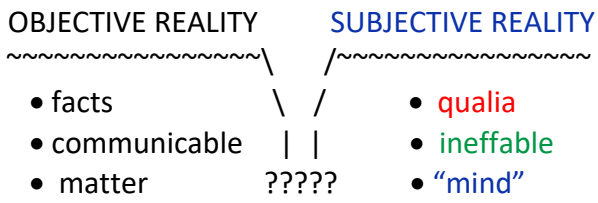
THE EXPLANATORY GAP AND THE "HARD PROBLEM" OF CONSCIOUSNESS

(**Reference:** Blackmore, *Consciousness*, ch. 1)

Objective and Subjective Reality

Suppose you call someone on your phone, and ask what they're aware of right now, concerning their surroundings and themselves. Much as Blackmore does in one paragraph, they might comment on the room or street where they are located, the houses, people, clouds, etc., they can see, and perhaps voices, cars, birds, etc., they can hear; and about themselves, they might mention things like sitting in a chair or standing in a street, and that they are talking on the phone and answering your strange question.

Now, those observations are all ones that someone could in principle verify. They are straight-forward objective facts. But perhaps your respondent also mentions bodily states like feeling chilly or warm, or having a headache, etc. Here it's much more difficult to know what that even *means,* in any objective sense -- they are subjective experiences that you can only understand by analogy, i.e., in terms of similar experiences you may have had. Actually, your respondent probably won't even mention the most compellingly, undeniably real things they are aware of right now, such as the color sensations, brightness sensations, or texture or shape perceptions that they are experiencing, or the sound of their own voice.



While objective reality leaves us with many unanswered questions (think of modern physics, biochemistry, neuroscience, sociology, climate science, exoplanets, medicine, etc.), we have a sense that objective knowledge can be extended bit by bit. By comparison, subjective reality just seems deeply mysterious. How are our undeniably, vividly real, private experiences -- our qualia -- connected to objective reality? That is the ***explanatory gap***.

Focusing on the subjective side, the problem is to explain qualia in a way consistent with our objective understanding of reality. That is David Chalmer's ***hard problem***.

Looking at this neuroscientifically, the problem is to explain how subjectively experienced qualia arise in neural circuits, even though the great majority of neural activity is not experienced at all.

Broad classes of "solutions"

Dualism

What is the classical, still popularly accepted answer? It is ***dualism***, as in most religions (but not Buddhism and Confucianism). Dualism is the idea that matter on the one hand, and mind/spirit on the other, are simply distinct realms of existence.

In René Descartes' version, the point of contact between the two realms is the *pineal gland*, a small gland located between the brain hemispheres. {BTW, it's ironic what the pineal -- pine-cone-shaped -- gland actually does: It produces melatonin, which regulates sleep patterns; so it *is* related to consciousness :}

But a point that Blackmore comes back to in the second chapter is this: If our mind, our spirit, can influence our physical behavior, this implies that our spirit has ***causal*** power. If that is so, then mind-stuff or spirit-stuff is ***physical*** stuff, because the whole point of our theories of matter is to account for ***everything*** that affects physical matter! It's just inconsistent to say that mind/spirit are nonphysical, but can have physical effects!

On the other hand, if we deny that mind/spirit influence physical behavior, then qualia are reduced to ***epiphenomena*** -- they somehow accompany brain processes, but exert no influence at all, even less than the motion of your shadow influences the world as you walk along. As Blackmore says, then why do we even notice qualia at all??

The Big Switch

A second class of "solutions" is to deny that there ***is*** a hard problem; i.e., subjective reality can be explained, and eventually will be explained, in objective terms. I call this way of thinking ***the Big Switch***.

It's a big switch, because as soon as you use objective, neural processing terminology to "explain" subjective experiences, you are no longer addressing the hard problem, but the easy problem -- you're no longer talking about subjective experience at all! Basically, it's just the claim that subjective experience doesn't need an explanation -- it's sufficient to predict the brain's -- and thereby our -- behavior.

There's a deep ethical dimension to this position: If a person's behavior is just the behavior of a bunch of molecules driven by physical laws, why should we be bothered by suffering in the world, or even by the use of torture? The apparent suffering, after all, is just a particular behavior of physical matter, no different from the squirming of a worm on a fisherman's hook, or the wilting of a flower in a drought.

