that the "folk" have different intuitions about people compared to their intuitions about chimpanzees and brain damaged people, but by saying this doesn't make sense, he is opening up the door to some very bad attitudes.

- **Scott**: You are really hung up on the magical part of free will.
- **Sam:** It's not hung up! It is what people mean when they feel that someone should be punished, really punished, because they deserve their punishment. That is "just deserts." That is someone who feels that the logic of retribution is anchored to libertarian free will.

It's sad to hear that Sam is still stuck repeating these points even though Dan Dennett took them apart several years ago. In **his review of *Free Will***, Dan noted that Sam said, "However, the 'free will' that compatibilists defend is not the free will that most people feel they have" (p16). But Dan countered, "First of all, he doesn't know this. [And experimental philosophy suggests he's wrong.] But even if it is true, maybe all this shows is that most people are suffering from a sort of illusion that could be replaced by wisdom. After all, most people used to believe the sun went around the earth. They were wrong, and it took some heavy lifting to convince them of this." And in *Just Deserts*, Dan and Gregg do lots of this kind of lifting. Both agree there are reasons to get rid of retribution and libertarian free will, and you can do so as a free will skeptic (Gregg's project) or as a compatibilist (Dan's project).

- **Scott:** It seems like people can do all the things they care about. If they think they care about making choices that are somehow uncaused, they just aren't literally understanding what that means, as you point out. What people really mean when they insist that free will is important is they don't want to feel coerced. They think of causes as sources of coercion, but that's a confusion. I think people want to make choices that are consistent with their own goals and be able to deliberate about the causes where their desires aren't totally clear, and they can do those things. And it's pretty clear their consciousness participates causally in that process.
- **Sam:** I would dispute that. ... For much of what we seem to do consciously, it remains mysterious why consciousness need be associated with any of these things. We can imagine building robots that could pass the Turing test that could do all of these things without there being something that it is like to be those robots.

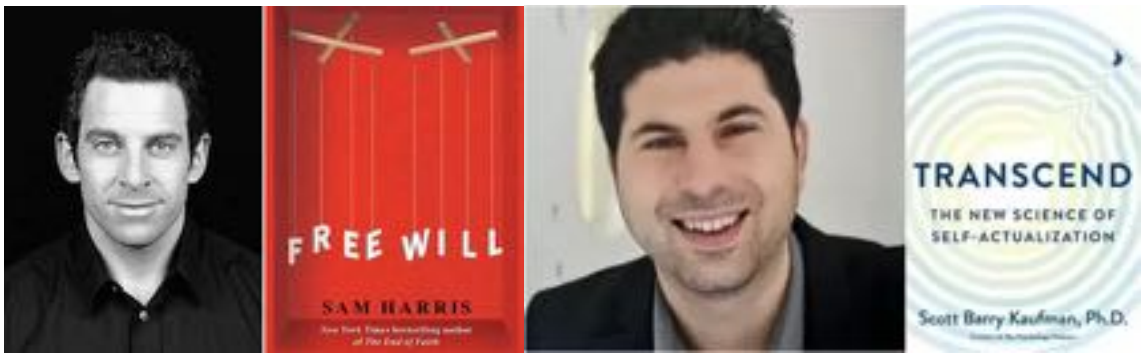
Ah, now we're getting to another root problem with Sam's view of the world. I think that all his meditation training focused on "consciousness and its contents" has left him with his

dualist language and a fundamental misunderstanding of what consciousness is. I've spent the past year looking at the definitions and studies of consciousness (**summarised here**) and found it requires a much more nuanced understanding of how consciousness emerges to varying degrees via a hierarchy of activities. Along the way, there is clear evidence that conscious awareness is required for certain types of learning. (This kind of awareness is possibly what Sam means by "consciousness" but that's not at all clear considering **his dabbling in panpsychism**.) Also, Dan Dennett's paper on **The Unimagined Preposterousness of Zombies** shows just how unimaginable Sam's robots really are.

- **Scott:** Of course we are just our biology. What else would we be? But isn't it the point that our biology encompasses all the interesting stuff that we are? You could still say that it means something for the robot to be a unique robot. But don't you think that the interesting thing is that the biology encompasses all the unique aspects of what Sam Harris is and who Sam Harris is, including your unconsciousness *and* your consciousness?
- **Sam:** Hmmm. I'm trying to think of how to make this point land…
- <<< ROLL PODCAST CREDITS >>>

Wow, what a cliffhanger ending! I shouted, "Yes!" to Scott's question, but let's see what Sam says next time.

Another Free Will Debate — Kaufman v. Harris (Part 2/2)



7 April 2021

In the **last post**, I covered **Part 1** of Sam Harris' recent conversation with Scott Barry Kaufman. Sam didn't come off very well in that one, but Scott left him with an excellent cliffhanger question so maybe he can redeem himself. Let's remind ourselves what that question was and then jump right into **Part 2** and pick out some interesting bits from the rest of their conversation.

- **Scott**: Of course we are just our biology. What else would we be? But isn't it the point that our biology encompasses all the interesting stuff that we are? You could still say that it means something for the robot to be a unique robot. But don't you think that the interesting thing is that the biology encompasses all the unique aspects of what Sam Harris is and who Sam Harris is including your unconsciousness and your consciousness?
- **Sam:** Hmmm. I'm trying to think of how to make this point land…
- <<< ROLL PODCAST CREDITS >>>
- <<< NEXT EPISODE PICKS UP RIGHT FROM THERE >>>
- **Sam:** All of the causes of what I'm conscious of were first unconscious. I'm not aware of what my brain is doing at the synaptic level. I'm not a dogmatic materialist [but] let's just talk in terms of materialism. ... So, my mind is what my brain is doing. … What we're talking about is information processing in a physical system. In my case, the computer is made of meat. In a robot's case, it's silicone. In neither case is there something extra which is emerging or being added which gives a degree of freedom beyond just the impressive complexity of the system in dialogue with its environment.
- **Scott:** I think there is. Let me try to pinpoint precisely what I think that extra thing is. Cognitive control includes things like implementation of intentions. ... You are right, in the moment we don't really have free will but we have the capacity to shift our behaviour in the future so that we can learn from our mistakes so that we can even make moral reasoning decisions. Turtles, chimps, apes, and robots right now don't have a great capacity for moral reasoning about an action they already made so that they can change their behaviour in the future. To me, that conscious control is free will. But I don't think I can convince you to use that label for that phenomenon.

I think Sam is correct here, but the "impressive complexity of the system in dialogue with its environment" is actually just a very good description of what Dan Dennett calls "the free will worth wanting." And Scott's "cognitive control" and "implementation of intentions" is just more of the same. In my **Summary of My Evolutionary Theory of Consciousness**, I gave the following short definition:

Consciousness, according to this evolutionary theory, is an infinitesimally growing ability to sense and respond to any or all biological forces in order to meet the needs of survival. These forces and needs can vary from the immediate present to infinite timelines and affect anything from the smallest individual to the broadest concerns (both real and imagined) for all of life.

So, when Scott notes that the capacity of humans to change their behaviour is much greater than the capacity for turtles, chimps, and robots, I would say he's describing points on the spectrum of what all of these (living or non-living) systems are able to process with their levels of consciousness. The more you can sense, and the more responses that are available to you, the more degrees of freedom you have. And compatibilists may wish to call this your free will. Much like consciousness, this isn't an on/off switch. It's an infinitesimally growing (or shrinking) amount of freedom. As Dan Dennett said towards the end of **his review of Sam's book**, "You can't be '*ultimately responsible*' (as Galen Strawson has argued) but so what? You can be partially, largely responsible." Equivalently, you can't have Ultimate Libertarian Free Will, but so what? There is a growing sense of freedom along the way towards that, which we might agree to call free will.

But Sam doesn't think we actually have that sense! And that is a big part of his argument that needs to be addressed.

- **Sam:** [People] think they are having an experience of being a self that can author its own actions. The experience of having free will and the experience of being a self...are two sides of the same coin. ... Meditation, successful meditation, absolutely proves to you from the first-person side that that is a false point of view. [The] point of view that gives motivation to this claim about free will [is] how you feel when you feel that you are the conscious upstream cause of the next thing you think and do. [But that is] because you are not noticing that the next thing you think or intend to do is simply coming out of the darkness behind you which you can't inspect. It is genuinely mysterious.

This is the kind of argument you make when you see consciousness as an on/off switch and you put far too much stock in meditating on conscious awareness (which is actually level 5 in **my hierarchy of consciousness**). Sam is right that "you" are not "the" conscious upstream cause of the next thing you think and do. But I would say that "you" are also an unconscious upstream cause! And these bleed back and forth into one another. There is bi-directional feedback between our unconscious activities and our conscious activities. If this was *genuinely mysterious*, the thoughts that came out of the darkness would be shocking and unrecognisable to us. But, of course, that's not what we experience. That only appears to happen in genuine cases of psychosis, which we diagnose and try to treat if that occurs. Why exactly does Sam think this way? He draws on two examples over the rest of the talk, so let's present them both at once and consider them together.

First Example:

- **Sam:** Take a moment of conscious deliberation. I have a glass of water and I can decide to pick it up and have a drink now or I can decide to wait. This is a prototypical case of me being in the driver's seat. I'm free to do this. No one's got a gun to my head. I don't have some kind of compulsive water-drinking behaviour. I'm a little bit thirsty, I'm conscious of thirst, but I can choose to resist my thirst. That seems to be me prosecuting my freedom there. But the more you pay attention to what it's like to make that choice out of your own free will, the more you will discover that it is *absolutely mysterious*, in every

particular, why and how you do what you do and when and how you do it. Subjectively, I have no idea why or how I do any of these things. I have no idea why or how one particular moment becomes decisive.

Second Example:

- **Sam:** [I can provide a long description of someone becoming a classically trained musician because of a love of Bach.] That's true of somebody. But not me. Why not? Why don't I care about Bach? All of these things have reasons, they have explanations, causally...
- **Scott:** Those are the things that make you who you are, even if you don't know why they were caused. [They are part of your] environmental and biological confluence.
- **Sam:** Yes. It's deterministic or random, but it's some pattern of causation. But so what does it mean to say that I am free to take a deep and all-encompassing interest in classical music? … The problem is, I have almost no interest in playing the cello. The fact that I don't is something that I did not author. ... I am as I am with respect to classical music. Now, just imagine that by force of this conversation, you said something that inspired me to be different than I'm tending to be, this would really be the ultimate instance of free will because this would be kind of a surmounting of all my prior tendencies into this new commitment. What would it be like for me to experience that awakening in my own consciousness? That would be totally compatible with the evil genius in the next room saying "We're going to give him the cello desire here." It would not demonstrate anything like free will. It would be like, "What came over me?" This would have come from outside of consciousness. It's not me.

These are not persuasive. In the first case, facts from our evolutionary history show that we humans are animals who only *generally* need water. We don't need to constantly drink, and there is a large range of hydration within which we can function perfectly well. Therefore, there is rarely, if ever, one instantaneous all-encompassing need to drink NOW. When Sam says he has "no idea why or how one particular moment becomes decisive," he is looking for something that just isn't there. Why not? Because it doesn't need to be there! Like **Buridan's Ass**, random noise is all that's necessary to decide to drink at any one second vs. another. However, let's say I'm a spy and I pre-arrange to have a drink in a bar at precisely 15 seconds after 8:00PM because that will be a signal to my counterpart that "everything has been arranged." Guess what. As long as everything goes as planned, I'm going to have that drink at precisely that time. And that particular action is going to feel very authored. Sam is trying to stack the deck with his meaningless example, but a meaningful counterexample drives an entirely different intuition.

Similarly, the second example isn't as mysterious as Sam claims either. A love for classical music and a drive to play the cello are very clearly driven by a bit of genetic variance (constitutive luck) and a bit of environmental conditions (situational luck). If you were born tone deaf and 500 years before the invention of the cello, you aren't going to have a drive to play the cello. If you are born with perfect pitch into a family of professional musicians who lead happy lives and have instruments all over the house, you may very easily develop a drive to play the cello. If your situation is somewhere in between these extremes but, at some point along the path of your life, cello-playing looks like a promising path to meet one or several of your **Maslow's hierarchy of needs**, then it is very possible a drive will develop to lead you down that path. That's how one might convince Sam to play the cello—by showing him he can, and that doing so would solve a very important need he has, over and above all the other need-fulfilling activities he already undertakes. That's a pretty high bar at this point in Sam's

life because of his particular **path dependence**. But if we managed it, these causal explanations would be nothing at all like an "evil genius in the next room saying 'We're going to give him the cello desire here.'"

Are we "Ultimately Free" to choose all of these factors in our lives? No. No one should ever think that we are. But is there freedom in discovering who we are and exploring the "impressive complexity of the system in dialogue with its environment"? Yes. And I think that's a satisfying way to look at life. What is stopping Sam from taking this big picture perspective? Towards the end of the podcast, Sam shows that it comes from his personal history placing a laser focus on meditation and the tiniest details of neuroscience.

- **Sam:** In certain cases, [conscious experience is] not descriptively mysterious at all. We know causally that we can tell a story about it. It's just two different levels of connecting to the phenomenology here. When I say mysterious, I mean like, I can move my hand, right. This is one of the most prosaic things about me that I can move my hand. I can do this. I have no insight into how I do this. If I suddenly couldn't do this, that would be flabbergasting. But the fact that I can do it is also flabbergasting. I have literally no insight. I know something about the neurology of this. I can talk about muscle fibres, actin, and the transduction in motor nerves and…I can vomit my concepts.
- **Scott:** You kind of get it.
- **Sam:** So, I'm not saying you can't have any insight into this, but there is still something, however deep you go, however atomised your experience consciously becomes of a phenomenon, there is just simply this fact of *first something wasn't there and then it's there*. You can shatter your subjective experience down to its atoms and notice that things are just appearing out of the darkness. Sights, sounds, thoughts, emotions, intentions, or their microconstituents. Things can get incredibly pixelated when you spend months on retreats doing nothing but paying attention to mostly sensory perception. It can break down, especially if you are doing it strategically, so as to look for its kind of smallest and briefest aspects, which is one style of meditation. Things become amazingly pixelated. You don't feel that you have a body anymore. You feel that you have a cloud of sensation, of temperature, and pressure, and movement, which just doesn't have the shape of a body at all. You don't feel "hand", you feel these micro-changes of primary sensation at each moment. But again, whatever you are noticing is there and then it's not there. And then something else is there and then it's not there. And "you" are not doing any of it. That's the crucial point. "You", the one who is witnessing, aren't doing any of it.

This perfectly captures Sam's walled-off dualism. If "you" are only "the one who is witnessing" then of course you aren't going to be able to understand everything moving in and out of that perspective. To expect differently would be like what the ecological philosopher Arne Naess calls "trying to blow a bag up from the inside." There are limits to what conscious awareness has access to and you have to examine the facts outside of those limits in order to understand it. And that's okay! The view from the interior—no matter how pixelated—only gives you so much. But a holistic view adds nicely to the picture, and it lets you understand more of the interior even if you don't have access to what is outside of it. For biological phenomena like us, **Tinbergen's four considerations** of (1) evolutionary histories and (2) personal histories, along with (3) functions and (4) mechanisms, add up to this big and informative picture.

In some ways, it's just philosophical wordplay to decide to call these perspectives *free will* or not, or *free will worth wanting*, but whatever label you use, the ideas you attach to that label have real consequences for the way you navigate through life. Let's examine a few of those.

I mentioned in **my last post** that Sam's views lead him to a very dehumanised place. In this episode, he puts that on display even further. Here are five examples of that which add up to something quite disturbing:

First Example:

- **Sam:** Almost no one understands this. Dan Dennett does not understand this. He obviously doesn't. He obviously feels like a self. And that is the string upon which all this controversy is strung. Most of the people listening right now are thinking, "what the fuck is he talking about?" But that voice in your head that says, "what the fuck is he talking about?"…that isn't you! That is not a self.
- **Scott:** What do you mean that's not you? It's you! Again, you're a dualist when you say that.
- **Sam:** It's no more you than the bead of sweat that drips down your forehead is you. It is an object.
- **Scott:** I disagree! People don't identify themselves with their hand, but they identify themselves with their conscious desires and motivations so we can have gradations of things, of parts of our body that people identify themselves with.
- **Sam:** From the point of view of consciousness, there is simply consciousness and its contents.

Second Example:

- **Sam:** [Trump's election] is a little bit analogous to if we elected a rhinoceros to be president. I'd be fucking tearing my hair out over how awful that is. At no point am I imagining that the rhinoceros can be anything other than a rhinoceros and at no point am I wishing suffering upon the rhinoceros. I don't hate the rhinoceros. The rhinoceros just shouldn't be president of the United States. That's a catastrophe to do that. And in some sense, we elected a rhinoceros president.

Third Example:

- **Sam:** Someone comes into your house and wants to kill you and your kids. By all means, shoot that person in the head. That is what guns are for. You should do it if it's a grizzly bear and you should do it if it's a person who seems to think he has free will to kill you and your kids. That's morally uncomplicated in my view.

Fourth Example:

- **Sam:** Hatred really does require an attribution to someone that they could and should have done otherwise. It's like you really do believe they are the authors of their bad actions. The moment you find that they have a brain tumour that makes them exculpatory then you change your response. You think, well I did hate Charles Whitman for getting up in that clocktower and killing all those kids but once they performed an autopsy on him and found that massive brain tumour pressing on his amygdala, well then, okay, I have to recognise that I can't hate the guy. He was as unlucky as the kids he shot. On some level that happens to everybody, once you recognise that free will is an illusion.

Fifth Example:

- **Sam:** Every instance of [voluntary control], the sufficiency of my strength of will in one case, the weakness of my will in another case, every bit of it is being determined by states in my brain which I didn't author, which I didn't create.
- **Scott:** It's still you! It's still you!

- **Sam:** But my liver is still me and it gives me absolutely no sense of free will. If my liver stops, if my liver is working exactly the way it is in this moment and no other way, if it works better tomorrow, or stops completely on Friday, I am a mere victim of those changes, or witness to their consequences. It's not within the domain of my autonomy or agency. But so it is with states of my brain. So it is with each instance of neurochemistry in my brain. And yet that produces everything that I experience including my preferences, my goals, my impulses that are in conformity with my goals, and then my sudden subversion of those impulses with some alternate impulse. That's getting piped up from below and ... the fact that that comes online in that moment and doesn't in another, that's mysterious. The fact that it comes on to the degree that it does, and not one degree further, is also mysterious. It's probably dependent on other things that seem completely adventitious to my character like whether I got enough sleep the night before or whether I had a full lunch or whether I got enough sunlight.

Bollocks! There are hard evolutionary facts that differentiate human minds from beads of sweat, rhinoceroses, bears, and brain tumors. Dan Dennett already answered this with an extended reply in **his podcast conversation with Sam**. It's helpful to read that in its entirety:

That's very useful. Tom Wolfe has this passage where he says what we've learned from neuroscience is that we're wired wrong. Don't blame me. Don't blame us. We're wired wrong. No! What neuroscience shows us is that we're wired. It doesn't show us we're wired wrong. Some people like poor Whitman are wired wrong. ... You're saying it's brain tumours all the way down. Well, I find that extrapolation doesn't move me at all. I don't think it's a logical argument. I think it is a mistaken extrapolation. It's like a mathematical induction gone wrong. [Free will libertarians also] say, we're all that way. Well, no. That's precisely what we understand — that we are not all disabled. Nobody's an angel. Nobody's perfect. So, if anything short of perfection counts as being disabled to the point of being exculpatorily disabled, then you're right. But that's a very strange view. The idea that you couldn't be able enough to be held responsible is the crux of the issue right now between us. I say that the boundaries are always porous, and as we learn more about neuroscience, we may very well move some people that are exculpated into the guilty / not excusable category and others will move in the other direction. But we'll still keep the distinction between those who are basically wired right and those that are wired wrong.

This is similar to a point I made in **my review of *Just Deserts*** about how a Tinbergen view of free will challenges the view that luck "swallows everything" in our considerations.

It isn't luck that I grew up to be a person rather than a horse. Once I was conceived, the evolutionary history (phylogeny) that led up to me put a lot of constraints on my personal development (ontogeny). Luck may explain all the differences between me and every other person out there, but we needn't worry about luck when describing all the things we have in common. There are hordes of characteristics that all humans share, but the one that is most important for this debate is our capacity to learn. The extreme neuroplasticity we have (a mechanism of free will) is what enables all but the most unfortunate humans to sense and respond to their environments (a function for free will) to the point where they slowly, slowly become a unique self.

Sam has taken the giant step-change introduced by Charles Whitman's brain tumor and tried to apply its conclusion to each and every step-change up or down the evolutionary ladder from there. To him, nothing is responsible all the way down to beads of sweat, and nothing is responsible all the way up to billions of average humans. This is a very blunt and useless view of the world. And in the wrong hands, it could be used to wipe away humans as easily as one would wipe away a bead of sweat. I'm not at all suggesting Sam is that kind of a monster, but it would take a weird view of morality to intervene here and save us from such

dehumanisation. As you might expect, Sam has exactly that kind of weird view.

- [**Scott and Sam** start to have a giant discussion about **the is-ought divide**. Sam thinks it's a language trick we should just ignore. He thinks the only thing you need to boot up morality is to agree that "we don't want the worst possible misery for everyone."]
- **Scott:** The way I think about it is that there is no "should" without "in order to," which is a goal. If someone says you should do X, that necessarily implies that you should do X *in order to* get Y. There can be no should without reference to a goal.
- **Sam:** What if the goal is to avoid the worst possible misery for everyone?
- **Scott:** But "better" or "worse" are value judgments. I don't know why you don't see that.
- **Sam:** Put your hand on a hot stove and then tell me that.
- **Scott:** If it was in order to achieve a broader goal, and putting my hand on that hot stove would help me to achieve that broader goal, I would do that and deal with the suckiness of the feeling.
- **Sam:** [...stammers, then...] To say that "the worst possible misery for everyone is bad is a value judgment," is to say *nothing*!
- **Scott:** You're accepting a particular definition of well-being.
- **Sam:** No, no, no. You're just not understanding my claim.
- **Scott:** How are facts going to lead me to action?
- **Sam:** The facts are that there are very different experiences on offer here and you will helplessly find yourself preferring the good day at Esalen over the rat-filled dungeon, just to take the fairly parochial differences that we can notice here on Earth.
- **Scott:** But good can only be used in relation to a goal. How are you divorcing it from the goal? You disagree with that?
- **Sam:** No, it's just the valence of certain experiences within consciousness that have no necessary reference to a goal. You can be so happy or unhappy that it has no reference point in past or future. You can have the best possible acid trip or the worst possible acid trip and there's no goal there. The sheer extremis of your physiology pushed to the breaking point.
- **Scott:** There are pleasurable and there's unpleasant. But I don't think they map onto good or bad in the way that you claim.
- **Sam:** Dial them up and give them enough time. What if existence was just that? [Sam then presents a poor analogy about how a physicist who doesn't believe in math doesn't get to have a vote at a physics conference, and he claims that is the same as the Taliban not getting to have a vote about morality.] They're imbeciles. They have a shitty culture. We know this. And it shouldn't be taboo to say this.

