The Hard Problem of Consciousness Made Slightly Easier

ABSTRACT: We can best understand the nature of consciousness through metaphor, by putting ourselves imaginatively in the place of the brain. Neural processes evoke sensation and meaning in the way that words evoke mental images. Conscious experience is (like) a virtual reality produced by the brain, guided by interaction with the external world. Explanation occurs in the field of view of the subject, so to speak. To explain the existence of that field inevitably involves circularity.

Except for sensations within the body, the senses are normally interpreted to give information about the external world. Firing of receptors on the retina of the eye, for example, is not experienced as taking place in the eye, but in the external world. Vibrations in the eardrum are not experienced as sensations in the ear, but as coming from a source at some distance from the body. For good reason, this is how the brain and body are naturally programmed. We simply couldn't exist if it were not so. The world outside the body is a natural object of attention, because it has crucial significance for the body's welfare.

This arrangement poses a problem when we wish to consider abstractions like the *self* or the *subject*—or, in science, the *observer*. It applies also to abstractions like the *world* or the *object.* For one thing, the physical body is part of what we call the external world. There is a world of objects outside the body, yet the body itself is an object in that same world. So, "external" can mean *outside the skin*, but it can also mean *outside the observing seat of awareness*. Physiology imposes a natural separation of subject and object. The subject is always *here* and the object is always *there.* The self or subject is a point of view, not an object to be seen or touched. This fact is central to the 'mind-body problem', more recently called the 'hard problem of consciousness'.

It is a problem because our strategies for understanding are naturally oriented toward the external world. It is "hard" because it cannot be solved in the terms in which questions about the external world are usually posed and answered. We try in vain to explain the seer in terms of the seen, the subject in terms of the object. That is, to explain consciousness—or the mind or the self—in terms of the mechanics of the external world. This strategy is built into our externally-oriented mentality as natural organisms. To the extent there is no way around it, there can be no strictly scientific solution to the problem posed by consciousness—that is, a solution which explains consciousness in terms of chemical or neurological events, for example. Fortunately, there is an alternative strategy, which we will pursue here.

The term 'neurological' has two components. There are *neural* events, such as the chemical discharges of nerve cells, propagated along axons. These can be viewed as events in the physical world that happen through causal processes in space and time. However, these are also *logical* events: something the body *does* in pursuit of its own agenda. They are part of the organism's survival strategy. In that sense, we may think of neurological events as intentional as well as causal. While the organism is a material object, it is also an *agent*. Because it is a physical thing in the natural world, it can be acted upon by other things. Yet, because it is a self-sustaining system, the organism can also act on those parts of the world that constitute its own physical being and on other things outside it as well. Its actions, whether internal or external, can be viewed as logical moves in a sort of game. As such, they do not take place in real space, but in logical space; they are not caused but intended.

Here we will view the organism as an intentional agent. But note that being intentional does not presume being conscious. If it did, we would simply be reasoning in circles. The experience of being conscious is what we hope to explain—not in terms of physical processes, or causes originating in the external world, but in terms of logical processes that originate within the organism. While we cannot formally explain consciousness by circularly invoking the very thing that is to be explained, yet perhaps we can allow ourselves to cheat a little by using metaphor, which inevitably *does* presume our own consciousness as language users.

Imagine you are standing before a blackboard. You wish to illustrate an abstraction: the general relationship between subject and object. You could write symbols on the blackboard, such as *S* for subject and *O* for object and try to devise some theoretical expression for a relationship between these concepts. In such an expression, one symbol may *imply* another, which is a logical relationship between them. But no symbol *causes* another. Rather, each is intended (and written) by you as an agent.

Alternatively, you might draw a picture of a brain, sealed inside the skull. You could diagram how it is connected, via nerve fibers, with the world outside the skull, emphasizing that there is no other way for information to enter or leave the brain. That is, the skull is a “black box,” whose functional content can only be inferred from comparing inputs and outputs. Furthermore, from a point of view within, the world outside the skull is equally a black box. The head is not a room with windows through which occupants can see the outside world, nor with doors through which they can exit to gain experience outside the room. Rather, our task is to explain *seeing* and *experiencing*, and to arrive at a concept of the *world*, without already presuming these. We must do this purely in terms of processes inside the black box. If we invoke the metaphor of a room, it is a room without portals and exits. If we invoke an occupant, it is an occupant who has never been outside the room and has no prior knowledge of a world outside (or that there is even such a thing as “outside”). With those provisos, we are allowed to presume a hypothetical agent who can explore this interior environment and do things within it. Such a hypothetical being is called a homunculus.

