Consciousness and the Interface Theory of Perception

by Donald D. Hoffman, Ph.D.

Based on evolutionary grounds, brain activity does not cause consciousness, and on mathematical grounds, consciousness is not identical to functional properties of the brain. Consciousness is fundamental and must be modeled precisely in its own right.

1. Introduction: The Mystery

"What is the biological basis of consciousness?" This is the version of the classic mind-body problem that is widely assumed in current research.[1] Much of this research is focused on finding neural correlates of consciousness (NCCs); an NCC is a minimal collection of neural events or mechanisms that is highly correlated with a specific conscious experience, such as an itch or a headache.[2] To its credit, recent research has found many NCCs. But to its dismay, it has failed to furnish a theory: It's a mystery how NCCs can be, or cause, or give rise to conscious experiences.

This mystery, in one form or another, has puzzled thinkers since before Plato. In 1866, it puzzled Thomas Huxley, who wrote, "How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of Djin when Aladdin rubbed his lamp."[3] Despite recent progress in unearthing NCCs, this mystery still puzzles researchers such as Christof Koch, who observes, "That is the universe in which we find ourselves, a universe in which particular vibrations of highly organized matter trigger conscious feelings. It seems as magical as rubbing a brass lamp and having a djinn emerge who grants three wishes."[4]

Why is the mind-body problem still a mystery? One answer, indeed the one most widely tendered, is that the key discovery that will solve this mystery has not been made but, when it is, the solution will be obvious.[5] This has happened before in the history of science. The mystery of life and inheritance, for instance, was solved by discovering the structure of DNA. Given the history of science, this reply is fair.

A second answer is that we have been short-changed by evolution. NCCs indeed cause or give rise to consciousness, but we have not been endowed by evolution with the concepts needed to understand how this happens. We don't expect spiders to possess the concepts needed to understand quantum physics; perhaps *Homo sapiens*

don't possess the concepts needed to understand the mind-body problem. This appears to be the view of Colin McGinn: "We know that brains are the de facto causal basis of consciousness, but we have, it seems, no understanding whatever of how this can be so."[6] Given what we know of evolution, this, too, is a reasonable reply.

2. Let's Question Our Assumptions

However, it's also possible that the mystery persists because our formulation of the problem harbors false assumptions. This has happened before in the history of science. For instance, the way that objects called "black bodies" radiate energy mystified classical physics, and was only understood when quantum theory rejected the classical assumption that energy varies continuously, replacing it with the counterintuitive assumption that energy, such as the energy in light and heat, comes packaged in discrete quanta.[7]

The problem with this proposal is that scientific theories, including current attempts at the mind-body problem, make many assumptions.[8] If false assumptions are indeed hindering progress, then it might be difficult to discern which are the offenders.

But it's worth a try. Here I question two assumptions of many current theories. (1) Natural selection favors true perceptions. (2) The mind is what the brain does.

Why these? One reason is that these assumptions are central. We assume, for instance, that there is a biological basis for consciousness in part because we assume that our perceptions of space, time, and physical objects (such as brains and neurons) are generally true. If it turned out instead that our perceptions of space, time, and physical objects are adaptive fictions, not genuine insights, then we would be less inclined to assume that some of those fictions, namely neurons, are the basis of consciousness.

A second reason is that both assumptions can be rigorously tested. The first can be tested using, for example, genetic algorithms and evolutionary games.[9–12] The second can be formulated as a mathematical proposition, and proven true or false.[13,14]

A third reason, which no doubt you've already guessed, is that I think both assumptions are, in fact, false. For the first assumption, a variety of evolutionary games and genetic algorithms demonstrate that true perceptions are dominated by simple heuristics that are tuned to fitness.[9–12] For the second assumption, a theorem

establishes that conscious experiences cannot be identical to functional properties of a complex system such as the brain. [13,14]

In what follows, I outline the evidence against the assumptions that natural selection favors true perceptions and that the mind is what the brain does, keeping mathematical discourse to a minimum. I then propose an "interface theory" of perception, [15,18] and a "conscious realist" ontology in which consciousness, rather than space-time and physical objects, is taken as fundamental. [19,20] These new assumptions transform the mind-body problem. Rather than being a puzzle about how matter gives rise to consciousness, it becomes the problem of how consciousness gives rise to space-time and matter. A scientific theory that starts with consciousness requires, of course, a mathematically precise theory of consciousness. I propose some ideas in that direction, aiming for a genuine theory that makes risky and testable predictions.