In fact, many movies portray scientists negatively because of the popular perception that scientists see everything in terms of physical objects and processes, with no regard for what human or animal subjects may be subjectively experiencing. [cf., the movie *The Shape of Water*]

Mysterianism

What's the third kind of "solution"? Well, it's not a solution at all, but rather it's the claim (Colin McGinn; more speculatively, Pinker) that the solution is beyond any human capacity for conceptualization and discovery. We may just not be smart enough to understand the nature of phenomenal consciousness, any more than a chimp can understand the nature of, say, its own cardiovascular system.

"Defining" Phenomenal Consciousness

So far, we've pretty much taken for granted that we all understand what it means to be conscious -- we know it when we experience it.

But philosophers and scientists try to "explicate" concepts (making them more precise and explicit) before trying to delve deeper. It's hard to study something if you haven't quite figured out what it is you're trying to study and understand.

The most famous attempt at a definition is Thomas Nagel's, in his provocatively titled 1974 article *"What is it like to be a bat?"* He thinks that bats may well conscious, i.e., there's ***something*** that it's like to be a bat; whereas rocks are not conscious, i.e., it's not like ***anything*** to be a rock.

> **NOTE:** Nagel seems to be saying that the only correct filler for the slot "--------" in the following sentence is "nothing":
>
> *"Being a rock is like --------"*
>
> But in fact there's another correct, albeit trivial filler, namely *"being a rock!"*. Similarly we can answer Nagel's question *"What is it like to be a bat?"* truthfully but trivially with *"It's like being a bat"*. But Nagel obviously doesn't want to allow for such trivial answers.

It seems to me that in using the phrase *"what is it like"* he really means *"what does it **feel** like"* -- what the subjective experience is like. In other words, the questions about the bat and the rock, put more explicitly, are *"What does it **feel** like to be a bat?",* or *"What does it **feel** like to be a rock?"*. And in the latter case, we'd indeed be inclined to answer, *"It doesn't feel like anything -- a rock doesn't have feelings"*. But now we're back to square 1 – saying *"Rocks don't have feelings (or other subjective experiences)"* is just a rephrasing of *"rocks aren't conscious!"*

So it seems that Nagel is really just equating being (phenomenally) conscious with *"having subjective feelings or experiences"*, which is essentially a dictionary gloss. As such it may be reasonable, but it certainly doesn't make the science or philosophy of phenomenal consciousness any easier.

In fact, Nagel thinks an objective understanding is not possible. He asks whether it is possible to know, truly ***know***, what another person, or another creature is feeling or experiencing? He thinks not, and argues that

> *"the objective viewpoint of scientific understanding, when applied to the mind, leaves out something essential. ...One learns and uses mental concepts by* ***being directly acquainted with one's own mind****, whereas any attempt to think more objectively about mentality would abstract away from this fact. It would, of its nature, leave out what it is to be a thinker."* (Wikipedia, my bold-face).

If he is arguing against The Big Switch, then I think he's right about that. Still, one may wonder what in the brain could distinguish "direct acquaintance" from (indirect?) objective knowledge. I will offer some thoughts on that later on.

In her book, Blackmore concludes the part of chapter 1 concerned with defining consciousness by asking whether or not consciousness, or qualia, are "extra ingredients", above and beyond thinking and perceiving, or whether instead they're just an intrinsic and inseparable part of being a perceiving, thinking creature.

I think we can agree with her that positing "extra ingredients" that have no causal role (epiphenomenalism) can be excluded by *Occam's Razor* (*Do not multiply entities unnecessarily*!) While if they do have a causal role, they are by definition physical and thus not "extra" after all.