Ugh! Scott is 100% right here. His "in order to" is another way of restating my argument on **how to bridge the is-ought divide with a want**. For example, I say: Life is. Life wants to remain an is. Therefore, life ought to act to remain so. Scott would put it: Life is. In order for life to remain an is, life ought to act to remain so. These are equivalent arguments that both require additional arguments as to why the "want" or the "goal" are correct.

Sam, on the other hand, has a viciously circular argument that tells us nothing about hard choices between where we are today and his worst possible outcome. How is one to judge whether one is moving towards or away from "helplessly preferring things that have no reference point"? He has no facts to consider for that! And his worst possible outcome is wrong too. I wrote about this in **my response to Sam's Moral Landscape challenge**, but I don't suppose he saw that. He was too blinded by the kind of thinking you get when you mistake a meaningless acid trip for profundity and then add in the woo-ey Buddhist claptrap that **emotions should only flow through you**. Sam wants to divorce morality from

consequences but that's just not possible. The "valence of certain experiences" were given to us by our evolutionary histories and they help us reach our evolutionary goals. Am I free to choose whether or not I experience **Jaak Panksepp's seven basic emotions** of FEAR, RAGE, SEEKING, LUST, CARE, PANIC, AND PLAY? No, I am not. But they drive competing needs that I must meet if I *want* to reach my goals (*in order to* reach my goals), which I have freedom to discover and freedom to choose between.

Are those goals and choices mine? Am I *responsible* for the choices that get made?
Not *ultimately*, whatever that is supposed to mean. I do not stand outside of life's evolutionary history. But those needs are felt by me, and those choices are not made anywhere else, so I don't see why they're not mine. Sam sees things differently, abdicating all goals, choices, and responsibilities (to the universe?), but that ends up tying him in knots and eventually making Dan Dennett's case even stronger.

- **Sam:** There are paradoxes here. The responsibility paradox is real, and I still don't know what I think about it. … When you take a truly competent person who then does something horrible, that person is really responsible. That's the true case of responsibility. But the paradox for me is that the more competent you make the person, the more their failures to behave well become inscrutable. … This is very clear in parenting. I have daughters who I'm certainly not browbeating about the illusoriness of free will. No, I'm trying to raise them to be competent self-regulating human beings. So, when I talk to one of my daughters, if I say, "you really should have done otherwise," …it's never a claim that in this instance, if I rewound the universe, they might have done otherwise. No, this is a causally determined outcome that was always the way it was going to be. But, it's a conversation about what I want them to do next time. And saying that is further input into the clockwork of their lives. So, that will change them. Ultimately, my daughters are going to become civilised human beings who will not behave the way they did at 7-years-old or 12-years-old when they are in their 40s. And those changes will be causally affected on the basis of demands imposed on them. But again, there's no place for the folk psychological notion of free will to land there.
- **Scott:** You wouldn't give your daughter any credit if she became president of the United States some day?
- **Sam:** I do feel like pride is a virtue that has an expiration date in a human life. Developmentally, there's like a critical period where pride is not an ethical error or a sign of psychological confusion. It's actually something you want to get into the code. … But at a certain point, I think you clearly want to outgrow it. … It's not a basis for compassion for oneself and others. … I don't feel pride about anything in my life now. I have all kinds of outcomes I prefer. Sometimes I realise them and sometimes I don't. And the obverse of pride is something like shame. Again, shame is an important thing to be able to feel, but ultimately, I think it reaches its shelf life. You want to be able to transcend shame. Not too early. This is an interesting topic. I'm not sure what I totally believe about it. … You're just telling yourself a story about the past in both cases. You're thinking thoughts in the present that nominally refer to the past and they're making you feel a certain way. It's like you are watching a movie about your past and you're being entranced by it and it's kindling an emotional response that has a certain half-life and it's incredibly boring. It's an incredibly boring thing to do with your attention. It's masturbatory on the pride side, a pseudo-source of gratification, which sets up a system of comparison between yourself and others that ultimately is not a source of well-being. If you are comparing yourself to others and feeling good about that, then five minutes later you are going to be comparing yourself unfavourably to other people who are doing yet more impressive things and you are going to feel bad about that. That pinballing between those two things is not the right

algorithm to live a truly self-actualised life. I do think pride and shame ultimately get outgrown. At what point, that's an interesting question.

Yes, Sam, that is an interesting question, because you are precisely describing the development of a person into a unique and responsible self. This is what Dan Dennett meant when he said that ***Freedom Evolves***. We have the freedom to learn from our experience. If I rewound the universe so that every brain state and environmental influence was exactly the same, it's true that "I couldn't have done otherwise," but that's not the point. You will never face the same exact situation twice. The universe moves on. But you can learn from the first instance and do something different the next time in a similar situation. This ability to review the past is one of the most important capabilities of consciousness that has developed. And it does not have to be boring, masturbatory, or self-flagellating when done correctly. You cannot change the past, but you can have a growth mindset about the future. You should not continually cry over spilt milk, but you are not doomed to be clumsy forever either. And the emotional feelings generated from your own introspection (or in reaction to those expressed by others), are mechanical cranes that help make the necessary changes in your neural wiring to help reach our goals. See my post on **where emotions come from** to understand this in more detail. This larger view renders the "fully competent person who does something horrible" much less inscrutable. They've usually just learned something from the past and decided to pursue a new goal.

Once again, you cannot choose the universe you were born into or the particular characteristics and situations that affect you, but the needs, desires, and goals that you feel do not belong to anyone else, so they are yours to own. The beliefs you hold about this are important drivers of your ability to learn and navigate the world. The emotions that drive us should not be too hot from believing in libertarian free will and ultimate responsibility, but they must not be too cold either, holding no one responsible for anything. Sam's arguments would literally drain the passion out of compassion for ourselves and others, which removes a crucial tool from our ability to learn and grow.

What terms should be used in the most helpful sets of personal beliefs about these issues? Perhaps the use of "free will" comes down to semantic choices between psychologists and philosophers. That's something Scott and Sam explored briefly.

- **Scott:** We can want to want things. You're not distinguishing between first-order goals and second-order goals. What gives us free will as a human species? … It's the wanting to want. It's our capacity to use implementation of intentions to get out of the bed in the morning and go to the gym even if we don't want to. I don't want to do that, but my freedom lies in my capacity to use my consciousness and change my environment in all sorts of ways so that it's easier, so that the constraints aren't as big. Don't you see that as an important part of free will that matters to people?
- **Sam:** I see no reason to call that free will.
- **Scott:** [After a short digression.] There's a really interesting paper about smokers and free will by Roy Baumeister. He found that in almost every case, people overestimated the extent to which they wouldn't be able to quit. They wouldn't be able to have free will [to eliminate] the urge, but it turns out humans have much more self-control than they realise they are capable of.
- **Sam:** There's a difference between voluntary and involuntary action. There's a difference between behavioural self-control and lacking that capacity. Let's say that…my goal is to stop smoking but I'm completely incapable of not smoking. That's one way to be. The other way to be is that I have a goal to stop smoking and I can actually veto the impulse

and stop smoking when it comes online. But every instance of this, the sufficiency of my strength of will in one case, the weakness of my will in another case, every bit of it is being determined by states in my brain which I didn't author, which I didn't create.

There goes Sam again with his dualist "I" sitting outside of his embodied self. But I read the **Baumeister paper** after I listened to this podcast and found it really interesting. I especially liked the following list of definitions from the front of the paper:

- **Agency** is the capacity to initiate and control action. It is related to the term *agent*, as in someone who acts. It encompasses choosing, initiating action on one's own, and accepting responsibility for one's chosen actions.
- **Voluntary control** has multiple meanings. For present purposes, it can be understood as indicating that the person is capable of choosing between performing the action and not performing it. Voluntary control means that the power to decide resides within the individual: the person is capable of making a conscious decision and implementing it. Loss of voluntary control means that the person is incapable of acting differently, either because of external forces or unconscious causes. With regard to addictive smoking, loss of voluntary control means that smokers cannot stop themselves from smoking.
- **Free will** is understood as the capability to act in different ways, subject to the person's own control and serving the person's reasons, goals, wishes, and choices. A recent and authoritative definition, based on an interdisciplinary committee working for a granting foundation, defined free will as the capability of performing free actions. Free actions, in turn, were defined in two ways. One was "any intentional action performed on the basis of informed, rational deliberation by a sane person in the absence of compulsion and coercion." The other invoked multiplicity of possible actions (i.e., the person could do two or more different things) in a given situation as constructed by all prior causes and events. Thus, in simple terms, free will is the capacity to act in different ways in the same situation. It thus overlaps considerably with voluntariness. **Shepherd (2012)** showed that most people do not accept unconscious free will, so free will entails conscious control of action. The term "free will" is a traditional usage but modern theorists generally do not postulate "will" as a distinct psychological entity, so it would be more precise to speak of free action.

I quite like these definitions. They are thoughtful, careful, fully drained of extreme libertarian notions, and compatible with the facts of a naturalistic and deterministic universe. They also overlap with a lot of what Sam thinks is going on in the world, despite his controversial and confused labelling.

- **Sam:** None of this is to deny that certain outcomes in life are better than others and worth wanting. None of this is to deny that there are ways to get what you want out of life and ways to fail to get what you want. None of this is to deny that there is this vast landscape of experience and we need to navigate one part of it so as to be happy and functional and we should avoid navigating so as to be captured by another part which leads to the worst forms of misery. All of that is true, and we can talk about how to do all of that. And all of that includes the prospect that people can learn, and people can improve themselves.
- **Scott:** I don't think what you are saying is wrong. I think you are confusing the hell out of people because you make such great points. The kind of free will that matters to humans —we have all of that. … My point is this. The cybernetic system wants to reach a goal that it desires. … Don't you think that's a sensible form of the term free will, that you have

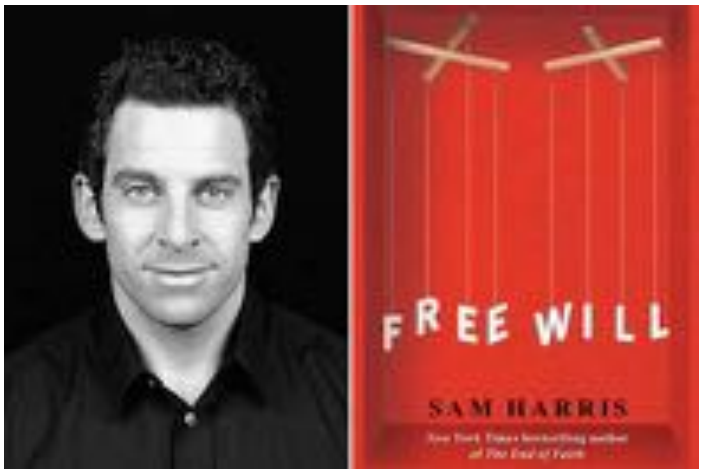
free will to write a book? You want to do so, and you use your consciousness to make that a reality. You don't see that as the kind of free will that people truly care about?

- **Sam:** People care about realising their goals in life. And there are causal ways to succeed at that, and causal ways to fail at that. Learning to play the cello is not going to happen by accident. My denying free will is not the same thing as endorsing fatalism. … This is how people misunderstand this criticism of free will. They think, well, if I have no free will, then why do anything? Why not just wait to see what happens? If I accidentally wait to see if I learn to play the cello, we know what's going to happen there. I'm not going to learn to play the cello. The only way to learn is to intend to learn, to practice, to seek instruction. All of that. People care about outcomes in life that are worth caring about. None of that requires free will to talk about that.

Well, it sure seems like we do need some notion of free will to talk about this stuff. As soon as you deny free will, fatalism, dehumanisation, and coercion creep into the conversation. So, until free will skeptics like Sam come up with a better term, I think we're stuck with Dan Dennett's *free will worth wanting* or the more clinical definitions of free will from psychologists like Scott and Baumeister.

Feel free to propose something different though! I always look forward to the opportunity to learn and improve my own beliefs. Next time, I'll take a brief look at Sam's "final thoughts" on all this and then I ought to be in a good position to offer my own current thoughts.

Some Thoughts on Sam Harris' Final Thoughts on Free Will



19 April 2021

In my last two posts (**1**, **2**), I examined Sam Harris' long appearance on *The Psychology Podcast* with Scott Barry Kaufman. Shortly after those aired, Sam released his **Final Thoughts on Free Will** on his own *Making Sense* podcast, which I thought I should take a look at before summing up my own current thoughts on this matter. I didn't find Sam to be very persuasive in his conversation with Scott, but let's see if he's had any better thoughts upon reflection (and in sole control). Since I've already spent a lot of time on Sam's ideas, I'll try to be quick about it and just pick out any new points that need to be made.

- The concept of free will touches nearly everything we care about: morality, law, politics, religion, public policy, intimate relationships, feelings of guilt and personal accomplishment. But the illusion of free will is itself an illusion. It is built on two things: the ability to choose otherwise and being the source of conscious awareness. But these are both wrong.

There are more things bound up in "free will" than just these two points. In particular, there's the question of avoiding external coercion, as well as the ability to carry out actions and plans that were made using conscious considerations. Both of those do a lot of work in building up feelings about the term free will. But even leaving those points aside for now, Sam's discussion of "the ability to choose otherwise" is a flawed repetition of ideas that have already been debunked. Dan Dennett knocked these down in **his review of Sam's book**, when he said:

"You can't assess *any* ability by 'replaying the tape.' ... This is as true of the abilities of automobiles as of people. Suppose I am *driving* along at 60 MPH and am asked if my car can also go 80 MPH. Yes, I reply, but not in *precisely* the same conditions; I *have* to press harder on the accelerator. In fact, I add, it can also go 40 MPH, but not with conditions *precisely* as they are. Replay the tape till eternity, and it will *never* go 40MPH in just these conditions."

So, looking backwards at decisions that *were* made just doesn't tell you everything you need to know about the ability to make decisions going forward. As for what Sam means about "being the source of conscious awareness", I'll have to hear more to understand his claims.

- There's no place for you to stand outside of the causal structure of the universe.

Agreed. But my unique genetic and environmental history forms its own cause. We witness that from the inside as we act. That is consistent with the embodied view of consciousness. As I said in **my review of *Just Deserts***, we may not *have* free will, as that makes it sound like free will is a possession that could be separated from our selves. But we *are* a will that has degrees of freedom.

- You aren't a self. You're not a subject in the middle of experience. You're not on the riverbank watching the stream of consciousness. As a matter of experience, there is only the stream, and you are identical to it.

That's right that we aren't a homunculus watching the Cartesian theatre unspool before us in some ethereal mind space. But that stream that we are identical to is *something*. It exists. And I don't see why we can't call that an everchanging, unique, and personal self.

- [Sam asks for you to choose a film. Any film.] We can't see how those choices are made. If free will isn't there, then it's not anywhere.

Bollocks! This is just like the point I made in **my last post** about looking for a decisive moment in the random noise of choosing when to drink water. Sam is stacking the deck in his favour by asking for a random film choice, but there are no identifiable interior mechanisms to make random choices. If, instead, I asked you to choose the top 20 films of all time in terms of their return on investment, you would immediately be flooded with ideas on how to act to solve that problem. (And if you are JT Velikovsky, you will have already written **a PhD thesis on this subject**!) Where did those thoughts come from? From some mysterious darkness that we have no access to? No! They would come from learned experience that I myself have experienced, plus maybe some creativity at putting together bits of experiences that I haven't thought about putting together before. This type of problem solving is one of the **13 types of cognition** that we have evolved to have. And that feels very much like a self acting in its own self-interest.

- Everything is springing to mind. What could free will possibly refer to?

To the ability to hold onto a train of thought rather than pinging among these random upsurges?

- Letting go of free will is the only thing that cuts through the desire to retributively punish people.

Not so! The fact that you can't change the past is another perfectly good reason to get rid of the desire to retributively punish people. From a consequentialist point of view, retributivism makes no sense.

- People ask, "if there's no free will, then why are you trying to convince anyone of anything? … Your very effort to convince people that they don't have free will is proof that you think they have it." Again, this is confusion between determinism and fatalism. Reasoning is possible. Not because you are free to think however you want, but because you are not free. Reason makes slaves of us all. To be convinced by an argument is to be subjugated by it. It's to be forced to believe it, regardless of your preferences.

Well, this certainly doesn't track with the history of reasoning with people about their beliefs. Sam hasn't responded to any of Dan Dennett's very good arguments. Why not? Because

people have their own unique personal histories, which drive their passions *and* their reasoning. These people are selves who act for their own self-determination.

- Not thinking about this clearly has consequences. In the United States, there are 13-year-olds serving life sentences in prison. Not because we have determined that they can't be rehabilitated, but because some judge or jury felt that they truly deserved this punishment as retribution because they were the true independent cause of their actions.

This is an abhorrent shame and it definitely needs to be corrected. It's possible that making the argument that "we don't have free will" could actually open many people's eyes to the problems with their retributivist thinking. But it's also possible that such arguments close off many people's minds because they think they definitely are autonomous agents, so free will skeptics must be out of touch with reality.

- At the moment, the only philosophically respectable way to defend free will is to endorse a view known as compatibilism and argue, in essence, that free will is compatible with the truth of determinism. Compatibilists like my friend the philosopher Dan Dennett generally claim that a person is free as long as he is free from any outer or inner compulsion that would prevent him from acting on his actual desires and intentions. So, if a man wants to commit murder, and does so because of this desire, then that's all the free will you need. But from both a moral and scientific perspective, this seems to miss the point. Where is the freedom in doing what one wants, when one's very desires are the product of prior events that one had absolutely no hand in creating? From my point of view, compatibilism is just a way of saying that a puppet is free as long as he loves his strings.

Well, there are definitely strings from our evolutionary history. And natural selection has generally produced beings who love them. The ones that don't tend to go extinct. In fact, in *Just Deserts*, Dan agrees with this and says, "I have adopted [this] sentence and reinterpreted it as indeed a pretty good definition of free will. … If you are lucky enough to be a responsible agent, you have an obligation to *love your strings*, protecting them from would-be puppeteers."

- Compatibilists tend to push back here. They say even if our thoughts and actions are the products of unconscious causes, they are still our thoughts and actions. Anything that your brain does or decides, consciously or not, is something that you have done or decided. So, on this account, the fact that we can't always be aware of the causes of our actions does not negate free will. Our unconscious neurophysiology is just as much us as our conscious thoughts are. But this seems like a bait and switch that trades a psychological fact, the subjective experience of being a conscious agent, for an abstract idea of ourselves as persons. The psychological truth is that most of us feel identical to or in control of a certain channel of information in our conscious minds, but we are wrong about this. The you that you take yourself to be isn't in control of anything.

This is not a bait and switch by compatibilists. It's a holistic understanding of our evolved and embodied selves. What's wrong with that? Sam is the one who insists on fighting a straw man by merely picking on the worst kind of dualist, Cartesian, libertarian free will.

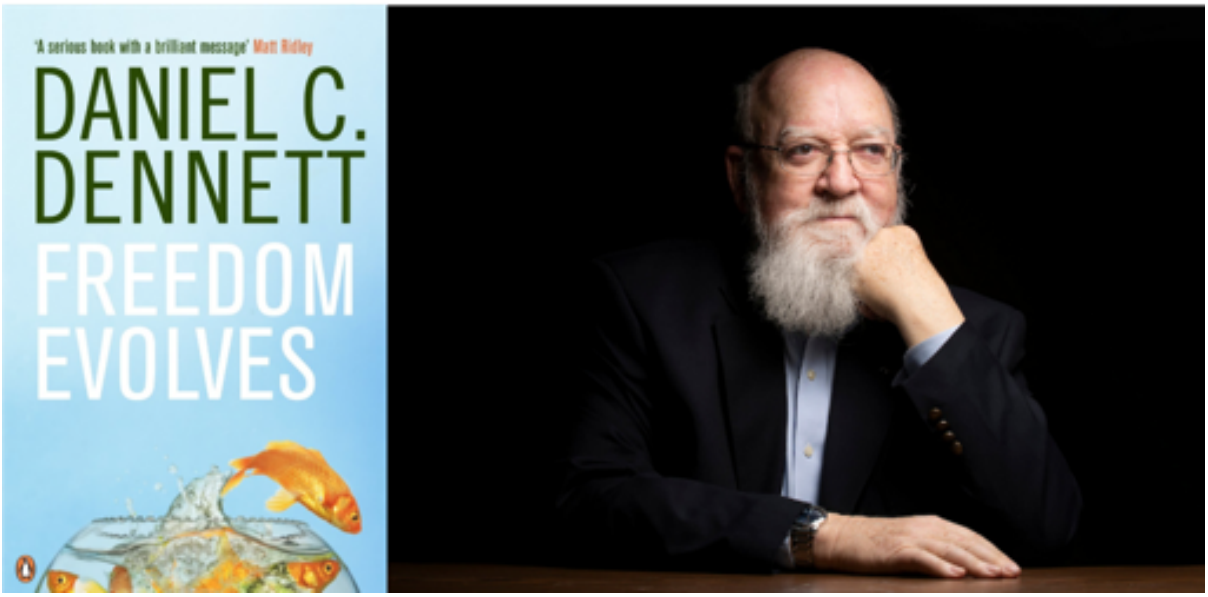
- Compatibilists try to save free will by asserting that you are more than your conscious self. You're identical to the totality of what goes on inside your body, whether you are conscious of it or not. But you can't honestly take credit for your unconscious mental life.

> In fact, you are making countless decisions at this moment with organs other than your brain, but you don't feel responsible for these decisions. Are you producing red blood cells right now? If your body decided to stop doing this, you would be the victim of this change, not its cause. To say that you are responsible for everything that goes on inside your skin because it's all "you" is to make a claim that bears absolutely no relationship to the feelings of agency and moral responsibility that have made the idea of free will a problem for philosophy in the first place.

And to treat red blood cell production the exact same way you treat conscious deliberation by human beings is (**as I said in my last post**) to sink to a level of dehumanisation that is truly troubling. To say *no one* is responsible for *anything* that goes on inside your skin also bears absolutely no relationship to the feelings and facts that have made free will a problem for philosophy. Guess what. We aren't responsible for it all. And we aren't responsible for nothing. Let the hard work of philosophy begin.

Okay, that's enough from Sam. He has helped me see more issues that need to be discussed, but it's time for me to put them all on the table in my next and final post in this short series about free will.