There is a simple way to dismiss the project just outlined: If natural selection favors untrue perceptions, then surely it favors untrue logic and math. If so, then this project refutes itself. It uses logic and math to conclude that logic and math are unreliable.

This would be a showstopper. I think, however, that the same evolutionary games that reveal selection pressures against true perception also reveal selection pressures toward reliable logic and math. This is an open issue, but I sketch reasons to be hopeful, similar in flavor to Dutch book arguments for the axioms of probability. [21]

3. Some Common Intuitions About Selection and Perception

Does natural selection favor true perceptions? Many vision researchers claim that it does. In his textbook *Vision Science*, Stephen Palmer tells the reader that "Evolutionarily speaking, visual perception is useful only if it is reasonably accurate…. Indeed, vision is useful precisely because it is so accurate. By and large, *what you see is what you get.* When this is true, we have what is called **veridical perception** . . . perception that is consistent with the actual state of affairs in the environment. This is almost always the case with vision…." [22]

Noë and Regan argue that "perceivers are right to take themselves to have access to environmental detail and to learn that the environment is detailed" and that "the environmental detail is present, lodged, as it is, right there before individuals and that they

therefore have access to that detail by the mere movement of their eyes or bodies." [23]

Geisler and Diehl suggest, "In general, it is true that much of human perception is veridical under natural conditions." [24]

Marr claims, "We…very definitely do compute explicit properties of the real visible surfaces out there, and one interesting aspect of the evolution of visual systems is the gradual movement toward the difficult task of representing progressively more objective aspects of the visual world." [25]

Physicists and philosophers have also weighed in on this issue. The physicist Abner Shimony, for instance, argues that "evolution has eventuated in animals which transform their sensitive reactions so that their resulting cognitive states are quite accurate indices of crucial distal characteristics of the environment." [26]
The philosopher Thomas Nagel argues, "If there is a mind-independent physical world, the systematic inability to detect the basic truth about our surroundings (setting aside more sophisticated scientific truth) would be disastrous for our reproductive fitness. Realism about the physical world is a fundamental aspect of any Darwinian explanation of our perceptual and cognitive faculties, as well as of our motives and capacities for action." [27]

The intuition behind these claims seems to be that truer perceptions are ipso facto more fit. In consequence, those of our predecessors who saw more truly had a fitness advantage over those who saw less truly, and were more likely to have offspring. We are the descendants of those who saw more truly, and thus can count on our perceptions to be generally accurate.

The problem with relying on this intuition is that evolution is complex, and intuitions are fallible guides to its workings. Fortunately, there are mathematical formulations of evolution, such as evolutionary game theory and genetic algorithms, which permit us to rigorously investigate whether natural selection indeed favors truer perceptions. [28,29]

3. What Is a Perceptual Strategy?

We want to carefully investigate what evolution entails about perception. What kinds of perception does evolution favor? What kinds are likely to go extinct? To try to answer these questions, we must have a clear idea about what we mean by "kinds of perception." We need to precisely define different kinds of

perception, so that we can see precisely what evolution will do with them.

There is a long and interesting history of philosophical debate about the nature of perception, which is a good source to draw on here.[30,31] But we must transform this debate into precisely defined *perceptual strategies* that we can then allow to compete in evolutionary games. These perceptual strategies might, in turn, aid philosophical debates by providing a precise language for discourse.

So we start by first specifying the classes of perceptual strategies to be tested in our evolutionary simulations. If we denote the objective world by some set *W* and the perceptions of an organism by some set *X*—where we assume for the moment that we know nothing about *W* or *X*—then a *perceptual strategy* is a function, call it *P*, from *W* to *X*. This is illustrated in Figure 1, and written $P : W \to X$.