So are qualia instead an intrinsic and inseparable part of being a perceiving, thinking creature? I believe this is heading in a more plausible direction, but is unsatisfactory as it stands. Here an AI perspective is useful: There *are already* AI agents that think and perceive to some extent. For example, our URCS **ETA** blocks-world agent perceives the blocks on the table and hears what is being said, and to some extent reasons about the relationships among blocks and how they were positioned and moved in the past, and can answer questions about that. But I don't think for a moment it has any subjective experiences. It could be that whether or not consciousness is present in a perceiving, thinking being depends on more detailed, subtle aspects of its perceiving and thinking "equipment", such as the particular kinds of internal organization, representations, and processes that it employs. If we can figure out what those structural and functional aspects are, *we would be able to infer* whether a given being has subjective qualia. In other words, we would have a general method of *predicting* the presence of subjective qualia, even though of course *only the things themselves* actually ***experience*** those qualia. (In a brief paper, I called this general problem of predicting the presence of subjective qualia the ***Pretty Hard Problem*** of consciousness, and later David Chalmers and Scott Aaronson independently came up with the same moniker.)

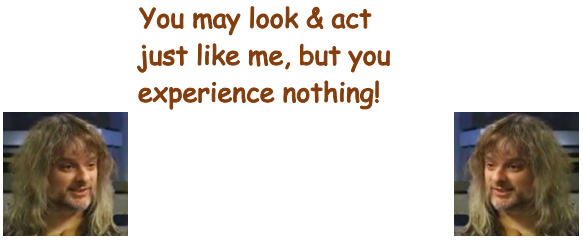
So it seems that to answer the question whether machines could be (or even already are) conscious, we'll need to figure out what those criterial internal structures, representations and processes are! We've already seen roughly how visual perception "works" in our current machines. Is this enough for actual subjective experience in robots -- phenomenal consciousness? Intuitively, the answer seems to be "no", perhaps **(a)** because the visual capabilities are not as sophisticated as ours, or perhaps **(b)** because Nagel's "acquaintance" connection between the

“thinking components” and the visual input stream is missing, or perhaps **(c)** because the robot can't "testify" that it is having subjective experiences. (Note this possible connection to self-awareness -- access consciousness.) As we construct more and more sophisticated robots, including ones whose internal thinking makes intimate and continuous contact with the visual (and other) input streams, and can report on their (putative) perceptual experience and feelings, would we concede that they have phenomenal consciousness -- qualia?

In other words, when we consider beings ranging from simple insects or vacuum-cleaning robots to humans or very human-like robots, is there a transition from being able to perceive and act without consciousness, to having subjective experiences while perceiving and interacting with the world? If so, then conscious qualia are not exactly "something extra", but rather the consequence of "something special" in the internal makeup of the creature or machine. Baars (our next author) tries to get at that "something special", but his answers only provide some general hints. Drew McDermott (also one of our reference authors) argues that being able to model one's perceptual & emotional experiences in a self-model, thus also being able to report on those experiences, is the key to phenomenal consciousness. (However, I hypothesize that the “acquaintance” aspect – an intimate and continuous connection between thinking and perceptual input – is criterial as well.)

If we can figure out what the special equipment or organization or capabilities are that lead to phenomenal consciousness, then we'd be closer to being able to judge to what extent a worm, a bat, a baby, or a given robot is conscious, and perhaps even what kinds of consciousness they have. Mind you, we'd still be left with the question of what these subjective experiences, these qualia, really are, what they really feel like!

Zombies (the "Conceivability Argument" or “Absent Qualia Argument”)



David Chalmers imagines an alternative world where there are physically indistinguishable copies of himself and others, who have no inner life – no subjective experience, no qualia – whatsoever, just behavior. These are Chalmer’s (philosophical) “zombies”. After all, it seems that we can in principle account for our functioning entirely in physical terms, so (he argues) the absence of qualia is conceivable. But since we *do* have subjective experiences, qualia must be something “extra”.