Summary of Freedom Evolves



25 June 2021

Time to get back to the subject of free will. If you remember, **I reviewed *Just Deserts* by Dan Dennett and Gregg Caruso for 3 Quarks Daily** back in March. Then, I shared **some passages** that didn't make the final cut for that article. Next, while I was on this topic, I reviewed parts **1** and **2** of Scott Barry Kaufman's debate with Sam Harris about free will. And finally, I shared **some thoughts on Sam Harris' "Final Thoughts on Free Will"** (that was the title of a podcast he posted in March). I finished that last post by saying:

"Okay, that's enough from Sam. He has helped me see more issues that need to be discussed, but it's time for me to put them all on the table in my next and final post in this short series about free will."

Well, as I began writing up that last post, I decided I really needed to go back and read Dan Dennett's full book from 2003, ***Freedom Evolves***. I had read several of his papers on free will, and I'd read *Just Deserts* very closely (which Dan himself **tweeted** was his "latest and best defense" of his position on free will), and I basically found that I agreed with Dan that free will is not the magic libertarian thing that many ordinary folks believe in. But neither is it the fatalistic determinism that these folks see as the only other choice. Instead, there is something in between these extremes where more and more degrees of freedom have evolved into something that explains the phenomenology of what we experience, which Dan calls "the kind of free will worth wanting." I think I have a few things to add to Dan's position on this, some details which make it clearer, but I needed to go check *Freedom Evolves* to be sure. So, here are the main quotes (about free will) that I pulled from that book, along with just a few comments from me about them as well.

- p. 25 Determinism is the thesis that "there is at any instant exactly one physically possible future" (Van Inwagen 1983, p.3).

This is the succinct definition that Dan lays out at the beginning of the book which all naturalist / physicalist / materialist philosophers must recon with. This is really the crux of

the free will issue. We humans feel that we have alternatives and that we make choices, but if there is only one physically possible future, then how real are these choices? If they are not real, then is free will really just an illusion?

- p. 59 [Dennett's imaginary foil Conrad says,] "Determined avoiding isn't *real* avoiding because it doesn't actually change the outcome." [Dennett replies:] From what to what? The very idea of changing an outcome, common though it is, is incoherent—unless it means changing the *anticipated* outcome. ... The *real* outcome, the *actual* outcome, is whatever happens, and nothing can change *that* in a determined world—or in an undetermined world!

Dan is making the point here that we cannot change the past, and we cannot accurately anticipate the future. So, a determined world feels exactly the same as an undetermined world and we shouldn't get so worked up about which one we are in. But what struck me from this passage was the question of *whose* prediction are we talking about here? If no one is actually able to anticipate the future (more on this later), then the determined outcome is literally non-determined. Ahead of time, no one has actually determined it. Therefore, to worry about determinism is like worrying about someone who never reveals their guesses about the future but still annoyingly insists on repeating after the fact, "I knew you were going to do that. I knew you were going to do that."

- p. 75 Now that we have a clearer understanding of possible worlds, we can expose three major confusions about possibility and causation that have bedeviled the quest for an account of free will. First is the fear that determinism reduces our possibilities.

That's right. Determinism doesn't remove any of the possibilities that have been opened up by previous actions in the universe.

- p. 84 Philosophers who assert that under determinism S* "causes" or "explains" C miss the main point of causal inquiry, and this is the second major error. *In fact, determinism is perfectly compatible with the notion that some events have no cause at all.*

What Dan really means here is that some events have no *known* singular cause. He uses some examples like stock market fluctuations or legal cases where there are multiple attempted murderers to show that many events are simply overdetermined by several various things, which makes it impossible *for us* to say that any *one* thing caused the event.

- p. 88 Consider a man falling down an elevator shaft. Although he doesn't know exactly which possible world he in fact occupies, he does know one thing: He is in a set of worlds *all* of which have him landing shortly at the bottom of the shaft. Gravity will see to that. Landing is, then, *inevitable* because it happens in every world consistent with what he knows. But perhaps *dying* is not inevitable. Perhaps in some of the worlds in which he lands headfirst or spread-eagled, say, but there may be worlds in which he lands in a toes-first crouch and lives. There is some elbow room.

That last sentence is, of course, a reference to Dan's 1984 book ***Elbow Room: The Varieties of Free Will Worth Wanting***, which argues that our human-specific evolutionary history has carved out quite a lot of (elbow room) space for decisions to be made beyond the determinable knee-jerk reactions of simpler animals. This sounds great, but what doesn't get emphasized from Dennett is that this perceived freedom is perhaps just due to ignorance. Could a super-intelligent being from another world scan the entire life history of

the man in the falling elevator and know for sure that he will try a toes-first crouch because he once saw it in a movie as a teenager? Sure. I guess that's possible. Does that matter to the choice the man is trying to make as he is falling towards his potential death? It shouldn't, because that man cannot possibly know about it.

- p. 89 At last, we are ready to confront the third major error in thinking about determinism. Some thinkers have suggested that the truth of determinism might imply one or more of the following disheartening claims: All trends are permanent, character is by and large immutable, and it is unlikely that one will change one's ways, one's fortunes, or one's basic nature in the future.

Well, those thinkers are just making an obvious error. A fixed future doesn't mean an unchanging future. It just means that the changes are conceivably all knowable ahead of time. So, no one should have **a fixed mindset vs. a growth mindset**.

- p. 91 Every finite information-user has an epistemic horizon; it knows less than everything about the world it inhabits, and this unavoidable ignorance guarantees that it has a *subjectively* open future. Suspense is a necessary condition of life for any such agent.

Coming back to the point made above, Dan is showing how our ignorance about the future is always guaranteed.

- p. 91 Footnote 6 Laplace's demon instantiates an interesting problem first pointed out by Turing, and discussed by Ryle (1949), Popper (1951), and McKay (1961). No information-processing system can have a complete description of itself—it's Tristram Shandy's problem of how to represent the representing of the representing of…the last little bits. So even Laplace's demon has an epistemic horizon and, as a result, cannot predict its own actions the way it can predict the next state of the universe (which it must be outside).

So, in fact, that ignorance is a logical fact of every enclosed system. Nothing can get outside of everything it knows in order to truly know everything that *might* affect it. Therefore, not even **Laplace's demon** could determine the future of its determined universe. And that kind of ignorance is vital to our feelings of freedom. This ends up being similar to something I said in my article "**Mortality Doesn't Make Us Free Either**":

"If there is any hope for the kind of spiritual freedom that Hägglund longs for, it could only be in the epistemological uncertainty that exists between certain mortality and certain immortality."

- p. 92 Do fish have free will, then? Not in a morally important sense, but they do have control systems that make life-or-death "decisions," which is at least a necessary condition for free will.

This hints at the evolutionary development of free will, which I intend to expand upon in my next post in a way that also aligns it with **my summary of the development of consciousness**. Furthermore, according to **my view of evolutionary ethics**, these "morally important" decisions are *all* life-or-death decisions. We humans are just able to consider longer time horizons and wider circles of moral concern. But the decisions we make are still moral or immoral if they lead to more robust or more fragile survival. (That's my argument anyway. Lots of moralizers can be mistaken about what they think is moral or immoral.)

- p. 94 The question that interests me: Could Austin have made *that very* putt? And the answer to that question must be "no" in a deterministic world.

Correct. But no one knows which putts will be missed ahead of time, so we still plan and try to make them. And we learn from misses about what to do differently the next time we are in similar situations.

- p. 122 If you make yourself really small, you can externalize virtually everything. [See Footnote 6]
- p. 122 Footnote 6 This was probably the most important sentence in *Elbow Room* (Dennett 1984, p. 143), and I made the stupid mistake of putting it in parentheses. I've been correcting that mistake in my work ever since, drawing out the many implications of abandoning the idea of a punctate self.

Great point. This is exactly the trap that Sam Harris falls into when he refuses to see consciousness as embedded in our entire bodies with lots of unconscious processing. He has a very tiny (dualistic?) view of the self.

- p. 125 The idea that someone who has been tested by serious dilemmas of practical reasoning, who has wrestled with temptations and quandaries, is more likely to be "his own man" or "her own woman," a more responsible moral agent than someone who has just floated happily along down life's river taking things as they come, is an attractive and familiar point, but one that has largely eluded philosophers' attention.

This is a great point that philosophers would not miss if they used the evolutionary framework of a Tinbergen analysis. The personal development of every individual (their ontogeny) is a vital part of the whole story of the development of free will.

- p. 127 We should quell our desire to draw lines. We don't need to draw lines. We can live with the quite unshocking and unmysterious fact that, you see, there were all these gradual changes that accumulated over many millions of years and eventually produced undeniable mammals. Philosophers tend to take the idea of stopping a threatened infinite regress by identifying something that is—must be--*the* regress-stopper: the Prime Mammal, in this case. It often lands them in doctrines that wallow in mystery, or at least puzzlement, and, of course, it commits them to essentialism in most instances.

Great passage! This is drawn out much further in Dan's paper about **Darwinism and the overdue demise of essentialism**.

- p. 135 Where is the misstep that excuses us from accepting the [incompatibilist's] conclusion? We can now recognize that it commits the same error as the fallacious argument about the impossibility of mammals. Events in the *distant* past were indeed not "up to me," but my choice now to Go or Stay is up to me because its "parents"—some events in the *recent* past, such as choices I have recently made—were up to me (because *their* "parents" were up to me), and so on, not to infinity, but far enough back to give my *self* enough spread in space and time so that there is a *me* for my decision to be up to! The reality of a moral me is no more put in doubt by the incompatibilist argument than is the reality of mammals.

This points out how incompatibilists attempt to rely on a version of **the Sorites paradox** to make their case, but that is an unsolved paradox for a reason! Imagine if I tried to start with

the claim that I am responsible for my decisions, and then went back and back and back and back, claiming my responsibility continued to hold for each small step along the way, until eventually I took responsibility for the Big Bang. That is of course nuts. But that is essentially the exact same logic that incompatibilists are using on their side of the argument. They are just using it in the opposite direction. But if that trick doesn't work for me, then it doesn't work for them either. A new approach to the problem must be used. (Read the link above on the Sorites paradox to see a glimpse into an approach informed by evolutionary logic.)

- p. 223 Love is quite real, and so is falling in love. It just isn't what people used to think it is. It's just as good—maybe even better. True love doesn't involve any flying gods. The issue of free will is like this. If you are one of those who think that free will is only *really* free will if it springs from an immaterial soul that hovers happily in your brain, shooting arrows of decision into your motor cortex, then, given what *you* mean by free will, my view is that there is no free will at all. If, on the other hand, you think free will might be morally important without being supernatural, then my view is that free will is indeed real, but just not quite what you probably thought it was.

This is an excellent synopsis of Dan's argument. And it is basically consistent with his strategy for consciousness too. He says *folk* notions of consciousness are an illusion, just as *folk* notions of free will are an illusion. I believe he's right that our definitions of these terms must evolve.

- p. 223 In my book *Brainstorms*, one of the questions discussed was whether such things as *beliefs* and *pains* were "real," so I made up a little fable about people who speak a language in which they talk about being beset by "fatigues" where you and I would talk about being tired, exhausted. When we arrive on the scene with our sophisticated science, they ask us which of the little things in the bloodstream are the fatigues. We resist the question, which leads them to ask, in disbelief: "Are you denying that fatigues are real?" Given their tradition, this is an awkward question for us to answer, calling for diplomacy (not metaphysics).

This is a great example of the confusion that arises when **Western languages use too many nouns**. As I said in my review of *Just Deserts*, "We may not *have* free will, but we *are* a will with an infinite degree of freedom (subject to certain restrictions)." It may help somewhat to consider this issue as the act of a verb.

- p. 225 I claim that the varieties of free will I am defending are worth wanting precisely because they play all the *valuable* roles free will has been traditionally invoked to play. But I cannot deny that the tradition also assigns properties to free will that my varieties lack. So much the worse for tradition, say I.

Yep! The tradition must evolve.

- p. 237 The conclusion Libet and others should draw is that the 300-millisecond "gap" has *not* been demonstrated at all. After all, we know that in normal circumstances the brain begins its discriminative and evaluative work as soon as stimuli are received, and works on many concurrent projects at once, enabling us to respond intelligently just in time for many deadlines, without having to stack them up in a queue waiting to get through the turnstile of consciousness before evaluation begins.

Yep again! I was very glad to see this as I independently arrived at the same conclusion in **my**

post about Libet. Good evolutionary thinking leads to the same places.

- p. 238-9 Conscious decision-making takes time. If you have to make a series of conscious decisions, you'd better budget half a second, roughly, for each one, and if you need to control things faster than that, you'll have to compile your decision-making into a device that can leave out much of the processing that goes into a stand-alone conscious decision.

I thought this was an interesting precursor to Daniel Kahneman's ***Thinking, Fast and Slow***.

- p. 243 As David Hume pointed out so vigorously several centuries ago, you can't *perceive* causation. You can't see it when it happens outside, and you can't introspect it when it happens inside.

Excellent observation.

- p. 273 A proper human self is the largely unwitting creation of an interpersonal design process in which we encourage small children to become communicators and, in particular, to join our practice of asking for and giving reasons, and then reasoning about what to do and why.

This is a nice point to make about our ontogeny. Morality concerns others. It is built by them too. We could not develop selves or morality in isolation.

- p. 279 The hard determinists say that our world would be a better place if we could somehow talk ourselves out of feeling guilty when we cause harm and angry when harm is done to us. But it isn't clear that any feasible "cure" along these lines wouldn't be worse than the "disease." Anger and guilt have their rationales, and they are deeply embedded in our psychology.

My analysis of **what causes our emotions** adds a lot of details to clarify this. Emotions (when they are working properly) do arise from reasons and we would be wise to recognize and hold on to the good reasons while discarding any poorly driven emotional responses. Properly aimed anger and guilt help shape individuals and societies to act towards more robust survival. Determinists think we can eliminate these and other emotions tied to notions of free will, but it is only the mistaken supernatural beliefs that need to go.

- p. 287 The self is a system that is *given* responsibility, over time, so that it can reliably be there to *take* responsibility, so that there is somebody home to answer when questions of accountability arise. Kane and the others are right to look for a place where the buck stops.

This is a nice description of how free will and moral responsibility are socially constructed in a bi-directional manner.

- p. 290 We now uncontroversially exculpate or mitigate in many cases that our ancestors would have dealt with much more harshly. Is this progress or are we all going soft on sin? To the fearful, this revision looks like erosion, and to the hopeful it looks like growing enlightenment, but there is also a neutral perspective from which to view the process. It looks to an evolutionist like a rolling equilibrium, never quiet for long, the relatively stable outcome of a series of innovations and counter-innovations, adjustments and meta-

> adjustments, an arms race that generates at least one sort of progress: growing self-knowledge, growing sophistication about who we are and what we are, and what we can and cannot do.

Yes! This makes for a good summary of the evolutionary steps that both free will and our understanding of it take. Next time, I'll do my best to help grow that knowledge and sophistication just a tiny bit further.

Not My Final Thoughts on Free Will

In case you haven't been following Sam Harris closely, that title for this post is a subtle dig at Sam's "**Final Thoughts on Free Will**" podcast back in March. Evolutionary thinkers can never (as far as we know) claim to have reached a final truth, so they ought not to say they've ever reached a "final position" on any topic. However, we do come to conclusions *for now*, and it is time now for me to wrap up my posts on free will. As a quick reminder, that series has included:

- **My Review of *Just Deserts* by Daniel Dennett and Gregg Caruso**
- **A Few Further Thoughts on *Just Deserts***
- **Another Free Will Debate — Kaufman v. Harris (Part 1/2)**
- **Another Free Will Debate — Kaufman v. Harris (Part 2/2)**
- **Some Thoughts on Sam Harris' Final Thoughts on Free Will**
- **Summary of *Freedom Evolves***

If you read the 17,500 words in all those posts, you'll have seen that there is already a large zone of agreement on this issue between hard incompatibilists like Caruso and Harris and compatibilists like Dennett, Kaufman, and myself. From my review of *Just Deserts*:

- Both are naturalists (*JD* p.171) who see no supernatural interference in the workings of the world. That leaves both [sides] accepting general determinism in the universe (*JD* p.33), which simply means all events and behaviours have prior causes. Therefore, the libertarian version of free will is out. Any hope that humans can generate an uncaused action is deemed a "non-starter" by Gregg (*JD* p.41) and "panicky metaphysics" by Dan (*JD* p.53). Nonetheless, both agree that "determinism does not prevent you from making choices" (*JD* p.36), and some of those choices are hotly debated because of "the importance of morality" (*JD* p.104). Laws are written to define which choices are criminal offenses. But both acknowledge that "criminal behaviour is often the result of social determinants" (*JD* p.110) and "among human beings, many are extremely unlucky in their initial circumstances, to say nothing of the plights that befall them later in life" (*JD* p.111). Therefore "our current system of punishment is obscenely cruel and unjust" (*JD* p.113), and both [sides] share "concern for social justice and attention to the well-being of criminals" (*JD* p.131).

My previous six posts also led to this conclusion in my summary of *Freedom Evolves*:

- I basically found that I agreed with Dan that free will is not the magic libertarian thing that many ordinary folks believe in. But neither is it the fatalistic determinism that these folks see as the only other choice. Instead, there is something in between these extremes where more and more degrees of freedom have evolved into something that explains the phenomenology of what we experience, which Dan calls "the kind of free will worth wanting." [And] I think I have a few things to add to Dan's position on this, some details which make it clearer.

Another way to see the need for this compatibilist conclusion would be to look at a word cloud for all of the issues that get discussed during free will debates. I don't have the time or resources to put lots of relevant texts into a computer program that would generate such a cloud showing the frequency with which each idea is used, but I did at least gather a list of

many of the relevant concepts while I was going through the books and papers and interviews I've covered in this series. Please don't read this entire list, but a quick scan is helpful:

---

choice, options, coercion, fatalism, emotion, reactive attitudes, reasons, learning, neuroplasticity, brain chemicals, neural wiring and firing, causation, manipulation, prediction, inevitability, evitability, personal responsibility, attributable, answerable, accountable, causal, individualism vs. communitarian ties and influences, motivation to act and to be responsible, personal behaviour, personal judgment of self, personal judgment of others, societal judgment of others, Justice System, determined by forces beyond our control, predetermined all the way back to the Big Bang, praise, blame, guilt, pride, gratitude, resentment, punishment and reward, suffering, the "mitigation effect" of realising there is causal determinism in the universe, romantic love, friendship, civility, voluntary, involuntary, self, self-control, Cartesian dualism, homunculus, mind, robots, puppets, Turing tests, consciousness, meaning, planning, hoping, promising, goals, desires, intentions, agency, morality, moral responsibility, Prisoners' dilemma, Iterated Prisoners' dilemma, Sorites paradox, vagueness, fuzzy boundaries, Tinbergen questions, folk notions of free will vs. folk notions of determinism, psychology definitions of free will, free action, agency, voluntary control, evil, complex belief systems, brain tumours, pragmatism, experimental outcomes for different beliefs in free will, luck, constitutive luck, present luck, moral luck, luck swallows all, control, self-control, locus of control, reasons-responsiveness, the ability to do otherwise, freedom, degrees of freedom, moral assessment, slow revision vs. sudden revolution, desert, just desert, basic-desert moral responsibility, familiar sense of desert, God's gift of free will, punishment, limiting liberty, public health-quarantine model, consequences, prevention, deterrence, retribution, restoration, infinite regress, demonisation, fixed or growth mindset, respect for law, stability in a state

---

Anyone trying to carve a neat and tidy definition of free will out of that mess—either to reject free will or to accept it—will forever be faced with a bunch of "**whataboutism**" from people holding other positions. There are just too many concepts bound up here. Any simple affirmation or denial of the phrase "free will" is going to feel too blunt to cover it all. To me, following the standard playbook of analytical philosophy and "defining one's terms" just is not going to get us very far. Consider the following quotes from the world of biology where free will is clearly located. (My emphases added in bold.)

- "In the distant future I see open fields for far more important researches. Psychology will be based on a new foundation, that of **the necessary acquirement of each mental power and capacity by gradation**." (Charles Darwin, *On the Origin of Species*)
- "Neither Mayr nor Tinbergen provide a detailed account of how to integrate different areas of biological inquiry, but both provide enough discussion to make it clear that they have in mind a general practice that philosophers of science have characterized in some detail under the label 'functional analysis'. The canonical account of this practice among philosophers of science is Robert Cummins' (1975, 1983) account, according to which **functional analysis consists in breaking down some capacity or disposition of interest into simpler dispositions or capacities, organized in a particular way**." (**Conley**)
- "Reduction, unlike analysis, ignores a system's organization (1982), which Mayr characterizes as the interaction between components (Mayr 2004). Organization explains the emergence of new characteristics that could not be predicted from knowledge of the isolated components of a system, but analysis provides a middle ground between reductionism and holism (Mayr 1982). Mayr claims that '**all problems of biology, particularly those relating to emergence, are ultimately problems of hierarchical organization**' (Mayr, 1982, p. 64)." (**Conley**)

So, for free will, we need a deep "functional analysis" where elements of that emerging property are listed out for separate consideration. In this way, nuances can be captured and

lassoed into an evolving understanding of all the issues. Now, where have we seen a hierarchical organisation of a complicated emergent biological process before?? Hmmm. This quote from one of **Dan Dennett's papers** should help you remember:

- "It is no mere coincidence that the philosophical problems of consciousness and free will are, together, the most intensely debated and (to some thinkers) ineluctably mysterious phenomena of all. As the author of five books on consciousness, two books on free will, and dozens of articles on both, I can attest to the generalization that you cannot explain consciousness without tackling free will, and vice versa."

In my nearly finished series on consciousness (**summarised here**), I explained how a Tinbergen analysis is the proper way to explore and explain that complex emergent phenomenon. And since free will and consciousness are so tied together, a Tinbergen analysis is useful here too. This is the extra detail I would add to the free will debate beyond Dan Dennett's generally excellent contributions that I have discussed so far. I hinted at this in **my review of *Just Deserts*** with the following passages:

- [M]ost philosophers [rely] on classical logic, which says *A* is *A*, *not-A* is *not-A*, and **the law of the excluded middle** says there is nothing else possible in between. Such rigid definitions work well in the precise worlds of mathematics and Newtonian physics, but not in the fuzzy world of biology. In that realm, the ethologist Nikolaas Tinbergen gave us his **Four Questions** which are now the generally accepted framework of analysis for all biological phenomena. To understand anything there, Tinbergen says you have to understand its function, mechanism, personal history (ontogeny), and evolutionary history (phylogeny). As a very simple example, philosophers could tie themselves in knots trying to define 'a frog' such that this or that characteristic is *A* or *not-A*, but it's just so much clearer and more informative to include the stories of tadpole development and the slow historical diversion from salamanders. So, is free will more like a geometry proof or a frog?
- Tinbergen's perspective gives us a few additional tricks. It isn't luck that I grew up to be a person rather than a horse. Once I was conceived, the evolutionary history (phylogeny) that led up to me put a lot of constraints on my personal development (ontogeny). Luck may explain all the *differences* between me and every other person out there, but we needn't worry about luck when describing all the things we have *in common*. There are hordes of characteristics that all humans share, but the one that is most important for this debate is our capacity to *learn*. The extreme neuroplasticity we have (a mechanism of free will) is what enables all but the most unfortunate humans to sense and respond to their environments (a function for free will) to the point where they slowly, slowly become a unique self.