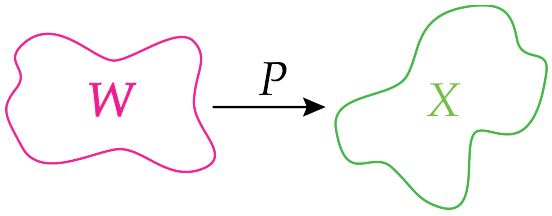

Figure 1. Perceptual strategy. The region labeled *W* represents possible states of the objective world. The region labeled *X* represents possible perceptual states of an organism. A perceptual strategy is a way of mapping the states of the world onto the perceptions of an organism, and is labeled *P*.

We can distinguish different classes of perceptual strategies by their assumptions about *W*, *X,* and *P*. The strongest assumption, which we call *naïve realism*, claims that our perceptions are identical to the world; that is, *X* = *W* and *P* is the identity function.

A weaker assumption, which we call *strong critical realism*, claims that our perceptions are identical to a subset of the world; that is, $X \subset W$ and *P* is the identity function on this subset.

A yet weaker assumption, which we call *weak critical realism*, claims that our perceptions need not be identical to any subset of the world, but that the relationships among our perceptions accurately reflect relationships in the world; that is, it is allowed that $X \not\subset W$, but required that *P* is a so-called homomorphism of the structures on *W*.

A yet weaker assumption, which we call *interface perceptions*, claims that our perceptions need not be identical to any subset of the world, and that the relationships among our perceptions need not reflect relationships in the world except measurable relationships (i.e., relationships needed to describe probabilities); that is, it is allowed that $X \not\subset W$, and that *P* is not a homomorphism of the structures on *W* (except for measurable structures).

Finally, the weakest assumption, which we call *arbitrary perceptions*, claims that our perceptions need not be identical to any subset of the world, and that the relationships among our perceptions need not reflect any relationships in the world; that is, it is allowed that $X \not\subset W$, and that *P* is not a homomorphism of any structures on *W*. The relationship among these classes of perceptual strategies is illustrated by the diagram in Figure 2.

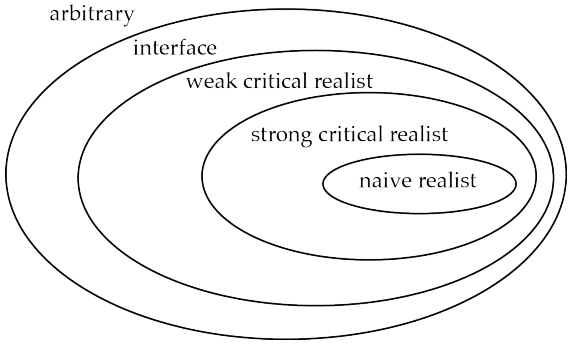

Figure 2. Venn diagram showing the inclusion relationship among the five classes of perceptual strategies.

4. Could Our Perceptions Be Like a User Interface?

Few researchers are naïve realists, because there is reason to believe that we do not perceive some aspects of the objective world. The human eye, for instance, only sees light whose wavelengths lie in a window between 400 and 700 nanometers, whereas the electromagnetic spectrum extends well beyond this window. Despite such evidence, some philosophers still defend versions of naïve realism. [30]

Many researchers are strong critical realists, and assume that our perceptions are, in the normal case, identical to a *part* of the objective world. According to them, when you see a round, white

baseball, there really is a round, white baseball that exists even if you don't look.

Some researchers are weak critical realists, and assume that our perceptions need not be identical to *any part* of the objective world, but that they do accurately portray its true structure. Colors, for instance, might not exist apart from our perceptions, but colors nonetheless accurately convey aspects of the world that exist even when we don't look.

Few researchers buy the interface theory, which allows that our perceptions are not identical to any part of the objective world, and that they do not accurately portray its true structure (apart from the structure needed to describe probabilities).[15–17] Indeed, when I introduce the interface theory in lectures at universities and conferences, the audience finds it amusing, obviously false, and almost beneath dignifying with a response. After all, they argue, if our perceptions need not be accurate, even about a part of the objective world, then how could they be useful? Illusions would not be the exception, they would be the rule.