Thinking about (philosophical) zombies is really a way of dramatizing the question of whether consciousness -- qualia -- are something extra, something that a perceiving, thinking being might conceivably ***not*** have. Churchland, Dennett, & Blackmore think this is nonsense ("daft"). For

example, Chalmers could find himself arguing with this doppelgaenger, each claiming to have subjective experiences and denying it to the other. Personally I tend to agree with those who say that a zombie exactly replicating a person physically, ***including in their internal structure and functions***, and behaving identically but lacking qualia, really makes no sense. However, what about …

Partial Zombies

Imagine a deaf person who is wearing special glasses linked to a "speech-and-sound" recognizer and displaying *in text form* what is being said to the wearer, as well as descriptively annotating a speaker's vocal qualities, such as timbre, intonation, loudness, etc.; and we assume that this special equipment also notates other sounds impinging on the wearer, such as street noises, bird song, music (perhaps showing musical notation, etc.). The deaf person wishes to keep their impairment private, and acts exactly like a hearing person, both in conversation and in reacting to ambient sounds. Since we're assuming that this person is totally deaf, not experiencing any *auditory sensations*, we can say that this is an example of an "auditory zombie" – experiencing no auditory qualia, yet acting just like someone who does experience such qualia.

This seems entirely conceivable, and as such provides a more plausible version of Chalmers' thought experiment. So it seems that auditory sensations are "something extra", in the sense that actually experiencing sounds is something more than just extracting relevant information from those sounds and reacting to that information. But it's important to understand that *we're not assuming internal structural and functional equivalence of a hearing person and an auditory zombie*: The auditory zombie is using technical equipment to do the information extraction from sound, and providing it to the wearer in a symbolic visual form; whereas the hearing person does this information extraction via their ear drum, inner ear, auditory nerve, brain stem, and auditory cortex.[1] So the challenge is to try to understand, in general terms, what the difference is between mere "abstract understanding" of perceptual inputs and a combination of "direct experience" of these inputs and abstract understanding of them. (This is related to Nagel's notion of "acquaintance" with perceptual inputs.) That's an issue we'll return to.

Note that we can also conceive of other sorts of partial zombies. For example, we can imagine a visual zombie – a blind person (with good hearing) who also wears special hi-tech sunglasses; but in this case the glasses are equipped with sophisticated computer vision technology that whispers verbal and perhaps other acoustic descriptions of the scene confronting the wearer into the wearer's ears. Though this is well beyond current technology, we can imagine such a visual prosthetic rapidly whispering everything that's relevant to enabling the wearer to act just like a sighted person. This would include colors of (normally) visible entities (cars, shops, trees, etc., etc., if relevant), identities of people, their posture and actions, what they're wearing, where they're looking, etc. Assume that the wearer has become very good a navigating with these glasses (that's perhaps the most difficult aspect, with symbolic auditory input alone), talking

[1] See https://www.nidcd.nih.gov/health/how-do-we-hear;
https://www.brainfacts.org/thinking-sensing-and-behaving/hearing/2020/how-hearing-works-011020

naturally with people, and generally behaving like a sighted person. One can claim that this person is a "visual zombie" – one who acts just like a sighted person, yet has no actual visual experiences -- no color qualia, etc. We could go even further and imagine persons lacking both sight and hearing (like Helen Keller), but with hi-tech prosthetics, perhaps tapping text with high-speed Morse code at points on the head and/or other body parts to convey symbolically what is going on visually and acoustically around the person, again allowing the person to behave much like a sighted and hearing person, but experiencing neither auditory nor visual qualia. This can probably be taken even further – at least, conceivably, eliminating taste and odor perception in favor of prosthetics that signal symbolically about tastes and odors to the impaired individual. Of course, ***some*** sensory connection to the world needs to remain, for information to be conveyed at all, but the point is that we can imagine radical reductions in actual sensory experience, with prosthetics compensating for these deficits through symbolically coded information streams. In this way a person could be enabled to act ***as if*** having sensory experiences of certain sorts without ***actually*** having those experiences.

We can carry over this thought experiment to robots: What if the robot's speech processor or vision system are ***separate modules*** that do sophisticated processing, perhaps using deep neural nets, and transduce the inputs into verbal descriptions (or similar high-level symbolic descriptions)? They then transmit this symbolic information to the robot's "behavioral" modules -- its navigation, thinking, planning, and conversational algorithms. Again suppose the robot behaves perfectly ***as if*** it were actually hearing sounds and had visual experiences, but does so by relying entirely on the ***symbolic*** information it receives from the speech and vision modules. Mind you, it would be difficult for a mobile robot to rely entirely on symbolic inputs (as in the case of the human visual zombie); nevertheless, this is quite conceivable, and arguably has been implemented a number of times, at least in stationary AI systems. In fact, many speech-based AI systems seem to be auditory zombies, as they receive inputs in symbolic form – words – from a speech recognition module, and base their responses on the words alone. Intuitively, such an AI agent would be deceiving us if it claims to hear or see, when all it is actually "experiencing" is a stream of symbols! The burning question then is, what would it take for a robot to ***actually*** have auditory or visual experiences , i.e., auditory or visual qualia?