For details on how I developed answers to Tinbergen's four questions for consciousness, you need to see posts **18** (Tinbergen), **19** (Functions), **20** (Mechanisms), **21** (Ontogeny), and **22** (Phylogeny). Luckily, there's no need to go into so much depth for free will now. Since the groundwork has been laid for consciousness, a quick sketch will suffice to show how free will folds very neatly into this view and then expands perfectly logically during the developments of consciousness. Essentially, it is clear that degrees of freedom only open up for living organisms, and they expand along as more and more levels of consciousness are developed. I don't expect that to sound controversial, but the details are hopefully helpful to the discussion.

TINBERGEN ANALYSIS OF FREE WILL WITHIN HIERARCHIES OF CONSCIOUSNESS

Unifying concept: subjects sensing and responding to the world for the ultimate goal of the survival of life.
Governing Evolutionary Laws — Natural Selection & Sexual Selection
Guiding Biological Forces — Consumption, Predation, Niche Competition, Conspecific Rivalry, Potential Invasion

Hierarchies of Consciousness	Functions	Phylogeny (Appearance in History)	Ontogeny (Develop in Self)	Mechanisms	Fulfilling the Evolutionary Hierarchy of Needs
1. Origin of Life	No Free Will at First	***	***	***	Existence
2. Affect	In moment reflex choice of good/bad	Once life established	Innate valence	Molecular Forces ("Pandynamism")	Durability
3. Intention	Choice of present based on past	Complex multicellularity	Learned reactions	Chemical messages / nerve systems	Interactions
4. Prediction	Choosing between alternate futures	Brains	Improving guesses	Cortex	
5. Awareness	Decisions based on wider view of self	Vertebrates, cephalopods	Individual goals and preferences	Local modules / global signals	Identity
6. Abstraction	Expands freedom to abstract influences	Human language	Episodic and strategic planning	Specific neocortex regions	Purpose

I think it's easiest to grasp this table by focusing on the Functions column. Going from top to bottom, there is (1) no free will before the emergence of life. Once (2) life is established, the phenomenon of affect provides innate valences for making in the moment reflex choices between good or bad options for life. As (3) complex multicellularity develops mechanisms to learn and act on (unconscious at first) intentions, then life gains the freedom for choosing different actions in the present based on things it has learned in the past. Continuing on, the (4) development of brains enables modelling predictions of the world, which gives life freedom to choose between alternate futures. As all of these abilities lead to (5) the dawning of self-awareness, living organisms can begin to develop autobiographical narratives that inform choices over longer and longer time horizons depending on the quantity and quality of memories and predictions that have been developed. Finally, in the (6) realm of human language, we *Homo sapiens* have gained the freedom to be influenced by an infinite array of abstract representations. At this level, we can now see strategic planning of actions for decades of a life, which clearly drives the feelings of free will that exist in folk psychology.

This brief rundown does not begin to address all of the items in the word cloud shown above for the free will debate. But I've already touched on most or all of those in my other posts, so hopefully this final summary just provides a "hierarchical organisation of capacities" (*a la* Mayr via Conley above), which helps us see the slow step-by-step emergence of degrees of freedom that starts from absolutely nothing but eventually grows to the enormous range that philosophers have contemplated for millennia. Slapping a line on this chart and declaring "here lies free will" or "you must be taller than this degree of freedom in order to be free" would seem to be a very silly exercise. Yet that appears to be what people do when they declare "free will" to absolutely exist or not. Taking all of the facts together, however, by using a "functional analysis" that is typical of the philosophy of science, there is hopefully now a bit more grandeur in the evolutionary view of the emergence of free will.

ADDENDUM

The FAQs of Consciousness and Free Will

3 October 2021

Here it is! Finally, after 19 1/2 months, I've reached the end of my series on consciousness and free will. This project started on a bit of a whim when I was looking for something interesting to dig into during the Covid lockdown. But I also had a hunch that there were some big evolutionary ideas to uncover about this topic. I had been listening to a lot of podcasts about consciousness and I felt like the time was right for a quick exploration.

Boy was I wrong!

This has been by far the hardest philosophical topic that I've focused on during my 10 years of writing. And after all that, I shared the **summary of my evolutionary theory** in my last post about consciousness. I think this could really make an important contribution (no one I know of has attempted a Tinbergen analysis of this phenomena before), but did that summary answer all of the questions about this topic? Hardly! So that's what I'm sharing here now to wrap up this series and finally turn my attention to other things.

During my research I gathered a huge list of questions that typically arise about consciousness. I whittled them down and felt they could best be organised into 5 groups: introductory questions, those from impartial sources, those coming from other naturalists, questions coming from those who doubt or disbelieve naturalism, and finally the many questions that have come from David Chalmers. Answering these questions in this order takes us on the best journey, but my answers ended up filling 43 pages with over 23,000 words. That's a lot even for me!

Rather than string these out over several digestible posts, I decided it was better to be able to see all the questions at once. The answers are provided after that, so you can go and read them in whatever way you prefer, in whatever order you like, and on as many questions as you care about.

Thanks to everyone who came along with me on this journey. As always, questions and comments are very much appreciated in the comment section. I hope you enjoy reading this as much as I did writing it!

INTRODUCTORY QUESTIONS

1. What would a good definition of consciousness look like?
2. What's your definition?

QUESTIONS FROM IMPARTIAL SOURCES

3. Why do we think consciousness is a physical phenomenon?
4. How could minds possibly arise from matter?
5. Does consciousness contain non-physical information?
6. And what about Hume's missing shade of blue?
7. Is consciousness so mysterious that it is beyond our ability to understand it?
8. What about Zombies?
9. How is our conscious experience bound together?
10. What can the neural correlates of consciousness tell us?
11. Are other animals conscious?
12. Can machines be conscious?
13. So, "what is it like" to be conscious?
14. Do we have immortal souls?
15. Do we have free will?

QUESTIONS FROM OTHER NATURALISTS

16. Can't we just get by with a very rough definition of consciousness?
17. What about the various parts of living systems? Which ones are conscious?
18. Is the United States conscious?
19. How do we know we don't have "inverted qualia"?
20. How do you solve the mind-evolution problem?
21. Does consciousness have a purpose?

QUESTIONS FROM THOSE WHO DOUBT OR DISBELIEVE NATURALISM

22. Why doesn't a chair feel my bottom?
23. How can consciousness survive sleep?
24. How could consciousness have possibly emerged from lower organisms?
25. Is conscious experience outside of the realm of science?
26. Are minds everywhere? What about panpsychism?

QUESTIONS FROM DAVID CHALMERS

27. What are the easy problems of consciousness?
28. What is the hard problem of consciousness?
29. What does it take to solve the easy problems of consciousness?
30. Is the hard problem really different than the easy ones?
31. Can we see an example? Is the binding problem hard or easy?
32. How have people tried to answer the hard problem?
33. So, what else is needed and why do physical accounts fail?
34. Is this the same problem we faced with vitalism?
35. So, is consciousness just fundamental?
36. If we accept consciousness is fundamental, then what?
37. Is this fundamental view a sort of dualism?
38. If consciousness is fundamental, shouldn't it be simple to describe?
39. What about Chalmers' own theories?
40. Is consciousness all about information processing?
41. So, can we make progress and answer the hard problem of consciousness?

INTRODUCTORY QUESTIONS

1. What would a good definition of consciousness look like?

First off, this is not going to be a simple thing. In the 16th post in this series, I went through **a (sorta) brief history the definitions of consciousness**. As I noted there, these have ranged "all the way from it being something as small as the private, ineffable, special feeling that only we rational humans have when we think about our thinking, right on down to it being a fundamental force of the universe that gives proto-feelings to an electron of what it's like to be that electron." The Stanford Encyclopedia of Philosophy entry on **consciousness** stated, "There is unlikely to be any single theoretical perspective that suffices for explaining all the features of consciousness that we wish to understand. Thus, a synthetic and pluralistic approach may provide the best road to future progress." And as Dan Dennett noted in **one of my favourite papers**, among many philosophers, their typical "demand for essences with sharp boundaries blinds thinkers to the prospect of gradualist theories of complex phenomena, such as life, intentions, natural selection itself, moral responsibility, and consciousness."

So, I believe it's clear we ought to be looking for a gradualist theory of the emergence of all the complex phenomena associated with consciousness. The famed evolutionary biologist Ernst Mayr who coined the proximate/ultimate distinction **claimed** that "all problems of biology, particularly those relating to emergence, are ultimately problems of hierarchical organization." Thus, trying to reduce consciousness to a single thing is impossible, but that still leaves open the possibility of analysis. What kind of analysis? The philosopher of science Robert Cummins gave **the canonical account** of a *functional analysis*, which "consists in breaking down some capacity or disposition of interest into simpler dispositions or capacities, organized in a particular way."

Therefore, we need a deep functional analysis of consciousness where elements of that emerging property are listed out for separate consideration. In this way, nuances can be captured and lassoed into an evolving understanding of all the issues.

2. What's your definition?

That's a tough question, but I wrote a full summary of my response in **post 23** of this series. In that post, I started with the background of my metaphysical hypotheses, which is just standard naturalism. Then, I laid out the theories that I like best for the two biggest mysteries for this topic — 1) the emergence of life, which I accept as happening using something like the RNA-world hypothesis, and 2) the hard problem of consciousness, which I think is most simply explained using my hypothesis of *pandynamism* (see question 4 below for details). Once chemistry makes the jump to biology, then the resulting proto-lifeforms have a defined self *and* they begin to compete for resources with other potential entrants, substitutes, or conspecifics in order to self-replicate and survive. They react to the world as if they know what they are *and* what they need. These are the building blocks for expanding the properties of subjective experience. Thus:

Consciousness, according to this evolutionary theory, is an infinitesimally growing ability to sense and respond to any or all biological forces in order to meet the needs of survival. These forces and needs can vary from the

immediate present to infinite timelines and affect anything from the smallest individual to the broadest concerns (both real and imagined) for all of life.

This is intended to be a comprehensive and therefore very broad definition. Anything that is able to act to remain alive does so using aspects of consciousness. There are infinite varieties of scope and scale within this definition, so in order to map these contours I spent several posts conducting a **Tinbergen analysis** of the **functions**, **mechanisms**, **ontogeny**, and **phylogeny** of consciousness. This is the standard procedure in evolutionary studies for coming to know all of the elements of any biological phenomenon, and I believe it is therefore the best method to perform a *functional analysis* as described above in question 1. The hierarchical organisation that emerged from this review is supported by logical requirements as well as empirical data from across the history of all life. That hierarchy is:

1) Origin of Life
2) Affect
3) Intention
4) Prediction
5) Awareness
6) Abstraction

In order to further elaborate this definition of consciousness, I finished my summary post by providing definitions of the following common terms in consciousness studies, which sometimes differ between technical and folk usages:

- Accessible, Attention, Bottom-up vs. Top-down, Cognition, Communication, Conscious vs. Unconscious, Emotions vs. Feelings, Evolutionary Hierarchy of Needs, Evolutionary Epistemology Mechanisms, Exteroception vs. Interoception, Intentionality vs. Intentional Stance, Involuntary vs. Voluntary, Language, Mind, Qualia vs. Something-it-is-like vs. Subjective Experience.

All of this is in keeping with the epigraph for *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness* by Peter Godfrey-Smith. That epigraph came from William James in *The Principles of Psychology* from 1890:

- "The demand for continuity has, over large tracts of science, proved itself to possess true prophetic power. We ought therefore ourselves sincerely to try every possible mode of conceiving the dawn of consciousness so that it may not appear equivalent to the irruption into the universe of a new nature, non-existent until then."

Godfrey-Smith **has been interpreted** as saying that "we need a theory [of consciousness] based on continuities and comprehensible transitions; no sudden entrances or jumps." This, of course, aligns with Darwin's observation that nature does not jump, and I believe my theories and definitions of consciousness fit this requirement.

QUESTIONS FROM IMPARTIAL SOURCES

Now that I've laid out what a theory of consciousness should look like and what my particular theory is, let's see how that addresses the standard objections raised against other theories of consciousness. Particularly, these questions come from the Stanford Encyclopedia of

Philosophy entry on **Naturalism** by David Papineau and the even more pointed Internet Encyclopedia of Philosophy entry on **Consciousness** by Rocco Gennaro.

3. Why do we think consciousness is a physical phenomenon?

According to **Papineau**, a majority of contemporary philosophers hold that physicalism will be able to explain consciousness, although a significant minority take two other options. The first is that "conscious properties are 'epiphenomenal' and do not exert any influence on brain processes or subsequent behaviour." The second route is "to embrace the 'overdeterminationist' view that the physical results of conscious causes are always strongly overdetermined" by both physical causes and by some other immaterial causes. Papineau declares that neither of these two positions are attractive. He says that they "posit odd causal structures," neither of which are observed anywhere else in nature, so we're not compelled to accept them here.

In my **summary post on consciousness**, I provide a more positive example of why the physicalist explanation for consciousness is more likely to be correct. I wrote:

- The psyche only originates and evolves along with life. This psyche expands as the living structures expand their capabilities of sensing and responding to [biological] forces. And the 'flavour' of experiences within this psyche are utterly dependent upon the underlying mechanisms of *which* particles of matter are being subject to *which* particular forces.
- For example, the retch of disgust from accidentally eating something harmful maps almost exactly onto the retch of moral disgust from accidentally witnessing something beyond the pale such as a mutilated dead body. These experiences come from very different sources, and they process very different bits of information, so we might expect them to feel very different, but we know from neuroscience that the brain has duct-taped the feelings of moral disgust onto the existing architecture for gustatory disgust and that is what explains the similar conscious experience. This is another striking bit of support for a materialist understanding of consciousness.

4. How could minds possibly arise from matter?

Gennaro lists this as the first standard objection to physicalist accounts of consciousness. It usually goes by one of two names. Joseph Levine (1983) coined the expression 'the explanatory gap' as a label for the idea that there is a key gap in our ability to explain the connection between subjective feelings (mind) and brain properties (matter). David Chalmers (1995) described something similar with the catchy phrase 'the hard problem of consciousness', which has come to dominate this discussion. (See an in-depth examination of this in questions 27 to 41.)

This is indeed a problem for the whole project of evolutionary explorations of consciousness. In a paper called "**The Difficulty of Fitting Consciousness in an Evolutionary Framework**", the author Yoram Gutfreund noted that "the question of how the mind emerged in evolution (the mind-evolution problem) is tightly linked with the question of how the mind emerges from the brain (the mind-body problem). It seems that the evolution of consciousness cannot be resolved without first solving the 'hard problem'. Until then, I argue that strong claims about the evolution of consciousness based on the evolution of cognition are premature and unfalsifiable."

In my **post 19 on the functions of consciousness**, I introduced my hypothesis for a solution to this. I wrote:

- The hard problem of consciousness is often phrased as wondering how inert matter can ever evolve into the subjective experience that we humans undoubtedly feel. I think this short-changes matter. Far from being inert, matter responds to the forces exerted on it all the time. Panpsychism says mind (*psyche*) is everywhere. But to me there can be no mind without a stable subject. In my current conception, the forces that minds feel and are shaped by are merely the chemical and physical forces that shape all matter. Until something else is found, what else could there be? So, mind is not everywhere, but forces are. The Greek for force is *dynami*, so rather than panpsychism, I would say the universe has *pandynamism*. The psyche only originates and evolves along with life.

In my **summary post on consciousness**, I further explained this when I wrote:

- We have subjective experience. Evolutionary studies have shown us that there is an unbroken line in the history of life. But water and rocks don't appear to have anything like consciousness. So, how can inert matter ever evolve into the subjective experience that we humans undoubtedly feel? Chalmers has proposed that subjective experience may be a fundamental property of the universe, like the spin of electromagnetism. I have come to accept that as a likely hypothesis. All matter is affected by the forces of physics and chemistry. But until that matter is organised into a living subject that is capable of responding to those forces in such a way as to remain alive, it makes no sense to talk of non-living matter as 'feeling' or 'experiencing' those forces. Inert matter has no structure capable of living through subjective activities. Panpsychism claims that minds (*psyche*) are everywhere, and they don't need physics and matter to exist. But this raises innumerable difficulties, including an enormous change to one's metaphysics that supposedly cannot be detected by science. What I hypothesise instead is that the forces of physics are everywhere, and it is a fundamental property of the universe that these forces are felt subjectively when subjects emerge. Since the Greek for force is *dynami*, I would say the universe has *pandynamism* rather than panpsychism. The psyche only originates and evolves along with life.

As an example, take a very simple force. What does it take to 'feel' gravity? For us humans, it's registering the difference between inner ear liquids as our movements in space accelerate or decelerate. Can a rock or a photon ever experience this? No. Why not? Because there is no structure in its makeup by which it could gain such information. Panpsychism is therefore a non-starter for me, but *pandynamism* could explain how subjectivity is a fundamental feature of the universe, yet only emerges as living organisms emerge, thus bridging the explanatory gap and providing a coherent answer to the hard problem.

Is this enough to overthrow all doubts from metaphysical dualists? Not likely. But Patricia Churchland provided a wonderful quote about this in her essay "**Neurophilosophy**", which was a chapter in the fantastic edited collection ***How Biology Shapes Philosophy: New Foundations for Naturalism***. She wrote:

- A methodological point may be pertinent in regard to the dualist's argument: however large and systematic the mass of empirical evidence supporting the empirical hypothesis that consciousness is a brain function, it is always a logically consistent option to be stubborn and to insist otherwise, as do Chalmers and Nagel. Here is the way to think

about this: identities—such as that temperature really is mean molecular kinetic energy, for example—are not directly observable. They are underwritten by inferences that best account for the mass of data and the appreciation that no explanatory competitor is as successful. One could, if determined, dig one's heels in and say, "temperature is not mean molecular kinetic energy, but rather an occult phenomenon that merely runs parallel to KE." It is a logically consistent position, even if it is not a reasonable position.

Thus, I believe conscious subjectivity appears to be another one of these identities of the universe rather than some occult phenomenon requiring an entirely new metaphysical realm.

5. Does consciousness contain non-physical information?

This is the second common objection according to **Gennaro** and it is usually labelled *The Knowledge Argument*. This is based on "a pair of very widely discussed, and arguably related, objections to materialism which come from the seminal writings of Thomas Nagel (1974) and Frank Jackson (1982, 1986). … The general pattern of each argument is to assume that all the physical facts are known about some conscious mind or conscious experience. Yet, the argument goes, not all is known about the mind or experience. It is then inferred that the missing knowledge is non-physical in some sense, which is surely an anti-materialist conclusion in some sense."

Luckily, I've already written about these arguments during **my series on 100 philosophy thought experiments**. When I tackled Jackson's thought experiment in **my physicalist response to Mary's Knowledge Problem**, I wrote:

- In logical form, the argument goes something like this:

  (1) Mary has all the physical information concerning human color vision before her release.
  (2) But there is some information about human color vision that she does not have before her release.
  Therefore
  (3) Not all information is physical information.

  Hogwash! The first premise is patently false because Mary does not have "all the physical information" and cannot know "all there is to know" about this subject without having experienced it first-hand. Why? Precisely because we live in a physical universe where mental imaginings are not enough to move the physical atoms that make up the nerves in our eyes and the synapses in our brains. In philosophical terms, there is a real epistemic barrier to what we can learn no matter how much we sit in our rooms and read and think.

Later, when I tackled the thought experiment about **what it is like to be a bat**, I wrote:

- If our epistemological stance is that knowledge can only ever come after sensory experience, then of course it would be impossible to know what it is like to be a bat because we do not share the sensory experiences of a bat. Nagel may have realised this, but he ducked the question. Buried in footnote number 8 in his **original paper**, there is this:

  "My point, however, is not that we cannot know what it is like to be a bat. I am not raising that

epistemological problem. My point is rather that even to form a conception of what it is like to be a bat...one must take up the bat's point of view."

But that's exactly the problem! The epistemological problem Nagel didn't want to raise explains the entire difficulty that his mind-body thought experiment supposedly raises. … So, to me, the fact that we can't know what it feels like to be a bat is actually an argument that *bolsters* physicalism, rather than questions it.

Read the entire posts for those thought experiments to dig in more deeply, but **Gennaro** clearly agreed with me when he wrote, "Indeed, a materialist might even expect the conclusion that Nagel draws; after all, given that our brains are so different from bat brains, it almost seems natural for there to be certain aspects of bat experience that we could never fully comprehend. Only the bat actually undergoes the relevant brain processes. Similarly, Jackson's argument doesn't show that Mary's color experience is distinct from her brain processes."

6. And what about Hume's missing shade of blue?

While we're at it. There is one more famous thought experiment that may undermine physicalism and is closely related to consciousness. This is **Hume's Missing Shade of Blue**, which I also wrote about. This didn't make the standard objections list, but let's cover it quickly here before carrying on. These are the relevant snippets from my post:

- Hume argued "that all perceptions of the mind can be classed as either 'Impressions' or 'Ideas'." He further argues that: *"We shall always find that every idea which we examine is copied from a similar impression. Those who would assert, that this position is not universally true nor without exception, have only one, and at that an easy method of refuting it; by producing that idea, which, in their opinion, is not derived from this source."*
- Just two paragraphs later though, Hume seems to provide just such a destructive idea that arises without a sense impression. He says:

  "There is, however, one contradictory phenomenon, which may prove, that it is not absolutely impossible for ideas to arise, independent of their correspondent impressions. I believe it will readily be allowed that the several distinct ideas of colour, which enter by the eye...are really different from each other; though, at the same time, resembling. Now if this be true of different colours, it must be no less so of the different shades of the same colour; and each shade produces a distinct idea, independent of the rest. ... Suppose, therefore, a person to have enjoyed his sight for thirty years, and to have become perfectly acquainted with colours of all kinds, except one particular shade of blue, for instance, which it never has been his fortune to meet with. Let all the different shades of that colour, except that single one, be placed before him, descending gradually from the deepest to the lightest; it is plain, that he will perceive a blank, where that shade is wanting, and will be sensible, that there is a greater distance in that place between the contiguous colours than in any other. Now I ask, whether it be possible for him, from his own imagination, to supply this deficiency, and raise up to himself the idea of that particular shade, though it had never been conveyed to him by his senses? I believe there are few but will be of opinion that he can: And this may serve as a proof, that the simple ideas are not always, in every instance, derived from the correspondent impressions; though this instance is so singular, that it is scarcely worth our observing, and does not merit, that for it alone we should alter our general maxim."
- So, despite Hume's uncharacteristic dismissal of such a singular instance, "scarcely worth observing," this stubborn little problem seems to undermine the whole underpinnings of empiricism and physicalism. And that's a really big deal!