But an analogy often helps. Consider the desktop of your laptop or mobile device. Suppose that there is a file icon on the desktop that is round, blue, and in the middle of the screen. Does that mean that the file itself is round, blue, and in the middle of the computer? Obviously not. Files have no colors or shapes, and their positions on the screen needn't mirror their locations in the computer. The colors, shapes, and positions of an icon are not true depictions of the objective properties of the corresponding file. Nor are they intended to be. It's not that the interface is trying to deceive you. It's simply that its purpose is not to depict objective reality, but rather to hide it. The reality is too complex, and understanding it is not necessary if one wants to delete a file or edit a photo. Indeed, if you were forced to deal explicitly with all the diodes, resistors, voltages, and magnetic fields that constitute the file, you might never finish editing that photo.

So here is a case where accurately perceiving the objective truth is not useful, it's an impediment. The interface theory of perception allows that natural selection might have shaped our perceptions to be analogous to interfaces that hide the complexity of objective reality and instead provide a useful guide to behavior. If so, then space-time could simply be our desktop, and physical objects with their colors, shapes, textures, and motions are just icons of that desktop.

5. Evolutionary Games: A Matter of Life or Death

There is a long history of philosophical debate about the nature of perception, [31] and recently this debate has included arguments from evolution. [32–36] Remarkably, until recently, no one formalized these arguments and tested them using evolutionary games. When this is done, interface perceptual strategies are typically more fit than realist strategies. [9–12]

Consider, for instance, a game in which two animals compete to obtain a resource, say water, that is in three distinct territories, as illustrated in Figure 3. An animal looks at each territory and chooses one, obtaining its resources and the corresponding fitness payoff. Once a territory is chosen, the other animal must choose one of the two remaining territories, and obtains its resources and its fitness payoff. On each trial of the game, we can randomly select which animal chooses first.

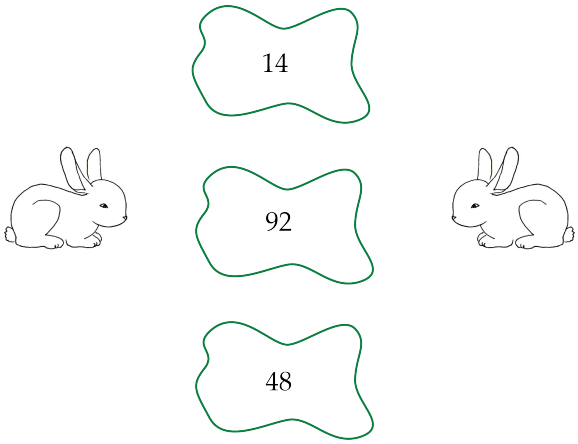

Figure 3. An evolutionary game. Two organisms (e.g., rabbits) compete for water resources in three territories. In this example, the quantity of water happens to be 14 in the first territory, 92 in the second, and 48 in the third. These quantities are not fitness payoffs. It might be, for instance, that the fitness payoff for 92 is less than for 48.

The quantity of water in a territory might vary, say, from 1 to 100, where 1 indicates little water and 100 indicates a lot. In different games, we can play with the statistics of water quantity, perhaps using a uniform distribution, a normal distribution, or some other distribution of interest.

In different games, we can also play with the fitness payoffs associated to different resource quantities. We could, for instance, consider games in which greater resources yield greater fitness

payoffs. But we could also consider games in which, say, resource values nearer 50 have higher fitness payoffs. This could model a case where too little water is bad for fitness (e.g., dying from thirst), too much water is bad for fitness (e.g., dying from drowning), and some intermediate quantity is just right.

Once we have the distribution of resources and the fitness payoff function, we can then compute the expected payoffs that different perceptual strategies would obtain when competing with each other. For instance, if some animals use an interface strategy (IS) and others a weak critical realist strategy (WS), we can compute the expected payoff to an IS animal when it competes with a WS, the expected payoff to an IS animal when it competes with another IS, the expected payoff to a WS when competing with an IS, and the expected payoff to a WS when competing with another WS. There are 2 x 2 = 4 such expected payoffs to compute. If there is a third strategy, say some animals use a strong critical realist strategy (SR), then we can compute the 3 x 3 = 9 different expected payoffs; if there is a fourth strategy, then there are 16 such payoffs, and so on.