Along rather different lines, sleepwalkers may be considered partial zombies. According to the Wikipedia article on sleepwalking, sleepwalking occurs during the deepest (slow-wave) stage of sleep, thus presumably without dreaming (not REM sleep). The article does call it a "state of low consciousness" (not an unconscious state), and goes on to say, "Sleepwalkers often have little or no memory of the incident, as their consciousness has altered into a state in which it is harder to recall memories". (There was a gruesome case in Canada, where a sleepwalker by the name of Parks drove 20 miles to his in-laws and bludgeoned his mother-in-law to death and nearly killed his father-in-law as well, all putatively in his sleep. But he did have some vague recall -- he continued on to a police station and said he thought he killed two people. Oddly, in has waking life he had gotten on very well with his in-laws.)

There are additional examples in Blackmore's chapter 2, notably *blindsight* [cf. Peter Watts' fascinating sci fi novel with that title].

Naturalistic Theories of Consciousness

Near the end of Blackmore's chapter 1, she enumerates some of the main naturalistic theories of consciousness (i.e., not relying on a spiritual realm). Let's briefly consider and critique these:

Epiphenomenalism . As discussed already, this seems to posit entities unnecessary for a natural, causal explanation of subjective experience.

Identity theory (functionalism). This identifies states of consciousness with certain physical states and neural system functions in the brain. But as such it relies on *The Big Switch* – it simply "defines away" any distinction between subjective experience and objective understanding of oneself and the world.

Delusionism. This is Blackmore's surprising and highly suspect (but honestly arrived at) conclusion. Since she finds none of the prevailing philosophical theories of consciousness plausible, she concludes that consciousness must be an illusion, somehow constructed by the brain – the brain is fooling itself. My view is that if our reasoning leads us to deny the most compelling and vivid aspects of perception and introspection – the very stuff of our awareness of the world and ourselves – our reasoning must have fallen off the cliff at some juncture. Try telling friends or family that neither you nor they are actually conscious – they're deluding themselves – and they'll assume that you're in need of psychiatric help. It's as if Descartes, instead of saying, *"I think, therefore I am"*, had said *"I am unable to confirm that any of my thoughts are valid, therefore I don't exist".*

I would say a couple of things about that. First, unfortunately Blackmore doesn't separate access consciousness from phenomenal consciousness (the distinction due to philosopher Ned Block). That's a natural omission for theorists who have little familiarity with computers and AI, and in particular are not aware of how easily one can equip an AI agent with some self-awareness by furnishing it with a symbolic self-model and (at least basic) world model. This shows how an agent can perfectly well report verbally on its own properties and relationships without self-delusion (even though self-delusion, via a faulty self-model, is definitely possible). Such an understanding can ground discussions about "self" and the like in a way that pure introspective speculation cannot.

Second, I think it may be possible to explain the intuitions that lead to (untenable) epiphenomenalism in another way, that we might call "**perspectivism**": It's not that subjective experience is somehow "extra", on top of ordinary physical reality. Rather, the subjective and objective perspectives are two perspectives ***on the same reality.*** The objective perspective is by definition restricted to the use of concepts and terminology that can be used for *sharing information* with others. That's a necessity for science! By contrast, the subjective perspective allows not only for communicable concepts and terminology, but also for subjective sensory and emotional experiences that just aren't explicitly sharable. Therefore it's simply a mistake to try to "shoehorn" the subjective experiences themselves into an objective perspective. The

subjective experiences are by no means extra, or lacking in causal consequence. On the contrary, they are what most directly and compellingly animates us. They just don't fit directly into an objective perspective. That's not to say that we couldn't give an objective account of what goes on in the brain when we are experiencing various qualia – that the "easy problem"; but the subjective perspective registers these processes in a different, non-sharable way.