- [After a thorough investigation of how our eyes and vision systems work, which you should read in depth if you are interested, I said that] now that we've got a consilient view of the problem across the disciplines of biology and philosophy, we understand how we can imagine particular shades of blue even if we haven't seen them yet. Physically, it's simply a matter of how excited our blue cones have been in the past. We may not be able to "know" what peaks on those cones might look like without seeing them, but we can easily imagine points in between levels of excitement we have seen. This is simply analogous to imagining what a 5 kg weight dropped on my toe would feel like once I have had a 2 kg and 10 kg weight dropped on it. We can fill in the gaps rather easily. Similarly, I might not "know" what a 200 kg weight dropped on my toe would feel like, but I could roughly extend my imagination to it once I have some experience in the matter.

So, once again, the mind is built from physical experiences and no exceptions have been found to refute that hypothesis.

7. Is consciousness so mysterious that it is beyond our ability to understand it?

This is the third standard objection noted by **Gennaro**. In short, "**mysterians**" believe that the hard problem of consciousness can never be solved because of cognitive limitations we humans face. Colin McGinn is the leading proponent of this idea and has suggested we may be in the same situation with consciousness as a rat or dog is with respect to calculus. McGinn also notes that we "access consciousness through introspection or the first-person perspective, but our access to the brain is through the use of outer spatial senses (e.g., vision) or a more third-person perspective. Thus, we have no way to access both the brain and consciousness together, and therefore any explanatory link between them is forever beyond our reach."

Gennaro notes that materialist responses to this are numerous. Rats have no concept of calculus whatsoever, so of course they cannot solve its problems. We humans, however, know a great deal about consciousness. Gennaro even quipped, "just see the references at the end of **this entry**!" We are clearly not in an analogous position with the ignorance of rats. And while we must acknowledge there are epistemological barriers to what any one person can know about their brains or the consciousness of others, we can "combine the two perspectives within certain experimental contexts. Both first-person and third-person scientific data about the brain and consciousness can be acquired and used to solve the hard problem." Scientists do this all the time.

More generally, my analysis of the evolution of consciousness places the ability for abstraction at the highest level of its hierarchy. Once the capabilities of this level are reached—and then expanded using the tools of language, writing, and symbol manipulation, which can be arranged in an infinite number of possibilities, and stored and analysed using powerful computers—it becomes very hard to see what, if anything, could limit the conceptualisations of such a consciousness. Certainty or indisputable proof for our theories of consciousness may be out of reach, but that is the case for all of our **knowledge**. We still get by with pragmatic hypotheses that prove to be extremely robust.

To read more about how Dan Dennett finds mysterianism an embarrassment for philosophy, read **his short review of one of Colin McGinn's books in the *Times Literary Supplement***.

8. What about Zombies?

We're really grasping at straws now. The fourth and last of the standard objections listed by **Gennaro** is the problem of zombies. Supposedly, these are "creatures which are physically indistinguishable from us but lack consciousness entirely. … The appeal to the possibility of zombies is often taken as both a problem for materialism and as a more positive argument for some form of dualism, such as property dualism."

I've written about this argument in several places now. First, in **my response to a thought experiment about zombies**, I focused on the poor logic in the zombie argument. The claim that zombies *may* be possible is supposed to prove that physicalism *is* false. But that's a flawed leap. It only proves that physicalism *may* be false, and thus may also be true. I also noted there how Richard Brown's "**zoombies**" (which are conceivable beings that are identical to humans in the *non-physical* realm but have *no* consciousness, therefore implying that consciousness *must* be physical) shows that zombie arguments are circular and could just as easily be constructed *against* dualism.

In this series on consciousness, I also covered a David Chalmers interview about **the Hard Problem** where he discusses his idea of zombies. And in my post covering all **the definitions of consciousness**, I traced the history of the idea and some prominent responses. Finally, in my post on **the functions of consciousness**, I focused quite a bit on Todd Moody's "**Zombie Earth**" and Dan Dennett's paper about **the unimagined preposterousness of zombies**, which both show just how untenable the idea really is. If zombies were truly "unconscious but indistinguishable from us," then they would display fear of upcoming public speaking events or be just as engrossed in sexual fantasies or show any number of other hallmarks of internal thought processing. They would even create and speak words in their language that describe these internal states. The fact that we think unconscious creatures couldn't do these things blocks the intuition that zombies are a possibility that we need to concern ourselves with. So, let's not.

Zombie proponents **Flanagan and Polger** thought these experiments "highlight the need to explain why consciousness evolved and what function(s) it serves. This is the hardest problem in consciousness studies." I agree it's hard, but fortunately that's what the rest of this series and these FAQs have helped to uncover.

9. How is our conscious experience bound together?

Moving on from the standard objections of philosophers, **Gennaro** next notes some prominent scientific holes that need to be filled. The first one listed is known as **the binding problem** and it relates to the unity of consciousness. In a nutshell, "How does the brain 'bind together' various sensory inputs to produce a unified subjective experience?"

This is a very difficult question to answer because examinations of the brain show there isn't any one spot that could possibly act as the unifying mechanism. From an evolutionary standpoint, that's exactly what one would expect to see in nervous systems and brains that have been built up incrementally over eons of time in lots of starts and stops down various paths of trials and errors. The philosopher Jonathan Birch even has varying degrees of unity as one of the variables in his "**Dimensions of Animal Consciousness**." In the summary of that fascinating paper, there is this observation:

- "For example, neuroanatomical considerations suggest that conscious experience in mammals (which have a corpus callosum) may be more highly unified than in birds (which do not) and that experience in birds may be more highly unified than in cephalopods."

Untangling all the brain structures in the animal kingdom is taking consciousness researchers decades. And reading up on this subject quickly dives into details of brain mechanisms like v1 regions of the visual cortex, gamma-band oscillations synchronized around 40 Hz for various neuronal signals, and electromagnetic fields generated by neuronal firing. (See **here** and **here** for plenty of details like this.) It all ends up with the current state where, "There are a wide range of views on just how real this 'unity' is" and "the nature of, and solution to, [the binding problem] remains a matter of controversy."

This is all okay for me and my philosophical theory of consciousness. I'm happy to wait and see how these mechanisms are mapped out. So far, the progress being made suggests that a physical solution will be found. In fact, a promising one was just discussed in April 2021 on the Brain Science Podcast with Ginger Campbell when **she interviewed Jeff Hawkins** about his new book ***A Thousand Brains: A New Theory of Intelligence***. The whole podcast is worth listening to, but here's the transcript of the 4-minute clip that specifically addresses the binding problem.

- [**Ginger Campbell at 29:33**] And you automatically solve the binding problem?
- [**Jeff Hawkins**] Yes! I didn't know if you wanted to go there or not but that's okay. So, there's a thing called the binding problem that's poorly defined because people interpret it differently. You can think of it as the following. The brain has all these different sensors. Your eye, your retina, is not really one sensor; it's thousands of sensors aligned with each other, just like your skin has thousands and thousands of sensors along your skin. Your ear has the cochlea, and it has thousands and thousands of individual sensors in there. So, you halve all this information streaming into the brain. They all have to be processed separately. All this stuff is going on, but we have this singular perception of the world. We don't have the feeling that I'm hearing something and I'm seeing something. Your not aware of all this complicated stuff going on in your head. You just look out into the world and say "there it is. I'm looking at something and I know what it is and what it's supposed to feel like; I know what it's supposed to sound like." The question is, where does all this information get brought together in the brain? Where does it get bound together into our singular percept? If you look at the brain, you don't see that. You don't see everything going into one spot, which is like "that's you." We see connections going all over the place. There doesn't seem to be any centralised anything. How could that be? Well, our theory, which I would be remiss in not mentioning that it is called *the thousand brains theory*, reflects the fact that you have these tens of thousands of models in your neocortex. The thousand brains theory says you have all these independent models. They're each modelling a part of the world that they can see. And they don't actually come together. But what they do, and we haven't talked about this yet, is they vote. So, most of the long-range connections in the neocortex that go all over the place—from one side to the other, up and down, just all over the place, just everywhere—form connections connecting different parts of the neocortex together. We believe they're *voting*. The different columns say things like, "I'm a touch column. I'm representing my finger's input. I think I'm touching a coffee cup. But I'm not certain about it." Another column in the visual column says, "well I'm looking at an edge in the scene out here and I'm trying to model it but I'm not sure if it's a coffee cup or it could be a chair." All these columns are not certain of what they're looking at, but they have information, and they can vote! These long-range connections really try to reach a common consensus which is consistent with what they are all experiencing. This

makes it so that all of a sudden everyone goes, "Yep! We're all agreeing that this thing is a coffee cup, or a computer, or a bird." So, the binding doesn't occur in one spot. It's essentially a voting mechanism that occurs across the brain and our perceptions are primarily of that voting. We're not aware of all the thousands of models that are guessing what is going on in the world. But we are aware of their consensus. And the consensus says, "yes, we all agree that this is something" and I can then drill down and say, well what does that look like, what does that sound like, what does that feel like. But we all agree that it's this bird or whatever. And so, this solves the binding problem by not binding it into one spot but by voting and reaching consensus. And so therefore we don't have to look for a spot in the brain where everything comes together.

- [**GC**] This also makes sense of the fact that most of what the cortex does is not conscious.
- [**JH**] Yes! We're almost totally unaware of most of what is going on in there. All the tiny inputs are constantly changing, but the consensus voting stays the same and that allows for continual experience. [**Clip ends at 33:50**]

This sounds very promising as it's easy to see how it would be built up gradually over time, bringing more and more representational voting into the overall picture. But for now, let's wait for the scientific method to play out before declaring any firm answers to this question.

10. What can the neural correlates of consciousness tell us?

The other major hole in our scientific understanding of consciousness that **Gennaro** discusses is the program to find the **neural correlates of consciousness** (NCCs). This project is based on the idea that consciousness originates in the brain, and "some credit for it must go to the ground-breaking 1986 book by Patricia Churchland entitled *Neurophilosophy*." In the paper "**What is a Neural Correlate of Consciousness?**", David Chalmers answers that title question thusly: "At first glance, the answer might seem to be so obvious that the question is hardly worth asking. An NCC is just a neural state that directly correlates with a conscious state." He goes on to elaborate, however, that, "A number of proposals have been put forward concerning the nature and location of neural correlates of consciousness. A few of these include:

- 40-hertz oscillations in the cerebral cortex (Crick and Koch 1990)
- Intralaminar nuclei in the thalamus (Bogen 1995)
- Re-entrant loops in thalamocortical systems (Edelman 1989)
- 40-hertz rhythmic activity in thalamocortical systems (Llinas et al 1994)
- Extended reticular-thalamic activation system (Newman and Baars 1993)
- Neural assemblies bound by NMDA (Flohr 1995)
- Certain neurochemical levels of activation (Hobson 1997)
- Certain neurons in inferior temporal cortex (Sheinberg and Logothetis 1997)
- Neurons in extrastriate visual cortex projecting to prefrontal areas (Crick and Koch 1995)
- Visual processing within the ventral stream (Milner and Goodale 1995)

(A longer list can be found in Chalmers 1998.)"

Looking at this list, you can readily understand why Gennaro said, "a detailed survey would be impossible to give here" and I would not attempt such a thing either. I'm happy to let the neuroscience play out for years to come as it maps what I think of as **the mechanisms of consciousness**, which is just one of **Tinbergen's four questions** about any biological

phenomenon. In the meantime, Chalmers' dense list of paths for this exploration serves to highlight the two main meta-problems with this project that Gennaro notes.

First:

- "One problem with some of the above candidates is determining exactly how they are related to consciousness. For example, although a case can be made that some of them are necessary for conscious mentality, it is unclear that they are sufficient. That is, some of the above seem to occur unconsciously as well. And pinning down a narrow enough necessary condition is not as easy as it might seem."

I think this problem of searching for a narrow condition comes from having too narrow a definition of consciousness. Researchers seem to be focused merely on conscious awareness, which comes in at level 5 in my hierarchy, and only arrived in biological life after the other levels below it were established. Such emergence never comes from a clear-cut break in evolution, so pinning down exact NCCs for that second C of "consciousness" may be a fool's errand. As detailed above in question 1, a *functional analysis* will be required which "consists in breaking down some capacity or disposition of interest into simpler dispositions or capacities, organized in a particular way." There just won't be one simple answer.

Second:

- "Another general worry is with the very use of the term 'correlate.' … Even if such a correlation can be established, we cannot automatically conclude that there is an identity relation. Perhaps A causes B or B causes A, and that's why we find the correlation. Even most dualists can accept such interpretations. Maybe there is some other neural process C which causes both A and B. 'Correlation' is not even the same as 'cause,' let alone enough to establish 'identity.'"

This is the same problem that Patricia Churchland answered for us above in question 4. I'll just repeat the relevant part of the quote here from her paper "**Neurophilosophy**":

- Here is the way to think about this: identities—such as that temperature really is mean molecular kinetic energy, for example—are not directly observable. They are underwritten by inferences that best account for the mass of data and the appreciation that no explanatory competitor is as successful. One could, if determined, dig one's heels in and say, "temperature is not mean molecular kinetic energy, but rather an occult phenomenon that merely runs parallel to KE." It is a logically consistent position, even if it is not a reasonable position.

So, the results of the NCC project will have their limits, but since they are not ruling out physicalism, that hypothesis continues to hold up with all of the evidence in the universe that has ever been gathered and tested.

11. Are other animals conscious?

Gennaro starts with the obvious (to me) concession that "in the aftermath of the Darwinian revolution, it would seem that materialism is on even stronger ground provided that one accepts basic evolutionary theory and the notion that most animals are conscious." But then he notes there is still much discussion around the question, "To what extent are animal minds

different from human minds?" Well, according to my definition, *all* living beings do indeed have *some* levels of consciousness, and I can use my comprehensive hierarchy as a guide to describe how much and of what kinds. These descriptions of a being's consciousness vary widely across all species, across individuals within a single species, and across the lifespan of individuals too.

An important outcome from this is to not think of consciousness as a single variable or an on-off switch. The philosopher Jonathan Birch has published an excellent example of this in his 2020 paper about the "**Dimensions of Animal Consciousness**" where he uses a **radar chart** (aka spider web chart) to illustrate what five dimensions might look for elephants, corvids, and cephalopods. Birch, however, recognises that this is just a starting example to get people thinking in the right way. Among the key challenges he discusses for mapping dimensions of consciousness, he says that "One is to find dimensions at the right grain of analysis. If our goal were to capture all interesting variation in conscious states, we would never have enough dimensions. We have to be pragmatic." I agree, although I probably would have started with the 13 types of cognition listed in **Pamela Lyon's paper** on the evolution of cognition (which I placed in my hierarchy when I mapped **the functions of consciousness**). That's a bit more difficult to plot, though, and Birch isn't trying to be comprehensive. I, however, do want my hierarchy to be comprehensive, so let's see how Birch's dimensions might be covered within my hierarchy.

1. *E-richness* (where the e stands for evaluative) is roughly equivalent to the cognition of *valance* within my level of affect, but it also looks at *motivation* according to Birch's chart of experiments for each of his dimensions.
2. *P-richness* (where the p stands for perceptual) is equivalent to the cognition of *sense perception* that sits within my level of affect. I see p-richness and e-richness going hand in hand because one must perceive something in order to evaluate it, and living beings evaluate everything they perceive (as positive, negative, or neutral). This is why I have them on the same level in my hierarchy. Birch is right, though, that they can change in independent directions from one another.
3. *Unity* or *integration at a time* relates to the binding problem noted above in question 9. This is an interesting dimension which Birch explores with examples such as humans with split-brain syndromes, dolphins and seals sleeping with one hemisphere at a time, and the fact that birds have no structure akin to a corpus callosum. He wonders, "Could there be two subjects within one skull?" This will come up again when I discuss the nesting problem below in question 17, but for now, I see the unity dimension as a way of looking at how a few of the cognitions in my intention level actually combine together. Just how intentional can one animal (or one consciousness!) act using the *attention*, *memory*, *pattern recognition*, and *learning* that it has at its disposal. Each of sub-categories can obviously vary from one another, so I consider Birch's *unity* as a meta-variable examining how these are combined.
4. *Temporality* or *integration across time* is another complex meta-variable to me. This one looks at just how short or long of a timespan can be considered to affect the conscious experiences and thoughts of animals. This integrates across several cognitions in my hierarchy — *sense perception* and *discrimination* in my level of affect, *memory* and *pattern recognition* in my level of intention, *anticipation* and *error detection* in my level of prediction, the *self-reference* in my level of awareness that creates the autobiographical self, and even the ability in my final level for *abstraction* to use symbols and language to help extend thinking into the distant past or future. So, I could probably make another interesting radar chart for this single variable in Birch's dimensions.

5. *Selfhood*, according to Birch, is "the conscious awareness of oneself as distinct from the world outside." This is equivalent to the cognitive ability for *self-reference* in my level of awareness.

So, all of Birch's dimensions can indeed be mapped onto my hierarchy, but is there anything of mine that he's left out? The only cognitive abilities I have listed which I don't think he covers are the ones for *communication* and *problem solving*. These both seem to be interesting abilities that can vary widely across different individuals and different species, so perhaps they too could make for useful considerations during future analyses of animal consciousness.

12. Can machines be conscious?

The short answer is yes. According to my definition, machine consciousness is possible, although it would certainly feel different than ours. (Think how much our own consciousness changes under the influence of a few chemicals and just imagine what an entirely different substrate might cause.) In order to describe any machine's consciousness accurately, we would need the same kind of comprehensive functional analysis as described above, which would map all of the dimensions throughout my hierarchy.

To explore this in more detail, let's consider three more questions that **Gennaro** raised in his IEP article on consciousness.

1. Could an appropriately programmed machine be conscious? Yes. In a material universe without souls imbued by gods, it's hard to see why not. My theory of *pandynamism* acknowledges that all matter feels forces, but minds arise when subjects emerge. An appropriately programmed machine could conceivably recreate the conditions for a living subject, which would then feel its physical changes.

2. Could a robot really subjectively experience the smelling of a rose or the feeling of pain? Once again, yes, but only if the above conditions are satisfied. You have to have a subject before what we call subjectivity can enter into it. Cameron Harwick's long article on "**What Computer-Generated Language Tells Us About Our Own Ideological Thinking**" makes an important point about this. Harwick states:

- Thus, the ancient question of what separates humans from animals is the inverse of the more recent question of what separates humans from computers. With **GPT**, computers have finally worked backward (as seen in animal terms), from explicit symbol manipulation to a practically fluent generative language faculty. The result might be thought of as a human shell, missing its animal core.

This is exactly right. And that "animal core" is my hierarchy level of affect, which is what Mark Solms **calls** the hidden spring or source of consciousness. Without this innate, evolved, built-in sense of judging what is good or bad or indifferent for a self, there is no way that the sting of pain or the sweet smell of a rose can make sense. Could that be programmed into a robot or machine? Yes. But with some interesting differences worth discussing in the next question.

3. How and when does one distinguish mere 'simulation' of some mental activity from genuine 'duplication'? This question is in reference to John Searle's famous Chinese Room. I have summarised **my response to this thought experiment** by saying "emotions, definitions for good and bad,

and the ability to learn to meet a hierarchy of needs are probably enough to create strong artificial intelligence. They are all we have ourselves." So, creating artificial subjects by simulating our own interactions with the world seems entirely possible, although that wouldn't *duplicate* our conscious experience of these interactions. Does that matter? Do the forces felt from the movement of my sodium-channel ions matter any more or less than the forces felt from the movement of a different set of chemicals? I'm not a bio-chauvinist so I don't see why that makes a difference morally, even if there is a difference in the raw feelings. So, duplication isn't the goal to me. Searle's Chinese Room is meant to pump the intuition that simulation of a function isn't enough to matter because "clearly" the man in the Chinese Room (or the Chinese Room as a whole system) isn't having the same experience as an individual human speaking Chinese. But that's because of all of the other activities that are also wired into our own speaking systems. If you could somehow remove all of the memories from a person, and all of the living, emotional, and other sense systems as well, but miraculously keep the auditory and speaking systems going all by themselves, would there be any "consciousness" there? Not in the way that Searle meant. Such a listening and speaking slice of a human would be just as dumb as a Chinese Room.

Here's another way to approach this issue. Let's say you programmed a computer to speak "ouch" when its vibration sensors moved too vigorously. That is *simulating* pain, but we don't think it is duplicating *our* pain. In a really sophisticated robot, would that pain matter? I don't see why not, if such a robot were programmed to be aware of its surroundings and able to learn from them, while also striving towards open-ended goals, and simply becoming irreplaceable because of its unique prior experiences and potential for even more. And yet, the exact same physical inputs in such a computer could be easily tinkered with and re-programmed to say "yum" or "blue hippopotamus" when it was shaken, which would render its conscious simulation utterly nonsensical. There just isn't the kind of singular match between changes in the world and felt states inside the computer that would persuade us to consider it conscious in the same way that we are. Such a computer could conceivably be constructed with a kind of consciousness that we care about, but it would be extremely fragile and fluid compared to our own. It would be subject to the whims of its programmers. Perhaps, however, we may one day learn the chemical coding that drives our own bodies to the point that we are as fluent in that as we are now in computer coding. Were such editing of our own biological codes to become so possible, our own consciousness could become just as fluid and changeable as the computer's. Would that erase our own consciousness? I think not. It would just change what else we need to include in order to describe it.

13. So, "what is it like" to be conscious?

Question 12 was the end of the standard objections and scientific holes in the IEP entry from **Gennaro** about consciousness, but there were three more things he touched on in his brief introduction that I thought were worth a quick discussion. This first one is of course a reference to Thomas Nagel's famous "what is it like" description of consciousness, which Gennaro called "perhaps the most commonly used contemporary notion of a conscious mental state." In Gennaro's retelling of this, "When I am in a conscious mental state, there is something it is like for me to be in that state from the subjective or first-person point of view. But how are we to understand this?"

One problem with this is that it is too narrow a search to provide much understanding. As I mentioned above in question 10 about the search for the NCCs,

- "I think this problem of searching for a narrow condition comes from having too narrow a definition of consciousness. Researchers seem to be focused merely on conscious awareness, which comes at level 5 in my hierarchy, and only arrived in biological life after the other levels below it were established."

This "what it is like" feeling that Nagel is describing disappears whenever we are rendered unconscious, and yet much of consciousness' processing still goes on to keep us alive (as it does when we are awake as well). In order to fully understand "what it is like", we have to look at the long history of emergence that got to that kind of on-again / off-again state. In my post on **the evolutionary history (aka phylogeny) of consciousness**, we can see the possibility that *awareness* of "what it is like" may go back a very long way. It is best tested using mirror recognition tests, which several non-human species have passed including mammals, birds, and fish, who shared a common ancestor 525 million years ago. And since cephalopods appear to have independently evolved awareness as well, it could be spread even farther and wider in the animal kingdom.