Let’s make the metaphor more tangible by likening the sealed chamber to a submarine without portholes, hatch, or periscope. As outside observers, we know that there is an underwater world outside the hull. Our homunculus, however, has no such knowledge to start with. Rather, the task is to gain that knowledge in the only way possible: through trial and error within the vessel. Let’s say that the inside of the submarine comes equipped with what we, as outsiders, recognize as “controls” and “instrument panels”—that is, with potential inputs and outputs. Our submariner can play with levers and switches to try to discover any patterned relationships between those actions and the readings on various gauges and dials. In other words: to see how doing something to the controls might bring about changes in the instrument readings.

In our outsider’s view, what connects these inputs and outputs is the real world outside the hull and the fact that the submarine is a part of that real world, can move through it, and can perform actions upon it (for example, with sonar and robotic arms). In other words, the submarine can affect and be affected by the underwater world. Pulling a lever inside may activate a propeller causing motion through the water; pushing a button might issue a sonar pulse whose echo is registered by an instrument. In that feedback loop, it is the real underwater world that establishes the patterns between input and output, by completing the loop, so that the patterns identified contain information about that world. It is thus possible for our occupant, by such experimenting, to create a model or map of the underwater environment outside the submarine, without ever seeing or touching it directly, simply by actions performed within.

Why would our occupant (homunculus cum submariner) bother to do any of this, apart from idle curiosity? The submarine obviously represents the living organism—in this case, a human body. But, of course, the submarine is *not* an organism but simply a machine. No machine (so far) has a vested interest in its own existence. An organism, by contrast, is *defined* by that vested interest. The organisms that exist have learned how to negotiate their environments and would not exist otherwise. Natural selection is the process of eliminating failures. So, we must imagine a corresponding principle whereby submarines that lack a "realistic" enough model of the underwater world are potentially eliminated. We must imagine a submariner who knows nothing of that principle yet, and simply succeeds or fails to preserve the submarine through experimentation. We must imagine generations of submarines that have adapted (or not) to the underwater world through some equivalent of natural selection.

What makes that environment seem *real* to the submariner is the fact that it *matters.* It has power of life and death over the submarine's existence. Otherwise, our homunculus would be engaged in a meaningless game or fantasy. *Realness* is thus both a property of hypothesized external things and events, but also a quality the homunculus attributes to the model. In short, the model comes to be experienced *as* the reality it represents—as a real external world. To consider something to be real is to believe its power to affect you for better or worse. And to so acknowledge its reality is to *experience* it as real. Thus, for example, we normally know the difference between a rapidly approaching bus and a fantasy or image of the bus, since the one can kill us, the other not.

Yet, if we have the ability to unilaterally imbue experience with realness, then we have some discretion over what to consider real. And, in that case, we can misattribute realness—indeed, project it indiscriminately at whim. This poses a secondary problem for the mariner. The original problem is to identify correlations to be taken seriously as real. The additional problem is doubt: how to know when apparent correlations or patterns are merely spurious.

Together these two challenges point to a general relationship between subject and object. Let us return to our blackboard, where we represented the subject or self by *S* and the object or world by *O.* Let us also presume, as we did for our homunculus, that *S* represents a conscious viewpoint. So, we must introduce a further symbol for the actual moment-to-moment experience *from* this viewpoint: the content of that consciousness, which includes sensation, feeling, and thought. Call it *E.* Then a general expression for the relationship we seek might look like this: $E = f(S,O)$. In plain English: "Experience is a function of both subject and object." We can assert this truism more generally by adding that *behavior* too is a function of both subject and object. In other words, all that we can experience, think, feel or do is jointly shaped by both self and world—always interacting together. Which factor is dominant at a given moment may vary, but neither is ever entirely absent. This may seem an obvious truth, but the normal outward focus of mind makes it easy to overlook. We might be convinced, for example, that some alarming real event *compels* us and *justifies* us to react in a certain way. However, we have some power to decide what is real and to choose a course of action. On the other hand, we might think that some fantasy image is a pure invention with no relation to reality. Yet, even imagination derives ultimately from experience of the real world.

Now that we have a formula for the relation of subject to object, we can visualize that relation more concretely by returning to our submarine metaphor. The challenge for our homunculus cum mariner is to interpret the inputs from instruments as evidence about a real external environment, and from that evidence to "visualize" that environment. In our metaphor, we imagine the homunculus with eyes to see the interior of the submarine and limbs to move about inside the submarine and manipulate controls. But all that is no more than a concession to the metaphor. In relation to the reality outside the submarine, our hypothetical mariner is *blind* and *uncoordinated*. Like for a baby, the task is precisely to learn to see and to navigate the surrounding world.