Perhaps a physics analogy could persuade a skeptical reader that the same reality can sometimes be viewed from different perspectives, where one seems to contain entities that the other does not contain. There is, for instance, "wave-particle duality": From a mathematical perspective, the world is described by the Schrödinger wave equation. From that perspective, there are strictly no particles, such as electrons. When you place a double slit in front of an electron source, each "electron" corresponds to a locally peaking wavelet within the "Ψ-function" determined by the global Schrödinger equation. This wavelet has non-zero values at both slits (and indeed everywhere). However, when you place a detector array behind the double slit, an apparent particle is registered at just one detector. Its probability of being registered there is given by the (amplitude squared of) the Schrödinger wavelet. So are there actually particles, or just waves? Well, both perspectives are tenable! If you try to combine the wave and particle perspectives, you get to a strange conundrum: You need to assume the wavelet "collapses" to a particle upon observation (the "Copenhagen interpretation" due to Nils Bohr), but no-one has a convincing account of the collapse mechanism – especially since the wave equation itself, which everyone regards as correct, predicts no such collapse![2]

In the same way, trying to amalgamate the subjective perspective on consciousness into the objective perspective, in the name of being "scientific", is fraught with conundrums. Nonetheless, even keeping those perspectives apart, there remains the mystery of just why the subjective qualia we experience have the particular characteristics they do – e.g., the particular characteristics of perceived colors, sounds, pains, pleasures, emotions, etc.

The "Theater of the Mind" Metaphor

A naïve idea about how we perceive the world, make sense of it, and act, is to imagine a little "self" inside one's head that receives images projected through the eyes onto a kind of theater screen, and interprets those images, and reasons and commandeers your behavior accordingly. This is the "homunculus" that Gilbert Ryle denied, in criticizing Descartes' dualism. Daniel Dennett amplifies the arguments against the homunculus in the theater, saying it leads to an infinite regress.

But while it's easy to see the infinite regress in positing a little "self" -- a "homunculus", the existence of a kind of projection screen is not so far-fetched. Keep in mind that retinal images (including colors, motion, etc.) in fact get represented in the visual cortex at the back of the

[2] Everett's "many worlds" interpretation of wave mechanics provides an alternative interpretation, but it assumes the existence of an infinite number of branching worlds – again entities that don't exist in other interpretations.

brain . In fact Bernard Baars' theory of consciousness is centered entirely around a theater metaphor, as we'll see.

What is Dennett's own view of how consciousness arises? Well, he talks about our conscious awareness as a product of competition among multiple processes, leading to "multiple drafts" interpreting our current sensory inputs, one of which is dominant, and as such conscious, at any given time. But for scientific or AI purposes, we can't learn much from that. Just what are those processes, how do they relate to the knowledge we possess, and what could the "boundary" between conscious and unconscious mental processes be? And what ***is*** subjective experience anyway? Actually, Dennett just rejects subjectivity – but that's the Big Switch Strategy.

Since a "screen" or "stage" are not so far-fetched, the problem with the naïve theater metaphor must lie with the little "self" or "inner self". Blackmore argues that when you say "I" or "me", you are referring to some essential inner self that somehow owns and controls the body. (Eventually though, in Ch.5 of her book, she concludes that the "self" is a fictitious construct.) She says that the notion of an inner self is built into language, because we say "my body", "my head", "my brain", etc., suggesting ownership of the body by a distinct self. But suppose you have a friend who is very committed to physical and mental fitness, and you say, *"**He** keeps **his** body as well as **his** mind very fit".* Are you referring to a "he" that somehow exists separately from, or interior to the friend's body and mind? It seems more plausible to me that you're just referring to the ***person***, as a whole. Would you agree? Then isn't it the same when you say *"**my** body"* and *"**my** mind"?* I think it's important to recognize that the English possessive has multiple meanings, one of which is indeed possession, while another important one is "having as a part". E.g., when I say "*My laptop has an Intel processor",* I don't mean it *owns* it – it's just a *part* of it. In the same way, I feel that when I say *"I have a brain"*, I just mean it's part of me, as whole!