These estimates are, of course, 3rd-person conjectures using the best tools science has for studying consciousness. We cannot experience "what it is like" to be in another subject, so I should also repeat here briefly what I mentioned above in question 5 about **my post** reacting to Nagel's thought experiment.

- The epistemological problem Nagel didn't want to raise explains the entire difficulty that his mind-body thought experiment supposedly raises. … So to me, the fact that we can't know what it feels like to be a bat is actually an argument that *bolsters* physicalism, rather than questions it.

14. Do we have immortal souls?

The second extra issue raised by **Gennaro** with "the problem of consciousness is…related to major traditional topics in metaphysics, such as the possibility of immortality." The possibility of immaterial souls that go on forever has no evidence behind it and lots of other evidence to the contrary. Physicalists reject this idea, although I believe that ending the aging process in our human bodies in order to live **indefinitely long lives** is definitely an idea worth thinking and writing (**a novel**) about.

15. Do we have free will?

For the last of these issues raised by **Gennaro**, there is the point that "the problem of consciousness is…related to major traditional topics in metaphysics, such as…the belief in free will." Quite luckily, while I was writing this series on consciousness, I was asked if I wanted to review Gregg Caruso and Dan Dennett's book on this subject (***Just Deserts***), which spurred me to dive deeply into the free will debate. After 6 posts exploring other people's positions, I wrote **a summary of my own thoughts**. In a nutshell, I say we don't have the freest will imaginable, but we do have significant degrees of freedom, and that provides a kind of "free will worth wanting." Adding another functional analysis here using Tinbergen's four questions sheds a lot of light on the emergence and expansion of these freedoms, which are completely aligned with the emergence and expansion of consciousness. This linkage makes sense since **Dan Dennett** noted:

- "It is no mere coincidence that the philosophical problems of consciousness and free will are, together, the most intensely debated and (to some thinkers) ineluctably mysterious phenomena of all. As the author of five books on consciousness, two books on free will, and dozens of articles on both, I can attest to the generalization that you cannot explain consciousness without tackling free will, and vice versa."

Agreed. And tackled now.

QUESTIONS FROM OTHER NATURALISTS

In addition to the standard questions listed in the online encyclopedia articles that I cited above, I have found a few other questions worth discussing that have been raised by other naturalist philosophers. Let's go through those here.

16. Can't we just get by with a very rough definition of consciousness?

The philosopher Eric Schwitzgebel maintains an excellent blog called **The Splintered Mind**, which often touches on topics in the field of consciousness. In 2016, Schwitzgebel published a paper called "**Phenomenal Consciousness, Defined and Defended as Innocently as I Can Manage**" in which he argued that the best approach for defining consciousness right now may be a "definition by example" which can work well "if one provides diverse positive and negative examples and if the target concept is natural enough that the target audience can be trusted to latch onto that concept once sufficient positive and negative examples are provided."

Let's see how this works in practice. Here are some of the positive examples Schwitzgebel lists: sensory and somatic experiences; conscious imagery; emotional experience; thinking and desiring; and dream experiences. Does that help yet? Here's a passage about the negative examples to keep at it.

- "Not everything going on inside of your body is part of your phenomenal consciousness. You do not, presumably, have phenomenally conscious experience of the growth of your fingernails, or of the absorption of lipids in your intestines, or of the release of growth hormones in your brain. Nor is everything that we normally classify as mental part of phenomenal consciousness. Before reading this sentence, you probably had no phenomenal consciousness of your disposition to answer 'twenty-four' when asked 'six times four'. … If a visual display is presented for several milliseconds and then quickly masked, you do not have visual experience of that display (even if it later influences your behavior). … [And] we normally think that dreamless sleep involves a complete absence of phenomenal consciousness."

Now that these have been introduced, Schwitzgebel concludes, "I suggest that there is one folk psychologically obvious concept, perhaps blurry-edged, that fits the positive and negative examples while leaving the contentious examples open and permitting wonder of the intended sort. That's the concept of phenomenal consciousness."

Is that very helpful, useful, or interesting? I don't really think so. Can we say more? Sure, but Schwitzgebel doesn't want us to go too far. He says, "At this point, it is tempting to clarify by making some epistemic or metaphysical commitments—whatever commitments seem

plausible to you. You might say, 'those events with which we are most directly and infallibly acquainted' or 'the kinds of properties that can't be reduced to physical or functional role'. Please don't! Or at least, don't build these commitments into the definition. Such commitments risk introducing doubt or confusion in people who aren't sure they accept such commitments."

Okay, now we're just backing away from any of the hard work of understanding consciousness, and it's obvious that Schwitzgebel is only concerned with the very narrow conception of *conscious awareness*, which is level 5 in my hierarchy. Worse still, he's looking for the least common denominator that everyone can agree to. I'm afraid that will end up with as tiny and useless a definition as possible when more and more opinions are brought into the discussion. In fact, Schwitzgebel acknowledges, "My definition did commit me to a fairly strong claim about folk psychology: that there is a single obvious folk-psychological concept or category that matches the positive and negative examples." But this is exactly the type of essential on/off switch that Dan Dennett warned about in his paper "**Darwin and the Overdue Demise of Essentialism**."

Sorry, but I don't think an overly simple and deliberately narrow definition will do. Far better to work on the comprehensive functional analysis that helps put everyone's various opinions in their own place and shows the relationships they all have to one another. Such an analysis helps us see the building blocks of consciousness and how they all emerge over evolutionary timescales. Schwitzgebel's definition does none of that work. I've set myself a lofty goal for my consciousness studies, but I do think its attainable.

Fortunately for us, Schwitzgebel's "innocent" and narrow definition doesn't actually stop him from exploring wider issues with consciousness, which I'll cover in questions 17 and 18.

17. What about the various parts of living systems? Which ones are conscious?

In a November 2020 post on his blog, Schwitzgebel laid out **the nesting problem for theories of consciousness**. In this question I'll look at the nesting problem going down, and in the next one I'll consider it going up. First, though, what are we talking about exactly?

Schwitzgebel starts with the background that "in 2016, **Tomer Fekete, Cees Van Leeuwen, and Shimon Edelman** articulated a general problem for computational theories of consciousness, which they called the Boundary Problem. The problem extends to most mainstream functional or biological theories of consciousness, and I will call it the Nesting Problem." Then, he gives this as a quick explanation:

- "Consider your favorite functional, biological, informational, or computational criterion of consciousness, criterion C. When a system has C, that system is, according to the theory, conscious. … Unless you possess a fairly unusual and specific theory, probably the following will be true: Not only the whole animal (alternatively, the whole brain) will meet criterion C. So also will some subparts of the animal and some larger systems to which the animal belongs."

This, then, yields some questions for Schwitzgebel:

- First: *Are* all these subsystems and groups conscious?

- Second: If we want to attribute consciousness only to the animal (alternatively, the whole brain) and not to its subsystems or to groups, on what grounds do we justify denying consciousness to subsystems or groups?
- Or maybe instead of a threshold, it's a comparative matter: Whenever systems nest, whichever has the most connectivity is the conscious system. … Or maybe it's not really C (connectivity, in this example) alone but C plus such-and-such other features, which groups and subsystems lack. … Or maybe groups and subsystems are also conscious — consciousness happens simultaneously at many levels of organization.

Schwitzgebel doesn't think these questions are unanswerable, just that, "this is a fundamental question about consciousness which is open to a variety of very different views, each of which brings challenges and puzzles—challenges and puzzles which philosophers and scientists of consciousness, with a few exceptions, have not yet seriously explored."

This discussion really shows the beauty of having a comprehensive hierarchy of consciousness rather than a singular, restrictive, narrow definition. This issue started off as *the boundary problem*, but since we are dealing with a biologically emergent property, there are no clear boundaries here! It's obvious that any singular criterion C will have trouble moving up and down the story of consciousness. For me, that's not a problem.

Are all these subsystems conscious? No, your kidneys or autonomic nervous system have not reached conscious awareness in my hierarchy, but they do have the properties of consciousness that are included in my levels of affect and intention. According to Jeff Hawkins' **thousand brains theory**, they may also have local abilities for prediction too. And these subsystems contribute pieces of consciousness to other systems that may reach higher levels in my hierarchy. The point is that all of these elements can be analysed and understood for their contribution back and forth to the various levels of consciousness in living systems.

On what grounds do we justify denying consciousness to subsystems or groups? We don't deny them *all* of the levels of consciousness. We can just be clear about which ones they have and which ones they contribute to other systems that may or may not reach different levels of consciousness.

Maybe consciousness happens simultaneously at many levels of organization. That's right, as long as your definition of consciousness is as wide and flexible as mine is, yet capable of offering enough precision to describe the various varieties of consciousness that are on offer as well.

18. Is the United States conscious?

This question essentially extends the nesting problem in the upwards direction, although it is based on a paper from Schwitzgebel that is six years older than his post on the nesting problem. That paper is called, "**If Materialism Is True, the United States Is Probably Conscious.**" That title sounds ridiculous on the face of it, but let's give Schwitzgebel some benefit of the doubt and explore his claims in a bit of detail rather than just dismiss them. The explorations prove fairly illustrative for the benefit of taking an evolutionary approach here.

Schwitzgebel starts off by introducing us to two sci-fi scenarios that are meant to disabuse us of a prejudice he calls *contiguism*, which apparently stops us from believing in spatially distributed consciousnesses. The first are *Sirian Supersquids*. Here is their story:

- They can detach their limbs. To be detachable, a supersquid limb must be able to maintain homeostasis briefly on its own and suitable light-signal transceivers must appear on the surface of the limb and on the bodily surface to which the limb is normally attached. … [Also], the limb-surface transceivers developed the ability to communicate directly among themselves without needing to pass signals through the central head. … Despite their spatial discontinuity, they aren't mere collections. They are integrated systems that can be treated as beings of the sort that might house consciousness.

I would agree that these creatures could be spatially distributed, yet consciously integrated, but only because the information from the various parts is being integrated in one place. Any signals not sent to the head would be analogous to the unconscious processing that goes on in our own bodies. Anyway, let's carry on. The second sci-fi creation are the *Antarean Antheads*. Here are the relevant bits of their story:

- [These are] a species of animals who look like woolly mammoths but who act much like human beings. … Here's why I call them "antheads": Their heads and humps contain not neurons but rather ten million squirming insects, each a fraction of a millimeter across. Each insect has a complete set of minute sensory organs and a nervous system of its own, and the antheads' behavior arises from complex patterns of interaction among these individually dumb insects. … Maybe there are little spatial gaps between the ants. Does it matter? Maybe, in the privacy of their homes, the ants sometimes disperse from the body, exiting and entering through the mouth. Does it matter? … You might think that the individual ants would or could be individually conscious and that it's impossible for one conscious organism to be constituted by other conscious organisms. Some theoreticians of consciousness have said such things—though I've never seen a good justification of this view.

I believe the answers to these questions come from careful considerations of evolutionary biology. It's not so much that "it's impossible for one conscious organism to be constituted by other conscious organisms." That depends very much on your definition of consciousness and my answer to the previous question shows how subsystems with lower forms of consciousness integrate into higher systems that achieve higher levels of consciousness based on the extra information that is available to them. That is possible and completely consistent with the functional analysis enabled by my hierarchical theory. However, based on the evolutionary biology that has been observed in our world, it seems impossible for creatures like the Antarean Antheads to ever emerge.

I draw this conclusion from ***The Origins of Life*** by John Maynard Smith and Eors Szathmary, which covers **the major transitions in evolution**. As I summarised in **a talk** I gave, "the big takeaway from this book is that each transition occurred when formerly separate and competitive biological elements figured out new ways to join up and cooperate with one another, and begin to evolve together." That sounds vaguely like what Schwitzgebel's antheads have done, but it runs afoul of this quote from p.19 of the book:

- "One feature [of major transitions] crops up repeatedly. Entities that were capable of independent replication before the transition could afterwards replicate only as part of a larger whole."

That reproductive integration is crucial! It's what actually enables natural selection to slowly work its magic on the shaping of these emergent new species. Evolution is said to require

three steps: variation, selection, and retention. But there's no way for the antheads to manage this as a coherrent species with such independent creatures like the ants in their heads. So, they make for a poor example whose seeming impossibility is unable to dissuade me from my so-called *contiguism*. Nevertheless, let's carry on with Schwitzgebel's paper as he combines some features from these two sci-fi creatures in order to investigate yet another one called the *Sirian Squidbit*, which brings up a few more issues.

- "The Sirian squidbits [are] a species with cognitive processing distributed among detachable limbs. … Let me tie Sirius, Antares, and Earth a bit more tightly together. As the squidbit continues to evolve, its central body becomes smaller and smaller—thus easier to hide—and the limbs develop more independent homeostatic and nutritional capacities, until the primary function of the central body is just reproduction of these increasingly independent limbs. Earthly entomologists come to refer to these central bodies as 'queens'. Still later, squidbits enter into symbiotic relationship with brainless but mobile hives, and the thousand bits learn to hide within for safety. These mobile hives look something like woolly mammoths. Where is the sharp, principled line between group and individual?"

Schwitzgebel is clearly referencing the eusocial species of ants here and trying to use the fact that they are considered superorganisms to make it seem plausible that there can be something like *superconsciousness*. But once again the issue is resolved by the separability of reproductive biology from the biology of consciousness. Eusocial ants are considered superorganisms because they cannot reproduce as individuals. That is why they are only selected for at the group level. But that says nothing about the consciousness of such a group of individuals. As discussed above in question 9 about the binding problem, and in question 10 about the neural correlates of consciousness, there are several candidates for physical structures and processes that integrate elements of consciousness together. There are no features like these in ant colonies which could bind the consciousness of the individuals together even though they must reproduce as a group and are therefore selected and shaped on the basis of their collective actions. Unless consciousness is immaterial, there is no reason to believe in the consciousness of an ant colony. In fact, since there is no "spooky action at a distance" from one ant individual to another, there is no evidence for an immaterial consciousness there that is sensing and reacting to the needs of the group as a whole. Note that this is the case even though ants have been part of fiercely competitive superorganisms for millions of years! If superconsciousness were going to arise anywhere, surely it would be there. Anyway, that is how the lines between groups and individuals can be understood in ants and Sirian Squidbits.

Okay, but what about the United States? By now you must see why this is also problematic, but Schwitzgebel raises a number of other questions here (in a kind of **Gish Gallop**??), so let me tackle them as quickly as I can in rapid fire succession.

- You might say: The United States is not a biological organism. It doesn't have a life cycle. It doesn't reproduce. It's not biologically integrated and homeostatic. Therefore, it's just not the right *type of thing* to be conscious.

It's not about the *type of thing*. I'm not a bio-chauvinist. It's about the fact that the United States doesn't have any mechanisms, phylogeny, or ontogeny—three of the four Tinbergen questions—which could contribute to any sense of a U.S. consciousness.

- Why should consciousness require being an organism in the biological sense? Properly-designed androids, brains in vats, gods—these things might not be organisms in the biological sense and yet are sometimes thought to have consciousness.

This point is fine. As I described above in question 12, consciousness may not require biology.

- Second, it's not clear that nations aren't biological organisms. … other types of coordination emerge spontaneously from the bottom up, just as in ordinary animals.

If you actually look at the detailed definitions of life and organisms, it's quite clear that the United States doesn't qualify as an organism. I trust I don't need to explain this further.

- Nations also reproduce—not sexually but by fission.

This badly confuses culture with biology. Nations are merely an abstract notion. They don't reproduce in any way comparable to organisms.

- According to a broad class of plausible materialist views, any system with sophisticated enough information processing and environmental responsiveness, and perhaps the right kind of historical and environmental embedding, should have conscious experience. My central claim is: The United States seems to have what it takes, if standard materialist criteria are straightforwardly applied without post-hoc noodling. It is mainly unjustified morphological prejudice that blinds us to this.

Eeks. This sounds like a blatant **category error**. Our "morphological prejudice" remains well justified, and my materialist criteria require no "post-hoc noodling" to deny the consciousness of the United States. I'll discuss the problems with linking consciousness to information processing alone in question 40 below.

- Consider, first, the sheer quantity of information transfer among members of the United States. … Our information exchange is not in the form of a simply-structured massive internet download. The United States is a goal-directed entity, flexibly self-protecting and self-preserving. The United States responds, intelligently or semi-intelligently, to opportunities and threats. … I am asking you to think of the United States as a planet-sized alien might, that is, to evaluate the behaviors and capacities of the United States as a concrete, spatially distributed entity with people as some or all of its parts, an entity within which individual people play roles somewhat analogous to the role that individual cells play in your body.

Yes, indeed, this is the mother of all category errors. The United States is not a "concrete, spatially distributed entity." It's just an abstract idea. We can't draw an abstract line around every imaginable group and declare it to have its own consciousness. We don't think there is a consciousness of "left-handed NBA fans" no matter how similar that group is to a nation.

Schwitzgebel asked us to consider the consciousness of the United States, but of course the same question is often asked of other super-entities such as ecosystems or the whole earth of Gaia. Well, all of the same arguments in this question apply to those situations as well and deny any likelihood of superconsciousness there either. Let me just add this extra quote from **John Maynard Smith and Eors Szathmary** as a final piece of evidence:

- "Consider a present-day ecosystem—for example, a forest or a lake. The individual organisms of each species are replicators; each reproduces its kind. There are interactions between individuals, both within and between species, affecting their chances of survival and reproduction. There is a massive amount of information in the system, but it is information specific to individuals. There is no additional information concerned with regulating the system as a whole. It is therefore misleading to think of an ecosystem as a super-organism."

19. How do we know we don't have "inverted qualia"?

This is a quick little issue that was mentioned in *The Guardian* in **a long review** of Mark Solms' recent book about consciousness. The author noted that, "the 'problem of inverted qualia' refers to the fact that the experience you call 'seeing green' could be identical to the one I call 'seeing red', and vice versa, and we'd never have any way of knowing."

Based on my response to question 13 about "what it is like" to be conscious, we physicalists admit that we can't actually know what others are experiencing. That barrier is completely consistent with physicalism, and in fact it is a consequence of the universe being confined to the physical. (If consciousness arose from immaterial mental properties, you'd think we would already have found a way to inhabit other physical bodies and therefore know what it was like in them.) However, the shared evolutionary history of all life, and the shared physical building blocks we are all made of precludes any reason to think any of us actually have inverted qualia. After all, the most common cause of **color blindness** is "an inherited problem in the development of one or more of the three sets of the eyes' cone cells, which sense color." Once again, changes in the subjective experience of consciousness are matched by physical changes in the body experiencing that consciousness. My reliance on evolution here leads us to the next question.

20. How do you solve the mind-evolution problem?

Based on its title, the neurobiologist Yoram Gutfreund wrote a really challenging paper for me called "**The Mind-Evolution Problem: The Difficulty of Fitting Consciousness in an Evolutionary Framework**." In that paper, Gutfreund described how,

- "Consciousness is one of the last biological phenomena about which we do not have a solid idea as to how and when it appeared and evolved in evolution. … The question of how the mind emerged in evolution (the mind-evolution problem) is tightly linked with the question of how the mind emerges from the brain (the mind-body problem). It seems that the evolution of consciousness cannot be resolved without first solving the 'hard problem' (Chalmers, 1995). Until then, I argue that strong claims about the evolution of consciousness based on the evolution of cognition are premature and unfalsifiable."

I agree with Gutfreund that this mind-evolution problem is tightly linked with the mind-body problem and the hard problem. In question 4 about how minds could have emerged from matter, I explained how my theory of pandynamism fits the evidence in the world where consciousness appears to emerge and grow along with the emergence of living subjects. This answers the hard problem by naming "felt forces" as an underlying identity in the universe. These felt forces, then, grow and change in subjective consciousnesses as the subjects grow and change their structures for sensing these forces. Changes in the world that affect my five

senses will change my conscious experience. If I lose a sense (e.g. if I go blind), then changes in the light around me no longer affect my consciousness.

Gutfreund had identified four possibilities for any attempts to fit consciousness into an evolutionary framework. Presumably, one of them will work for me if my theory is worth considering. Gutfreund's four possibilities are:

1. *Consciousness as a tool for behavior.* Is consciousness to an animal like wings are to a bird, i.e., a tool to enable an advantageous goal? If consciousness is a tool, what is the goal that it enables? Some answers include: to create a unified and coherent representation of all incoming information (Crick and Koch, 1998; Merker, 2005); to enable the learning of sensory and cognitive representations (Grossberg, 1999); to make complex flexible decisions (Earl, 2014); and more. … Difficulty with this notion is that cognitive behaviors are caused by the brain's neural circuits, without the necessity to introduce conscious states to the models.
2. *Consciousness as brain identity.* One escape route around this paradox is to suggest an identity between consciousness and neuronal states (Loorits, 2014; Smart, 2017), that is, some neuronal states are conscious feelings; the two are the same, described at different levels. The biological function of the neural state then becomes the function of the feeling (Searle, 2013). A problem with such an identity approach is that evolution operates at the level of the body and not at the level of the feelings. The only things that matter from an evolutionary point of view are the animal's actions, and the neural processes that choose and elicit the actions. … Therefore, the implication of an identity hypothesis is that consciousness becomes detached from any evolutionary theory.
3. *Consciousness as an advantageous goal.* What if consciousness is a goal in itself? In this case, neurons organized in specific ways in specific brain structures are the wings to support consciousness, and the property of being conscious improves the fitness of the animal in which it is installed, just like the properties of flying, swimming, or chewing. But, in what ways do feelings and emotions improve fitness? An antelope escaping from a lion needs to run quickly and efficiently. Why, from an evolutionary point of view, does it also need to feel the terrible feeling of fear? This is a puzzle and evolutionary theory has no answers.
4. *Consciousness as a by-product.* A different approach that bypasses the difficulties described above is to view consciousness as a byproduct of brain activity. In this case, consciousness doesn't affect behavior and has no function of its own. However, it has an adaptive value that stems from its association with a behavioral phenomenon, which in turn does have a function. … The pitfall of such an approach is that consciousness can be removed from the model without any influence on the flow of the model.

Once again, Gutfreund appears to only be considering "consciousness" as some narrow part of conscious awareness, and this makes it quite difficult to trace the evolutionary path and usefulness of that small piece. By tracing the history of the evolution of all forms of cognition, and embracing all of those associated functions and behaviours as different aspects of consciousness, I think it becomes easier to see the slow emergence of *consciousness as an identity with living systems* (i.e. #2, but not just for brains), which impacts all behaviour and therefore *acts as a tool* (i.e. #1 but with a much broader reach of enabling and improving survival across many different routes). Consciousness is not *a goal in itself* (#3) or *an epiphenomenal by-product* (#4) since it is just an unavoidable part of life, which is unavoidably shaped by evolution and natural selection.