The submariner thus has a dual focus: the model inside the submarine and the world outside it. In this metaphor, of course, *you* are the submariner, the submarine is your body, the undersea world is the external world you are tasked to perceive and navigate in a way that permits your existence. Attention is naturally and appropriately on that external world. Yet, we also live more immediately in an internal environment of appearances produced by the model. Attention can shift from outside to inside, from world back to model. And that is a useful ability, because the model of the external world is always guesswork; no matter how much confidence we gain, it can never be perfect or certain. So, it pays to be able to "go inside" and re-check the data, re-evaluate the model, question the validity of appearances, which are by nature inferences from sensory input. We recognize that our visualization of the external world cannot be taken for granted as literal truth. Rather, it is a creative effort of interpretation, an act of deliberate hallucination.

We live concurrently in the external world and in an interior one of appearances, represented by *E* in our blackboard formula. Since these appearances are a joint product of the external world and of the interpreting mind, our submariner is in the difficult position of having to arrive at a confident vision of the underwater world, while realizing that overconfidence is dangerous. In particular, it is important to know when to take experience seriously and when not. A specific skill in that regard is to deliberately regard *E* as a product of *S* rather than *O*: to treat an apparent reality as merely a subjective appearance, an apparent truth as merely a belief. In other words, to regard an experience not as something occurring in the watery world but as something occurring within the submarine. While the mariner's breakthrough achievement is to see straight through the hull, as though with x-ray vision, the compensating ability is to return attention to the process inside. These two abilities combined, as summarized in our formula, imply that we live with a chronic uncertainty about what in our experience comes from outside and what comes from inside.

It is easy to confuse these two realms, to confuse the created image with the real thing it is supposed to represent. Having created it, the mariner has direct access to the model or image that emerges through the feedback loop between instruments and controls, but no direct access to the world outside the hull, whose role in the feedback loop is merely inferred. There is no way to directly verify it by stepping outside to see for yourself and compare the image with the reality. (The brain is confined in the skull.) So, it is natural and inevitable to take the internal image for the external reality. The alternative is to question the validity of the image and test it through further rounds of feedback. To do that, the mariner must break out of the hard-earned trance of seeing the model as the reality, to see it once again as simply a model.

Jules Verne captured the 19th-century imagination with *20,000 Leagues Under the Sea*. At that depth, there would be little visibility, rendering navigation by instrument crucial. Modern submarines use sonar. But modern technology has given us an even better metaphor for the nature of consciousness: virtual reality. It has given us a metaphor for the brain itself: the computer. And, it has given us an updated version of the navigator's dilemma of uncertainty: how to tell reality from simulation.

The ideal of simulation is to be so like reality that one cannot tell them apart. Normally, a simulation is a computer program that convincingly imitates a real thing or experience. Our submariner's model is a simulation of the undersea world, achieved through a long learning process which could be computerized. Yet, the model does not *copy* the real thing or situation, to which there is no direct access. Let us imagine instead a simulation that is an original creation, not a copy of something else. Let us also suppose that this original creation is nevertheless guided by an external reality in the same way that the development of the mariner's model is guided by the interaction between controls and instruments: through a feedback loop that includes an allegedly real environment. Then our new metaphor merges with the old one, and we can say that conscious *experience* is a virtual reality created by the brain, yet guided and continually updated through interaction with a real environment outside the skull.

A conventional virtual reality is created for the purpose of entertainment. But the virtual reality created by the brain is a matter of life and death; only if it is considered esthetically, as an *artifact* and not as literally *real*, does it serve as entertainment. On the other hand, a simulation seems real to the degree it is convincing. Of course, a conventional virtual-reality headset can be put on or taken off at will by users, who normally will not forget their identity as human beings who can embrace or leave the experience by disconnecting from the equipment. This was not the case for our submariner, who could not leave the submarine and had never had a life outside it to remember. So, in this new metaphor we must imagine someone who grew up in the simulation, had never lived outside it, and cannot turn it off. Imagine, therefore, a simulation like in the *Matrix*, designed to be so convincing that it completely deceives its captive users. In the film, there is a true reality of passive bodies harvested for the electricity they generate. These poor creatures live in a hallucinated reality where they believe they are actively going about a normal life. By design, this simulation is supposed to be so seamless that you cannot know that you are living in a virtual reality. But the plot of the story requires that there be some way to detect the deception: a "glitch" in the computer code.