Given these expected payoffs, there are formal models of evolution that we can use to predict which strategies will dominate, coexist, or go extinct.[37–40] We can, for instance, use evolutionary game theory, which assumes infinite populations of competing strategies with complete mixing, in which the fitness of a strategy varies with its relative frequency in the population. In the case in which just two strategies, say S_1 and S_2, are competing, we can write down the four expected payoffs in a simple table, as shown in Figure 4. The expected payoff to S_1 is *a* when competing with S_1 and *b* when competing with S_2; the expected payoff to S_2 is *c* when competing with S_1 and *d* when competing with S_2.

	S_1	S_2
S_1	a	b
S_2	c	d

Figure 4. Expected payoffs in a competition between two strategies, S_1 and S_2.

Then it can be shown that S_1 dominates (i.e., drives S_2 to extinction) if $a > c$ and $b > d$; S_2 dominates if $a < c$ and $b < d$; they are bistable if $a > c$ and $b < d$; they coexist if $a < c$ and $b > d$; they are neutral if $a = c$ and $b = d$. Similar results can be obtained when more strategies compete, but new outcomes are possible. For instance, with three strategies, it might be that S_1 dominates S_2, S_2 dominates

S_3, and S_3 dominates S_1, as in the popular children's game of Rock-Paper-Scissors in which rock beats scissors, which beats paper, which beats rock. With four or more strategies, the dynamics can have more complex behaviors known as limit cycles and chaotic attractors.

Amy Knupp 8/5/2013 5:41 PM
Comment [1]: Note to formatter: I can't figure out how to get this comma to appear on the same line as the S with the 3 subscript

In a large series of evolutionary games, realist and interface perceptual strategies have been allowed to compete. The result is that interface strategies, in most cases, drive realist strategies to extinction.[9–11] One key reason is illustrated in Figure 5.

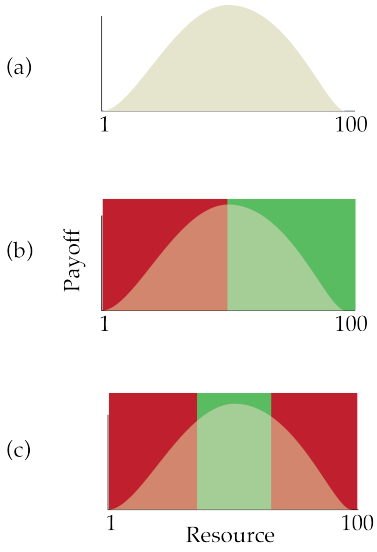

Figure 5. One reason true perceptions go extinct. (a) A resource that varies in quantity from 1 to 100, and a fitness payoff function that rewards intermediate quantities. (b) A realist perceptual strategy that sees resource quantities 1 to 50 as red, and 51 to 100 as green; it is a realist strategy because green truly indicates greater resource quantities than does red. (c) An interface perceptual strategy where green does not truly indicate greater resource quantities than red but does indicate greater fitness payoffs.

Figure 5a illustrates a fitness payoff function, in which the payoff varies as a resource varies in quantity from 1 to 100. The payoff is greatest for intermediate quantities of the resource. Figure 5b illustrates a realist perceptual strategy that can only see two colors, red and green. It is a realist strategy because the perceived colors accurately report information about the true resource quantity: all resource quantities seen as green are greater than those seen as red. Figure 5c illustrates an interface perceptual strategy that also sees only red and green. It is not a realist strategy because the

perceived colors do not accurately report information about the true resource quantity: all resource quantities seen as green are not greater than those seen as red. However, all resource quantities seen as green do have greater *fitness payoffs* than those seen as red. In consequence, when the interface strategy competes with the realist strategy in evolutionary games, the interface strategy will systematically reap greater fitness payoffs and drive the realist strategy to extinction.
The key point of this example is that fitness and truth are distinct. A perceptual strategy that is tuned to fitness will, in general, outcompete one that is tuned to truth. Truer perceptions are not, in general, fitter perceptions, and evolution "cares" only about fitness, not truth.