All of this is best traced in my post on **the functions of consciousness** where my hierarchy was first developed in full. Here are a couple of quick highlights:

- As soon as the origin of life takes hold in the first level of my hierarchy, the next tier of affect begins to get embedded as living entities *feel* their way through life and quickly develop associations between good and bad feelings as they relate to life and death. These are innately passed down through successful generations.
- Over time, adaptations from affective reflexes alone lead to capacities for cognition that are able to interrupt these reflexes. The capacities of attention, memory, pattern recognition, learning, and communication create a core self where organisms can be said to be acting with intention, which is the third level of my hierarchy.
- Once intentions exist, they can be taken into account. To do so is to use prediction (my fourth level) to think through what the result will be from any intentions. This requires the cognitive capacities of anticipation, problem solving, and error detection.
- As predictions and perceptions improve, organisms eventually make the connection that there is a self which has its own mind. The fifth level of awareness is achieved, along with the arrival of the cognitive capacity for self-reference. Such conscious cognition allows memories and thoughts built from the lived past and the anticipated future to create the autobiographical self. Note that this is often the level that neuroscientists concern themselves with and only a few extra abilities seem to emerge here such as "trace conditioning" and the recreation in thought of past events in order to learn from them anew in light of new information.
- Finally, in the sixth and final level of my hierarchy of consciousness, the ability of conscious and aware selves to make abstract connections gives rise to language, which immeasurably expands the scale and scope of one's thoughts for consideration.

Note that these final two levels address what Dan Dennett calls "**the hard question of consciousness.**" According to Dennett, "the so-called hard *problem* of consciousness is a chimera, a distraction from the hard *question* of consciousness, which is once some content reaches consciousness, 'then what happens?'. … The question, more specifically, is: *Once some item or content 'enters consciousness', what does this cause or enable or modify?* For several reasons, researchers have typically either postponed addressing this question or failed to recognize—and assert—that their research on the 'easy problems' can be seen as addressing and resolving aspects of the hard question, thereby indirectly dismantling the hard problem piece by piece, without need of any revolution in science."

Dennett is probably right that a focus on all the tools and functions of consciousness ends up dismantling the hard problem. As all of the details for this have rolled in, the only thing left for the hard problem to cover is why there is consciousness at all. Well, we can never answer all why questions. Some things appear to just be here, like gravity, or electromagnetism, or any other fundamental force in the universe. And now that we have listed out all the basic ingredients of consciousness and observed that they have been around since the very beginning of life, that makes it trivially easy for me to posit pandynamism as an underlying identity in the universe, which helps us see the bridge between the forces which affect all matter and the subjectivity those forces cause in subjects once subjects emerge. As for the question, "*what does this cause or enable or modify?*", it clearly causes survival behaviour in ever expanding capacities towards more and more robust survival. More on that in the next question.

21. Does consciousness have a purpose?

The great evolutionary biologist **Ernst Mayr** is perhaps best known for helping to define the **modern synthesis**, but as an evolutionary philosopher, I'm also very interested in the distinction he made between **proximate and ultimate causations**. Mayr used this to show that biology just cannot be reduced to one thing; it must instead by analysed holistically. Proximate causation "explains biological function in terms of immediate physiological or environmental factors" whereas ultimate causation "explains traits in terms of evolutionary forces acting on them." Some examples make this clearer.

- Proximate description: "A female animal chooses to mate with a particular male during a mate choice trial. A possible proximate explanation states that one male produced a more intense signal, leading to elevated hormone levels in the female producing copulatory behaviour."
- Ultimate description: "Female animals often display preferences among male display traits, such as song. An ultimate explanation based on sexual selection states that females who display preferences have more vigorous or more attractive male offspring."

Note that the behaviour in these two examples is exactly the same. We just come to understand the situation better when we look at all the levels of causation. Nicholaas Tinbergen divided these two causations even further when he developed his **Four Questions**, which I have found to be crucial for understanding the entire story of consciousness. But in a wonderful paper by the philosopher Brandon Conley about how to **disentangle and integrate Mayr and Tinbergen's views**, we can see how Mayr's simpler distinctions help address a longstanding issue in the philosophy of biology. Conley writes:

- "According to Mayr, 'The clear recognition of two types of causation in organisms has helped to solve an important problem in biology, the problem of teleology.' A hallmark of the scientific revolution was the rejection of ancient and medieval applications of teleological reasoning to the cosmos. In slogan form, physics progressed when it came to focus on causes rather than purposes. Biology, on the other hand, and evolutionary biology in particular, appears to require reasoning about what a given trait is for, or what good it does for the organism. Biological explanation appears to be ineliminably teleological, but according to dominant conceptions of scientific reasoning, teleological reasoning is unscientific. There are three possible responses to this: (1) claim that biological explanation is not really teleological, (2) admit that biological explanation is not really scientific, or (3) claim that teleological reasoning can be scientific after all. Philosophers and scientists have tried all three, but Mayr argues that the class of processes that have been labeled as teleological are not unified and a combination of all three strategies is necessary."

For a quick reminder of what **telos/teleology** is, this traces back to Aristotle and can mean *purpose*, *intent*, *end*, or *goal*. In particular, "Aristotle used it in a more specific and subtle sense—the *inherent* purpose of each thing, the ultimate reason for each thing being the way it is, whether created that way by human beings or nature." As noted in the passage above, modern physics made progress when this concept was removed from the field. But it's important to acknowledge that this was only possible because non-living matter simply *reacts* to the forces that are applied to it. Biology, on the other hand, deals with living things that can *act* too. By definition, living things act to stay alive. They have evolved an internal drive to maintain their lives. An external observer can look at these actions and say they *want* to stay alive.

Another way of describing this is by using what Dan Dennett called the *intentional stance*. In a nice **profile of Dennett in the New Yorker**, this term was explained in the following clear and helpful way.

- "During the course of his career, Dennett has developed a way of looking at the process by which raw matter becomes functional. Some objects are mere assemblages of atoms to us and have only a physical dimension; when we think of them, he says, we adopt a 'physicalist stance'—the stance we inhabit when, using equations, we predict the direction of a tropical storm. When it comes to more sophisticated objects, which have purposes and functions, we typically adopt a 'design stance'. We say that a leaf's 'purpose' is to capture energy from sunlight, and that a nut and bolt are designed to fit together. Finally, there are objects that seem to have beliefs and desires, toward which we take the 'intentional stance'. If you're playing chess with a chess computer, you don't scrutinize the conductive properties of its circuits or contemplate the inner workings of its operating system (the physicalist and design stances, respectively); you ask how the program is thinking, what it's planning, what it 'wants' to do. These different stances capture different levels of reality, and our language reveals which one we've adopted."

Getting back to the question of teleology or purpose in biology, we know that physical matter reacts to physical forces. And in my post taking us **from physics to chemistry to biology**, I identified a set of "biological forces" that are missing from our scientific description of the world, but which clearly cause biology to react in predictable ways. Unlike with mere matter, however, living systems don't simply react in perfectly repeatable and definitively knowable ways. Biological life learns, grows, and changes how it reacts to biological forces by using the various aspects of consciousness at its disposal to sense and respond to the environment in order to drive its behaviour toward the ultimate goal of survival. This, of course, isn't a goal that has been designed by anyone. Nor is it even apparent to any beings in the grips of proximate goals. This is actually why Darwin faced **problems with the term natural selection** — it vaguely implied a selector — and so he toyed with the idea of calling the central force in evolution "natural preservation" instead. But logically, the survival goal must be the ultimate necessary outcome in a universe where things change, and nothing lives forever. Any and all proximate goals that don't work towards this will end up going extinct.

With all of this in mind, we can now answer this question, and choose from among Mayr's three responses. Consciousness does indeed have a purpose or telos, but it is one that emerges from selection forces rather than intentional designs. Because living beings *act* as well as *react*, it is necessary to look at *underlying causes* (biological forces) as well as *teleological purposes* (survival, ultimately) if we want to holistically understand the way that life works. In this way, teleological reasoning can be scientific after all (Mayr's third choice), and in fact it is necessary for full scientific reckonings. (This is also why telos sits at the top of **my evolutionary hierarchy of needs**.) Consciousness, in all its emergent and expanding properties, helps life sense and respond to the world in order to maintain its survival and make it more and more robust.

QUESTIONS FROM THOSE WHO DOUBT OR DISBELIEVE NATURALISM

For this next batch of questions, I wanted to make sure I wasn't just preaching to the choir or responding to people who already held favourable dispositions toward the naturalist project. I wanted to make sure I properly understood objections from the other side. To that end, I have

some questions from Raymond Tallis and Philip Goff, which I'll cover in that order because it takes us through their points in increasing levels of difficulty and importance.

To start, I have three questions from Ray Tallis' recent book ***Seeing Ourselves: Reclaiming Humanity from God and Science***. Tallis is a retired physician and **patron of Humanists UK** who was once named as "**one of the top living polymaths in the world.**" A local philosophy group really likes his work so I've had a chance to meet him in person a few times and I was once scheduled to discuss *Seeing Ourselves* with him in a public Humanist meeting, but that event fell through. After reading the first few chapters and plucking out the questions below, you may see why this cancellation was for the best.

22. Why doesn't a chair feel my bottom?

It's hard to believe this is an actual question, so let's quote Tallis directly to see what he really means by this.

- "If energy exchange entirely accounted for touch then it would be as reasonable for the chair on which I am sitting to feel my bottom as for my bottom to feel the chair: the ontological equality of myself as an object among objects does not translate into a dialogue of equal partners. That ontological equality, however, is central to materialist naturalism." (***Seeing Ourselves*** p. 54)
- "The causal theory of perception, in which all parties are subject to the Dennettian edict of being subject to *the same physical principles, laws, and raw materials* that operate elsewhere in nature offers nothing to explain the differentiation between the perceiving subject and object of perception; between the perceiver and the perceived." (***S.O.*** p. 54)

What an absurd caricature of the naturalist position! Tallis loves taking cheap shots at Dan Dennett like this, even though he grossly misunderstands him. (See **my review** of *Seeing Ourselves* for more on that.) Can naturalists explain why the blind naked mole rat doesn't see me even though I can see him? Of course we can! It is not just that "stuff feels" which explains the sense of touch. It is the *structure* of that stuff ("that ontological equality") which enables subjectivity to emerge in subjects via various **mechanisms**.

23. How can consciousness survive sleep?

Here is yet another daft-sounding question that requires direct quotation for context.

- "One way of addressing the so-called combinatorial problem — the problem of explaining how sequins of consciousness spread through the world add up to a subject — is to deflate the subject. The subject is reduced to successive experiences, or time-slices of a flow of experience: there are no persisting subjects; each distinct experience has its own experiencer. This merely transfers the problem to that of explaining how experiences add up to a subject who has a sense of herself at a time and over time and is acknowledged to be a person by other subjects also acknowledged to be persons. It is not at all clear by what means, by who or what, the thin subjects are stitched together and how we would survive sleep or episodes of unconsciousness." (***S.O.*** note 84 on p.393 referring to p.66)

So, Tallis is really referencing the binding problem here, which I addressed above in question 9. Admittedly, we don't have full mappings of all the neuroscience in the animal kingdom yet

to give full explanations for how subjectivity is stitched together (to the extent that it is, anyway, since it **varies across the animal kingdom**). I think **Jeff Hawkins'** solution shows promise, but it's too early to say for sure.

What isn't a solution to the binding problem is "to deflate the subject [so] there are no persisting subjects." Once again, it is the *structure* of material that gives materialists their mechanisms for consciousness, and that structure clearly survives sleep and episodes of unconsciousness. (And the structures change slightly to cause those states of sleep and unconsciousness too.)

24. How could consciousness have possibly emerged from lower organisms?

There is another legitimate question, which I addressed in questions 4 and 20 above. I also described the actual evolutionary history of consciousness in much more detail in my post about **our shared history (phylogeny)**. Granted, this may be "**the hardest problem in consciousness studies**," but while Tallis grapples with it, he displays such a bewildering lack of understanding about evolution that it's no wonder he doesn't see the materialist argument. Some more direct quotes will show the paucity of Tallis' beliefs, which I'll just comment on briefly after each one so we can move on to better foes.

- "Darwinism highlights (if inadvertently) a serious objection to Darwinitis, namely, that Darwinism gives no account of the *emergence* of consciousness from the material world of which lower organisms are a part." (***S.O.*** p. 66)

I have now given just such a Darwinian account, so perhaps that should be considered, but let's be clear here, no other metaphysical theories have provided an account either. And we're still learning about the universe so there's no cause to dismiss naturalism as Darwinitis just yet. Also, how dare you say that you believe in evolution but place "lower organisms" in the material world separate from humans. When did that break occur exactly? This is sheer hubris, and it's dangerous to the survival of life too.

- "There are at least two major obstacles to a materialist evolutionary account of human consciousness: the first is the question of the nature of the supposed competitive advantage conferred by being conscious; and the second is the question of how, even if consciousness *did* confer survival benefit, it could have been generated between unconscious species; that, as a result of the blood bath of natural selection, the universe could get to know itself." (***S.O.*** p. 67)

That first major obstacle is legitimate and addressed in question 21 above. But that second obstacle is an embarrassment of logic by which it's hard to see how Tallis thinks *any* novel trait could emerge from evolution. Yikes.

- "Given that pre-conscious processes do so much work; there is not much useful work left for consciousness to do. To address this question properly, we need to go back to a putative moment when the first spark of consciousness was lit and ask what possible additional advantage would an organism with (say) an array of photosensitive cells gain from being *aware* of the light it is responding to? The "obvious" benefits vanish when we acknowledge: a) That the best route to replication of the genome must be via utterly reliable mechanisms based on the (by definition) unbreakable laws of nature rather than the vagaries of (conscious) decision making; b) Evolution should favour appropriate

action, but it is not evident that this should have to be mediated by true belief or indeed any belief; and c) That unconscious mechanisms have been perfectly adequate to bring about things that consciousness could not even dream of, such as the basis of the organism's self-maintenance (including its voluntary actions), the spectacular achievement of the development of the human brain *in utero*, and the entire evolutionary process." (***S.O.*** p. 67)

Wow. This is just a mishmash of very muddled thinking. First, Tallis appears to consider "consciousness" to just be "conscious awareness" which we've established above will always get you into trouble. In my theory, "pre-conscious" processes are just lower levels of consciousness. Conscious awareness cannot arrive for a mere "array of photosensitive cells" because there is no structure there to evaluate the affect, intention, and prediction levels that are further down in my hierarchy. But they must all be there before "the first spark" of conscious awareness emerges. I could forgive such confusion about the confusing terminology used in consciousness studies, but the three-part argument in the latter half of this quote is inexcusable. a) The best route to *survival* of genomes is *not* "utterly reliable mechanisms" because that would leave no room for change and adaptation. Perfectly repeated organisms (if they ever existed) would have gone extinct at the first sign of trouble. No laws of nature stop mutations and genetic drift from happening. And conscious decision-making (to focus only on the conscious awareness that Tallis is describing) allows beings who have attained that ability to conduct mental trials and errors so their ideas can perish rather than themselves. That is hardly a vagary of living successfully. b) Useful beliefs about the world improve one's actions. Those are favoured by evolution. c) Unconscious actions are indeed driven by lower levels of consciousness, particularly the cognitions in my levels of affect and intention. Conscious awareness emerges on top of those and enables yet further behavioural adjustments by these already very finely tuned biological machines.

- "If, say, consciousness is necessary for learning and plasticity, then we have to ask why is it not always necessary for learning and plasticity. In most cases, learning and plasticity do not require the conscious participation of the organism." (***S.O.*** p. 67)

As discussed above in question 20, conscious awareness seems to enable "trace conditioning" which is another form of learning that is added to all the unconscious learning that is possible using lower levels of consciousness. Not all learning and plasticity is equal.

- Natural selection can act only on what is already available. It seems inconceivable that it could *generate*, even less requisition, entirely novel properties such as consciousness. The clash between forms of organic matter over limited means to life seems hardly likely to give rise to something that goes beyond the material world, namely intentionality. (***Seeing Ourselves*** p. 69)

This is simply more evidence of Tallis' complete lack of understanding about evolution. This is embarrassing now, and fully illustrates why Tallis' objections are so easily cast aside. Time to move on and see what we can learn from better foes. The next two questions come from the philosopher Philip Goff who is the new poster boy for panpsychism. I covered his general views in **the fourth post in this series**, but let's take a look at some of his specific objections now that my own theory has been fully elucidated.

25. Is conscious experience outside of the realm of science?

This first question comes from a short paper by Goff titled, "**Why Science Can't Explain Consciousness**." This is clearly related to his longer book ***Galileo's Error***, but based on this paper (and the chance I had to personally hear Goff in a small meeting in Durham) I don't think it's necessary to read that. Let's look at just a few quotes from Goff to see if you agree before I hit back with my response.

- "Here is Galileo describing his conception of matter: '…Hence I think that tastes, odours and colours, and so on are no more than mere names as far as the object in which we place them is concerned, and that they reside only in the consciousness.'"
- "In taking the qualities of consciousness not to be instantiated by material bodies, Galileo seems to be taking the qualities of consciousness to reside in an immaterial substance."
- "This rough sketch of nature was a short time later turned into a rigorous metaphysical view by Descartes. For Descartes, colours and smells and odours result from the interaction of immaterial minds with physical bodies."
- "As the result of this radical new Galilean/Cartesian metaphysics, we have, perhaps for the first time in history, a picture of the material world such that its nature can be completely captured in mathematics. Sensory qualities—the taste of the lemon, the smell of the flowers—cannot be entirely captured in mathematical language. So long as philosophers took such qualities to reside in the physical world, the scientific revolution was impossible. But once the physical world had been divested of qualitative nature, the remaining quantitative nature, concerning the way in which objects fill space, could be entirely captured in geometry. By putting sensory qualities in the conscious mind, and putting the conscious mind outside of the physical world, Galileo and Descartes provided the metaphysical underpinnings of the scientific revolution."
- "Physics, for all its virtues, gives us a radically incomplete picture of the world. It provides a description of the world that necessarily abstracts from the one aspect of concrete reality we know for certain to exist: the qualities of consciousness that are immediately and indubitably known to each of us."

There is much more in Goff's paper (and presumably in his book too), but this is enough to see that he's relying on dualist metaphysics from the 1600's that was **very poorly argued** at the time and has largely been discarded by modern thinkers. There is no need to think we have all placed qualitative research into some immaterial realm just because Galileo may have written about it that way. In question 7 above about whether consciousness would always remain a mystery, I wrote:

- And while we must acknowledge there are epistemological barriers to what any one person can know about their brains or the consciousness of others, we can "combine the two perspectives within certain experimental contexts. Both first-person and third-person scientific data about the brain and consciousness can be acquired and used to solve the hard problem." Scientists do this all the time.

As for "the qualities of consciousness that are immediately and indubitably known to each of us," there are no such things and Goff's argument evaporates once this illusion is broken. I particularly like these two quotes from Dan Dennett's paper "**Facing Up to the Hard Question of Consciousness**" for dispatching this nonsense.

- "Over the past few centuries, our understanding of how vision is accomplished has grown magnificently, and one of the striking facts about what we have learned is that until scientists told us, we had no idea at all, no 'privileged access', to the complicated activities of the optic nerve, the occipital cortex, and even the activities of our eyeballs."

- "The fact is, the traditional claim that our conscious minds are immediately and maybe even perfectly known to each of us is wildly false. The psychologist Karl Lashley once suggested provocatively that 'no activity of the mind is ever conscious', by which he meant to draw our attention to the inaccessibility of the processing that we know must go on when we think. What 'we' do 'have access to' is the contents and apparent temporal order of those contents, but how these contents, these representations of properties, objects and events, manage to represent what they do, and how they are generated when they 'appear' to 'us' is completely off-limits to introspection."

26. Are minds everywhere? What about panpsychism?

Here's one more quick question, based on Goff's notorious essay "**Panpsychism is crazy, but it's also most probably true**." The argument is very basic, so let me give it in a few quotes.

- "According to panpsychism, the smallest bits of matter—things such as electrons and quarks—have very basic kinds of experience; an electron has an inner life. The main objection made to panpsychism is that it is 'crazy' and 'just obviously wrong'. It is thought to be highly counterintuitive to suppose that an electron has some kind of inner life, no matter how basic, and this is taken to be a very strong reason to doubt the truth of panpsychism."
- "Scientific support for a theory comes not merely from the fact that it explains the evidence, but from the fact that it is the *best* explanation of the evidence, where a theory is 'better' to the extent that it is more simple, elegant, and parsimonious than its rivals."
- "In fact, the only thing we know about the intrinsic nature of matter is that some of it—the stuff in brains—involves experience. We now face a theoretical choice. We either suppose that the intrinsic nature of fundamental particles involves experience, or we suppose that they have some entirely unknown intrinsic nature."

That's it?! Well, there's actually a third choice that is just as simple, elegant, and parsimonious, which doesn't result in the just obviously wrong and counterintuitive notion that an electron has some kind of inner life. That's my theory of pandynamism, which I explained above in question 4 about how we might be able to get minds from matter. It's not just that matter feels subjectivity. It's that you need the right *structure* for that subjectivity to emerge in actual subjects. I'll repeat my comparison of these two theories here:

- As an example, take the simplest force. What does it take to 'feel' gravity? For us humans, it's registering the difference between inner ear liquids as our movements in space accelerate or decelerate. Can a rock or a photon ever experience this? No. Why not? Because there is no structure in its makeup by which it could gain such information. Panpsychism is therefore a non-starter for me, but *pandynamism* could explain how subjectivity is a fundamental feature of the universe, yet only emerges as living organisms emerge, thus bridging the explanatory gap and providing a coherent answer to the hard problem.

QUESTIONS FROM DAVID CHALMERS

Okay, that's enough from those two foes of naturalism. Now for the full arguments of the man best known for throwing up stumbling blocks for consciousness studies, David Chalmers.

He coined the phrase *the hard problem*, but what's so hard about it and what else does he object to and worry about? To find out, I've carefully gone through his 30-page paper "The Problem of Consciousness" (**TPoC**, hereafter). According to the abstract, "this paper is an edited transcription of a talk at the 1997 Montreal symposium on 'Consciousness at the Frontiers of Neuroscience.'" I found it to be an incredibly useful paper and would like to finish up this long list of FAQs (and my entire consciousness series!) by going through it in detail.

27. What are the easy problems of consciousness?

According to **TPoC**,

- The easy problems of consciousness include those of explaining the following phenomena:
  - the ability to discriminate, categorize, and react to environmental stimuli;
  - the integration of information by a cognitive system;
  - the reportability of mental states;
  - the ability of a system to access its own internal states;
  - the focus of attention;
  - the deliberate control of behavior;
  - the difference between wakefulness and sleep.
- All of these phenomena are associated with the notion of consciousness. For example, one sometimes says that a mental state is conscious when it is verbally reportable, or when it is internally accessible. Sometimes a system is said to be conscious of some information when it has the ability to react on the basis of that information, or, more strongly, when it attends to that information, or when it can integrate that information and exploit it in the sophisticated control of behavior. We sometimes say that an action is conscious precisely when it is deliberate. Often, we say that an organism is conscious as another way of saying that it is awake.
- In each case, an appropriate cognitive or neurophysiological model can clearly do the explanatory work. If these phenomena were all there was to consciousness, then consciousness would not be much of a problem. Although we do not yet have anything close to a complete explanation of these phenomena, we have a clear idea of how we might go about explaining them. This is why I call these problems the easy problems.