There are actually many glitches in the brain's simulation. The science of cognitive psychology is founded on them. The very awareness that there *is* cognitive processing going on, and that the brain somehow *produces* our conscious experience, began with the recognition of perceptual anomalies. These are glitches in normal perception, such as illusions of shape and figure/ground, motion effects, experimental investigations of sensory adaptation, hallucinations, and cognitive illusions such as the rubber hand effect. If normal perception were seamless, we would all be naïve realists who simply believe that the world exists exactly as we see it and that the brain has nothing to do with the appearance of the world.

The skepticism aroused by such observations led early thinkers like Descartes to the dread conclusion that it is possible to falsify experience by hijacking the nervous system. You could be living in a simulation and not know it. It was exactly that suspicion which led to the brain-in-a-vat scenario and the *Matrix* films. Descartes' solution to the problem was to trust that God would not allow such systematic deception. In modern thought, we might instead trust that *nature* would not allow it—if by deception we mean a set of ideas that would lead to our elimination through natural selection. The human ideal, however, is precisely to tamper with and defy such natural restraints and to deliberately explore the possibilities of artificial reality, not to mention self-deception.

The appeal of virtual reality, simulation, artificial intelligence, and perhaps technology generally, is to be able to do what nature does, or what God does, and perhaps do it better. To be able, ourselves, to re-create what was created in the first place, either by natural or by divine power. So, we aspire to create artificial mind, artificial experience, and even artificial life. All of those possibilities blurr the distinction between the natural and the artificial, the genuine and fake, the found and the made. Paradoxically, to *live* in a virtual reality is to believe it real, to ignore or forget that it is fabricated. And that applies as well to the virtual reality naturally produced by the brain: in a day-to-day context, it serves us well to believe the illusion created by the brain. We question it only in circumstances where we suspect it may *not* serve us. In the *Matrix*, the hero is given a choice: take one pill and remain comfortably in the illusion of a normal life; take the other pill and be painfully aware of the actual situation. Which would you choose?

There remains the hardest part of the hard problem: how does neurological activity in the brain become conscious experience? We've already conceded that a causal explanation is inadequate, because causes do not account for the organism's purposive activity as an agent. Quite possibly, computation provides a better basis for an explanation of consciousness than physical or chemical processes. While you can examine the wiring of a computer and explain its functioning on a certain level in physical terms of electrical charges and flows, it is the *logical* organization of the device that makes it a computer and makes it seem to be thinking. It can mimic human thought processes because it was designed by human agents to do so.

No computer yet acts on its own agenda, for its own purposes. Present AI simply mimics some aspects of human agency and autonomy, without agency or true autonomy of its own. It may be convenient to see an insect as a sophisticated tiny robot and possible even to build one. But no machine, so far, is programmed to maintain itself and seek its own well-being in the way that the common house fly is. Natural creatures acquire their agency through natural selection, which means many generations of individuals adapting, who either survive or not according to how well their programming happens to work. They have built into them behaviors that permit them to exist; they may also be able to adapt through learning in the course of a lifetime. Humans, especially, are also able to adapt their environment to themselves. So far, machines do not have these abilities—nor should they ever.

To have conscious experience, a system must be self-maintaining and self-defining. It must have priorities. Events must *matter* to it, and this mattering is the foundation and prerequisite of sentience. In other words, the basis of consciousness is the brain's evaluating response to its sensory input. The input comes from the world, but the response comes from the organism. The neurological activity involved in perceiving and responding *can* be viewed causally—as physical events happening in the part of the world occupied by the organism. However, the behavior and the experience of the creature must be understood in terms of mental actions or connections it performs within itself for its own reasons, which have to do with its own best interests. A brain's program, like that of the submariner, reflects the internal connections it has made that permit its survival.

While that condition is necessary, it still is not sufficient for consciousness, which requires also a specialized inner agent with executive powers. We might think of that agent like the CEO of a corporation—or the captain of a submarine. It is this captain who "sees" the world outside the hull and gives orders for actions to be carried out on behalf of the ship—actions that are not automated. The seeing itself is an elaborate action of interpretation and evaluation carried out inside with the aid of a crew. Consciousness is internal communication about external input, translated into experience in much the way that words evoke mental images. The world *appears* to us in consciousness through connections within the brain and body, intentionally made by the organism. If that seems more like magic than science, it is the same magic we use every day in language.

If this seems hardly an explanation at all, it is because the subject is no object. Language, thought, and metaphor aim at possible objects, which the subject always eludes. "Explanation" concerns something within the field of view, where the subject never is. The world appears in that field of view and we can explain things in the world in terms of other things in the world. To try to explain the field of view itself, in terms of appearances within it, can only lead us in logical circles.