That second bullet point illustrates the very wide variance in the usage of the term *consciousness*, which is another reason why I've done my best to rope them all into a comprehensive hierarchy. Researching the Tinbergen history of all of these easy problems is precisely what enabled me to set out the hierarchy as I have, while still recognising there are plenty of details to be filled in yet.

28. What is the hard problem of consciousness?

According to **TPoC**,

- The really hard problem of consciousness is the problem of experience. When we think and perceive, there is a whir of information-processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is something it is like to be a conscious organism. This subjective aspect is experience.

- Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does.
- If any problem qualifies *as* the problem of consciousness, it is this one. In this central sense of 'consciousness', an organism is conscious if there is something it is like to be that organism, and a mental state is conscious if there is something it is like to be in that state. Sometimes terms such as 'phenomenal consciousness' and 'qualia' are also used here, but I find it more natural to speak of 'conscious experience' or simply 'experience'. Another useful way to avoid confusion (used by e.g., Newell 1990 Chalmers 1996) is to reserve the term 'consciousness' for the phenomena of experience, using the less loaded term 'awareness' for the more straightforward phenomena described earlier. If such a convention were widely adopted, communication would be much easier; as things stand, those who talk about 'consciousness' are frequently talking past each other. (**Chalmers**)

Agreed! As we've seen throughout this series, researchers and philosophers frequently are talking about 'awareness' while others have something else in mind for 'consciousness' so they are indeed talking past one another or not getting to the root of the problem. The entire "phenomena of experience" is what I'm after with my comprehensive hierarchy of consciousness. And when we see how those phenomena exists across the entire spectrum of life, and over life's entire evolutionary history, but it does not seem to extend into any non-living organic systems, then it makes sense to posit *pandynamism* (see question 4 for details) as the theory for why subjectivity is a fundamental identity of the universe but it only arises in subjects.

29. What does it take to solve the easy problems of consciousness?

According to **TPoC**,

- The easy problems are easy precisely because they concern the explanation of cognitive *abilities* and *functions*. To explain a cognitive function, we need only specify a mechanism that can perform the function. The methods of cognitive science are well-suited for this sort of explanation, and so are well-suited to the easy problems of consciousness. By contrast, the hard problem is hard precisely because it is not a problem about the performance of functions. The problem persists even when the performance of all the relevant functions is explained.
- Once we have specified the neural or computational mechanism that performs the function of verbal report, for example, the bulk of our work in explaining reportability is over. … All it could *possibly* take to explain reportability is an explanation of how the relevant function is performed; the same goes for the other phenomena in question.

Not quite! **Functions** and **mechanisms** are only half of **Tinbergen's four questions**. We gain a lot of insight from looking through the **ontogeny** and **phylogeny** of these phenomena too. Seeing these evolutionary histories is precisely how we see the logic and empirical data for putting everything into the ordered hierarchy as I have done.

30. Is the hard problem really different than the easy ones?

According to **TPoC**,

- When it comes to conscious experience, this sort of [easy] explanation fails. What makes the hard problem hard and almost unique is that it goes beyond problems about the performance of functions. … even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience—perceptual discrimination, categorization, internal access, verbal report—there may still remain a further unanswered question: *Why is the performance of these functions accompanied by experience?*
- If someone says, "I can see that you have explained how DNA stores and transmits hereditary information from one generation to the next, but you have not explained how it is a gene", then they are making a conceptual mistake. All it means to *be* a gene is to be an entity that performs the relevant storage and transmission function. But if someone says, "I can see that you have explained how information is discriminated, integrated, and reported, but you have not explained how it is *experienced*", they are not making a conceptual mistake.
- We know that conscious experience does arise when these functions are performed, but the very fact that it arises is the central mystery. There is an *explanatory gap* (a term due to Levine 1983) between the functions and experience, and we need an explanatory bridge to cross it.

When I first discussed the hard problem in **post 3** of this series, I noted that "I'd like to make a distinction for Chalmers' hard problem between the how and the why. *How* do physical processes lead to subjective experience? *Why* do physical processes lead to subjective experience? The ultimate why is ultimately an impossible problem." Chalmers' hard problem is clearly a why problem, and perhaps an impossible why.

In the opening of **a recent Brain Science podcast**, the neuroscientist Anil Seth said much the same and pushed back on Chalmers by saying,

- "It's not essential for a branch of science to explain why the phenomenon is there in the first place. Physics…doesn't tell us why there is a universe in the first place to explain. We often set a higher bar for consciousness than we do for other things. Partly because we are conscious. We want that intuitive *a ha* that makes sense. There's absolutely no reason why a scientific account of consciousness should be intuitively satisfying. It would be nice if it were, but that's not strictly necessary."

In some respects, Chalmers is playing a game of eternal regression here by just continuing to ask *why* for consciousness. But by doing so, he ends up driving home the point that perhaps the experience of subjectivity is just fundamental to the universe. More on this later.

31. Can we see an example? Is the binding problem hard or easy?

According to **TPoC**,

- Binding is the process whereby separately represented pieces of information about a single entity are brought together to be used by later processing, as when information about the color and shape of a perceived object is integrated from separate visual pathways.

- Crick and Koch hypothesize that binding may be achieved by the synchronized oscillations of neuronal groups representing the relevant contents. When two pieces of information are to be bound together, the relevant neural groups will oscillate with the same frequency and phase.
- Such a theory would be valuable, but it would tell us nothing about *why* the relevant contents are experienced. … Even if it is accepted, the explanatory question remains: *Why* do the oscillations give rise to experience?

So, the binding problem is unsolved for now, but it is still easy. This passage perfectly illustrates how Chalmers uses the question of *why* to keep the hard problem out of reach.

32. How have people tried to answer the hard problem?

According to **TPoC**,

- In placing this sort of work with respect to the problem of experience, a number of different strategies are available. It would be useful if these strategic choices were more often made explicit.
- The first strategy is simply to *explain something else*. Some researchers are explicit that the problem of experience is too difficult for now, and perhaps even outside the domain of science altogether.
- The second choice is to take a harder line and deny the phenomenon. According to this line, once we have explained the functions such as accessibility, reportability, and the like, there is no further phenomenon called "experience" to explain.
- In a third option, some researchers *claim to be explaining* experience in the full sense. These researchers (unlike those above) wish to take experience very seriously; they lay out their functional model or theory and claim that it explains the full subjective quality of experience (e.g., Flohr 1992 Humphrey 1992). The relevant step in the explanation is usually passed over quickly, however, and usually ends up looking something like magic.
- A fourth, more promising approach appeals to these methods to *explain the structure of experience*. … At best, it takes the existence of experience for granted and accounts for some facts about its structure, providing a sort of nonreductive explanation of the structural aspects of experience (I will say more on this later). This is useful for many purposes, but it tells us nothing about why there should be experience in the first place.
- A fifth and reasonable strategy is to *isolate the substrate of experience*. … the strategy is clearly incomplete. For a satisfactory theory, we need to know more than *which* processes give rise to experience; we need an account of why and how.

I agree with Chalmers that these strategies do not answer his hard problem. But that is, of course, because he has probably placed it out of reach with his infinite regression of why questions. Still, it is interesting to see the various strategies that have been employed so far by people who don't seem to fully grasp what Chalmers is getting at. I don't believe the theory I have developed in this series misunderstands Chalmers' hard problem, however, nor does it 1) explain something else, 2) deny the phenomenon, 3) pass over it like magic, 4) take it for granted, or 5) assume it is isolated to one substrate.

33. So, what else is needed and why do physical accounts fail?

According to **TPoC**,

- To account for conscious experience, we need an *extra ingredient* in the explanation. This makes for a challenge to those who are serious about the hard problem of consciousness: What is your extra ingredient, and why should *that* account for conscious experience?
- At the end of the day, the same criticism applies to *any* purely physical account of consciousness. For any physical process we specify there will be an unanswered question: Why should this process give rise to experience?
- A physical account can *entail* the facts about structures and functions: once the internal details of the physical account are given, the structural and functional properties fall out as an automatic consequence. But the structure and dynamics of physical processes yield only more structure and dynamics, so structures and functions are all we can expect these processes to explain.

This is the same issue for all fundamental properties of the universe. We don't know *why* matter, gravity, or electromagnetism exist and behave the way that they do. One cannot get outside of all frames of reference to understand what is going on inside them. To paraphrase the eco-philosopher Arne Næss, one cannot blow a balloon up from the inside. This appears to be the same issue for explaining the subjective phenomena of consciousness. It just seems to happen in all living things, and my theory of *pandynamism* explains why this might be so for us, but not be so for non-living things.

34. Is this the same problem we faced with vitalism?

According to **TPoC**,

- This might seem reminiscent of the vitalist claim that no physical account could explain life, but the cases are disanalogous. … Once it turned out that physical processes could perform the relevant functions, vitalist doubts melted away. … With experience, on the other hand, physical explanation of the functions is not in question. The key is instead the *conceptual* point that the explanation of functions does not suffice for the explanation of experience.

Chalmers is right that the phenomenon of life (the fact that living beings act as living beings) is an objective observation that can be explained away once the mechanisms of life are understood. The internal subjective feeling of consciousness is not like this. There is an abundance of evidence for subjectivity in living organisms, as explained in question 8 above about zombies, but it is not an obvious phenomenon from the outside and we certainly cannot crawl into another's physical embodiment to truly know "what it feels like" to be them. Still, understanding the physical processes of life melted away any thoughts of extra non-physical ingredients for life. And similarly, understanding the physical processes for all aspects of consciousness in my comprehensive hierarchy is melting away any thoughts for any extra non-physical ingredients for consciousness. What is left? Just the simplest observations that subjectivity *does* occur in living things, and it does *not* appear to occur in non-living things.

35. So, is consciousness just fundamental?

According to **TPoC**,

- Although a remarkable number of phenomena have turned out to be explicable wholly in terms of entities simpler than themselves, this is not universal. In physics, it occasionally happens that an entity has to be taken as *fundamental*. Fundamental entities are not explained in terms of anything simpler. Instead, one takes them as basic, and gives a theory of how they relate to everything else in the world. For example, in the nineteenth century it turned out that electromagnetic processes could not be explained in terms of the wholly mechanical processes that previous physical theories appealed to, so Maxwell and others introduced electromagnetic charge and electromagnetic forces as new fundamental components of a physical theory. To explain electromagnetism, the ontology of physics had to be expanded. New basic properties and basic laws were needed to give a satisfactory account of the phenomena.
- Other features that physical theory takes as fundamental include mass and space-time. No attempt is made to explain these features in terms of anything simpler. But this does not rule out the possibility of a theory of mass or of space-time. There is an intricate theory of how these features interrelate, and of the basic laws they enter into. These basic principles are used to explain many familiar phenomena concerning mass, space, and time at a higher level.
- I suggest that a theory of consciousness should take experience as fundamental. We know that a theory of consciousness requires the addition of *something* fundamental to our ontology, as everything in physical theory is compatible with the absence of consciousness.

I agree with Chalmers that the subjective feeling of consciousness is fundamental in this way. But, as explained above, it does not arise in non-living matter because there is no structure there that constitutes a *subject*, which could then experience *subjectivity*. Our physical theories are compatible with the *reaction* of all physical matter to physical forces. But Chalmers is wrong about our biological observations. Those require something else to explain the *actions* that living organisms take. (See question 8 above for a discussion of the preposterousness of non-conscious zombies.) Defining consciousness as I have ("an infinitesimally growing ability to sense and respond to any or all biological forces in order to meet the needs of survival"), and then explaining what these biological forces are, and how *pandynamism* gave rise to feeling them, gives us a coherent physical explanation for all of our observations—both the objective ones and subjective ones, in physics, chemistry, and biology.

36. If we accept consciousness is fundamental, then what?

According to **TPoC**,

- We might add some entirely new nonphysical feature, from which experience can be derived, but it is hard to see what such a feature would be like. More likely, we will take experience itself as a fundamental feature of the world, alongside mass, charge, and space-time. If we take experience as fundamental, then we can go about the business of constructing a theory of experience.
- Where there is a fundamental property, there are fundamental laws. A nonreductive theory of experience will add new principles to the furniture of the basic laws of nature. These basic principles will ultimately carry the explanatory burden in a theory of consciousness.
- Just as we explain familiar high-level phenomena involving mass in terms of more basic principles involving mass and other entities, we might explain familiar phenomena involving experience in terms of more basic principles involving experience and other entities.

- Of course, by taking experience as fundamental, there is a sense in which this approach does not tell us why there is experience in the first place. But this is the same for any fundamental theory. Nothing in physics tells us why there is matter in the first place, but we do not count this against theories of matter. Certain features of the world need to be taken as fundamental by any scientific theory. A theory of matter can still explain all sorts of facts about matter, by showing how they are consequences of the basic laws. The same goes for a theory of experience.
- Nothing in this approach contradicts anything in physical theory; we simply need to add further *bridging* principles to explain how experience arises from physical processes. There is nothing particularly spiritual or mystical about this theory—its overall shape is like that of a physical theory, with a few fundamental entities connected by fundamental laws. It expands the ontology slightly, to be sure, but Maxwell did the same thing.

Yes! This is the route I have taken, and I have started to sketch these new principles and fundamental laws of pandynamism and biological forces.

37. Is this fundamental view a sort of dualism?

According to **TPoC**,

- In particular, a nonreductive theory of experience will specify basic principles telling us how experience depends on physical features of the world. These *psychophysical* principles will not interfere with physical laws, as it seems that physical laws already form a closed system. Rather, they will be a supplement to a physical theory. A physical theory gives a theory of physical processes, and a psychophysical theory tells us how those processes give rise to experience.
- This position qualifies as a variety of dualism, as it postulates basic properties over and above the properties invoked by physics. … If the position is to have a name, a good choice might be *naturalistic dualism*.

Except that it's not dualism! There isn't a dualism of matter + space-time + electromagnetism + any other fundamentals of physics. It's all just the list of properties in a monist physical universe. Adding subjectivity as a fundamental feeling that emerges in physical material once that material attains the form of self-sustaining life does not change this monistic view.

Furthermore, Chalmers is right that "*a physical theory* gives a theory of *physical processes*" but he is wrong about what *a psychophysical theory* then gives us. To extend the comparison logically, *a psychophysical theory* gives us …wait for it… a theory of *psychophysical processes*! That is exactly what my theory of biological forces helps us to understand—the *psychophysical processes* going on in living organisms, which drives their actions over and above the simple reactions of the physical and chemical laws of nature. If subjective consciousness is truly taken as fundamental, there is no need to "tell us how processes give rise to experience." That's fundamental!

38. If consciousness is fundamental, shouldn't it be simple to describe?

According to **TPoC**,

- If this view is right, then in some ways a theory of consciousness will have more in common with a theory in physics than a theory in biology. Biological theories involve no principles that are fundamental in this way, so biological theory has a certain complexity and messiness to it; but theories in physics, insofar as they deal with fundamental principles, aspire to simplicity and elegance. The fundamental laws of nature are part of the basic furniture of the world, and physical theories are telling us that this basic furniture is remarkably simple. If a theory of consciousness also involves fundamental principles, then we should expect the same. The principles of simplicity, elegance, and even beauty that drive physicists' search for a fundamental theory will also apply to a theory of consciousness.
- Finally, the fact that we are searching for a fundamental theory means that we can appeal to such nonempirical constraints as simplicity, homogeneity, and the like in developing a theory. We must seek to systematize the information we have, to extend it as far as possible by careful analysis, and then make the inference to the simplest possible theory that explains the data while remaining a plausible candidate to be part of the fundamental furniture of the world.

Yes! I think my theory is pretty simple. I'm glad that is a feature and not a bug.

39. What about Chalmers' own theories?

According to **TPoC**,

- In what follows, I present my own candidates for the psychophysical principles that might go into a theory of consciousness. The first two of these are *nonbasic principles* — systematic connections between processing and experience at a relatively high level. These principles can play a significant role in developing and constraining a theory of consciousness, but they are not cast at a sufficiently fundamental level to qualify as truly basic laws. The final principle is my candidate for a *basic principle* that might form the cornerstone of a fundamental theory of consciousness.
- The principle of structural coherence: this is a principle of coherence between the *structure of consciousness* and the *structure of awareness*. … If we accept the principle of coherence, the most *direct* physical correlate of consciousness is awareness: the process whereby information is made directly available for global control. … This principle reflects the central fact that even though cognitive processes do not conceptually entail facts about conscious experience, consciousness and cognition do not float free of one another but cohere in an intimate way.
- The principle of organizational invariance: this principle states that any two systems with the same fine-grained *functional organization* will have qualitatively identical experiences. If the causal patterns of neural organization were duplicated in silicon, for example, with a silicon chip for every neuron and the same patterns of interaction, then the same experiences would arise.
- The double-aspect theory of information: I understand information in more or less the sense of Shannon (1948). Where there is information, there are *information states* embedded in an information space. An *information space* has a basic structure of *difference* relations between its elements, characterizing the ways in which different elements in a space are similar or different, possibly in complex ways. … To borrow a phrase from Bateson (1972), physical information is a *difference that makes a difference*. The double-aspect principle stems from the observation that there is a direct isomorphism between certain physically embodied information spaces and certain *phenomenal* (or experiential) information spaces.

Regarding Chalmers' first principle of structural coherence, this is so typical of a philosopher to focus on such a high level of conscious awareness rather than starting at the bottom of consciousness. Chalmers says "the most *direct* physical correlate of consciousness is awareness: the process whereby information is made directly available for global control" but by tracing the evolutionary history of consciousness, we see that this comes far after all the cognitive abilities in my hierarchies of affect, intention, and prediction. In fact, I would go so far as to say that awareness *can only* arise after these other abilities are present. I agree that "consciousness and cognition do not float free of one another," so the general principle of structural coherence is fine, but you have to do a Tinbergen analysis to see *all* of the cognitions that are built into consciousness. And this affects Chalmers' other theories.

The second principle of organizational invariance is very hard to accept given the impact that tiny bits of chemical drugs can have on our conscious experience. The matter seems to matter! Perhaps the carbon lifeforms that have slowly, slowly arisen over the billions of years of Earth's evolutionary history have found their way here precisely because their structure yields experiences that drive towards survival and away from extinction. Maybe a silicon-based replica would love the feeling of electricity coursing through its body too much and would quickly zap itself into oblivion like a moth to a flame. We certainly don't know that, but it seems just as possible as Chalmers' speculation. And given the fact that no other substrates for life *have* arisen here, it seems more likely that functional organization is not enough for "qualitatively identical experiences."

Finally, I see Chalmer's basic principle about information as a simple truism. Yes, there is "a direct isomorphism between certain physically embodied information spaces and certain *phenomenal* (or experiential) information spaces." But this is exactly because the universe is physical. Any changes in experience *are* associated with physical changes. And both of these can be expressed as information. But information can be abstracted from everything! There isn't anything that follows from this about information itself. More on this in the next question.

40. Is consciousness all about information processing?

According to **TPoC**,

- This [basic principle] leads to a natural hypothesis: that information (or at least some information) has two basic aspects, a physical aspect and a phenomenal aspect. This has the status of a basic principle that might underlie and explain the emergence of experience from the physical. Experience arises by virtue of its status as one aspect of information, when the other aspect is found embodied in physical processing.
- If the principle of organizational invariance is to hold, then we need to find some fundamental *organizational* property for experience to be linked to, and information is an organizational property par *excellence*.
- Wheeler (1990) has suggested that information is fundamental to the physics of the universe. According to this "it from bit" doctrine, the laws of physics can be cast in terms of information, postulating different states that give rise to different effects without actually saying what those states *are*. It is only their position in an information space that counts. If so, then information is a natural candidate to also play a role in a fundamental theory of consciousness. We are led to a conception of the world on which information is

truly fundamental, and on which it has two basic aspects, corresponding to the physical and the phenomenal features of the world.

To me, this has it exactly backwards. The bit comes from the it! Information is just abstraction from the physical world. And I found abstraction to be the final level of consciousness that emerges in my hierarchy. The ability to have abstract thoughts (i.e., the ability to represent the physical with language and symbols) is what gives a consciousness the freedom to think infinitely far and wide, formulate imagined hypotheses about the world, and communicate our thoughts about all of these thoughts. Only physicists, mathematicians, and philosophers who spend their whole lives in this abstract world could actually think that this is the primary cause of reality. We must not follow them down this hole.

In **a long Psychology Today post** about the spirituality of **Integrated Information Theory**, we see the trouble this leads people into. The author there said,

- "Let's follow the logic of this idea and see how it holds up. We know that certain brain states feel like something. Brain states are just information states. Therefore, information feels like something. Sounds pretty solid.

No! Brain states are not *just* information states. They are *specific* information processors that are processing *specific* information. Information and information states are everywhere because everything in reality (and in imagination!) can be abstracted. One could similarly claim that boulders are *just* information states, therefore information feels like *nothing*. The bad logic is the same! So, contrary to what Chalmers states, information is not an *organizational property par excellence*. Information is that map in a joke which says "**Scale: 1 mile = 1 mile**."

Chalmers says, "Experience arises by virtue of its status as one aspect of information, when the other aspect is found embodied in physical processing." I find it much easier to say that experience arises from physical processing, but this subjectivity only emerges and grows along with the emergence of subjects as the requisite structures form that can capture and register these experiences.

41. So, can we make progress and answer the hard problem of consciousness?

According to **TPoC**,

- Most existing theories of consciousness deny the phenomenon, explain something else, or elevate the problem to an eternal mystery. I hope to have shown that it is possible to make progress on the problem even while taking it seriously. To make further progress, we will need further investigation, more refined theories, and more careful analysis. The hard problem is a hard problem, but there is no reason to believe that it will remain permanently unsolved.

Yes, but then no. We can indeed make progress on all the easy problems of consciousness, and that appears to melt away the hard problem in the same way that a hard problem of electromagnetism melted away. Given that all the building blocks of consciousness are widespread across the entirety of life throughout its evolutionary history, it is likely that the experience of subjectivity is a fundamental property of our universe. But any further questions about *why* this universe, or all universes, are like that appear to be permanently unsolvable. One cannot always get outside of one's frame of reference in order to understand

everything within that frame. **Godel's incompleteness theorems** and **Tarski's undefinability theorem** are good examples of this principle. But hopefully, once answers like the ones I have proposed to these frequently asked questions about consciousness have been developed, debated, and widely accepted, then the hard problem of consciousness will no longer be considered any more of a mystery than gravity.