6. The Mind *Is Not* What the Brain Does

Is it possible that the colors you see are quite different from the colors I see? This question occurs to many imaginative kids, and it occurred to John Locke who, in his 1690 *Essay Concerning Human Understanding*, asked if it's possible that "the idea that a violet produced in one man's mind by his eyes were the same that a marigold produced in another man's, and vice versa." This so-called spectrum-inversion question continues to be debated, because the fate of many theories of consciousness turns on its outcome. These theories propose that it is the *functional* properties of complex systems, such as brains, that are responsible for the presence and properties of consciousness. [42–47]

These theories come in two classes: reductive functionalism and nonreductive functionalism. Reductive functionalist theories pick out some particular functional properties of, say, the brain, and propose that conscious experience is *identical* to those functional properties, where they mean identical in the same sense that 12 and a dozen are identical: they're just different names for one and the same thing. Nonreductive functionalist theories also pick out some particular functional properties of, say, the brain, and propose that those functional properties *cause* or *give rise to* conscious experience.

Functionalist proposals that are nonreductive incur a promissory note: They owe us a theory that explains how and why the particular functional properties that they specify can cause, or give rise to, conscious experience. This promissory note has never yet been paid by any nonreductive functionalist theory, and the relevant bank accounts look pretty empty.

Functionalist proposals that are reductive incur no such promissory note: they owe us no causal or emergence theories because they make no claims about cause or emergence. Their claim is one of *identity*: "The mind *is* what the brain does" is the informal and popular statement of this claim. Now, of course, such a claim is intended to be a scientific hypothesis, not mere armchair speculation, and so it must, in principle, be falsifiable.

How could it be falsified? One approach is to use imagination. If a reductive functionalist proposes that some functional property *F* of neural activity is identical to our conscious experience of, say, a particular shade of red, then one can try to imagine the experience of red happening when *F* does not occur, and vice versa. If one can imagine this, then it is logically possible, and the identity claim fails. If, for instance, one could imagine a triangle that didn't have exactly three sides, this would falsify the claim that triangles are identical to three-sided polygons. (Good luck trying!)

The problem with this approach is that it is not conclusive. If a theory proposes that *F* is identical to some conscious experience, and someone claims that they can imagine otherwise, then a supporter of the *F* theory can simply reply that their opponent didn't really succeed in imagining what they claimed to imagine. This leads to fruitless debates about intuitions.

There is a better approach. One can formulate reductive functionalism as a specific mathematical claim and then try to disprove it. Then one can have profitable debates about the assumptions made by the mathematics and about the correctness of the disproof.

Reductive functionalism has been mathematically formulated and disproven.[13,14] The disproof is called the *scrambling theorem.* Each reductive functionalist theory of the mind-body problem is therefore false. This is not the place to give mathematical details of the scrambling theorem, but a simple example can convey the key ideas.

Suppose, for simplicity, that Jack and Jill each have only two color experiences, say *red* and *green*; and that they each can say only two words, *red* and *green;* and that they each only look at two objects, ripe tomatoes and ripe limes. Every time we show Jack and Jill a ripe tomato and ask them what color it is, they each say "red"; every time we show them a ripe lime, they each say "green." Thus, there are functional mappings that relate tomatoes and limes to the

conscious experiences *red* and *green* and to the verbal reports “red” and “green.”

Now consider a reductive functionalist claim that the experiences *red* and *green* are identical to these functional mappings. This would entail, for instance, that whenever Jack is shown a tomato and says “red,” he necessarily has the same color experience that Jill has when she is shown a tomato and says “red.”

But could Jack and Jill be functionally identical, and yet have different color experiences? Indeed they could, as illustrated in Figure 6. Here Jack and Jill each see tomatoes and limes and, in consequence, have color experiences *red* and *green* and give verbal reports “red” and “green.” However, as the straight arrows in the middle indicate, Jack’s color experiences are not identical to Jill’s, but instead they are swapped. When Jack, for instance, sees a tomato, he has the color experience *red*, but when Jill sees a tomato, she has the color experience *green*. Nevertheless, Jack and Jill each report that the tomato is “red” and the lime is “green.” They are functionally identical, even though their color experiences are inverted.

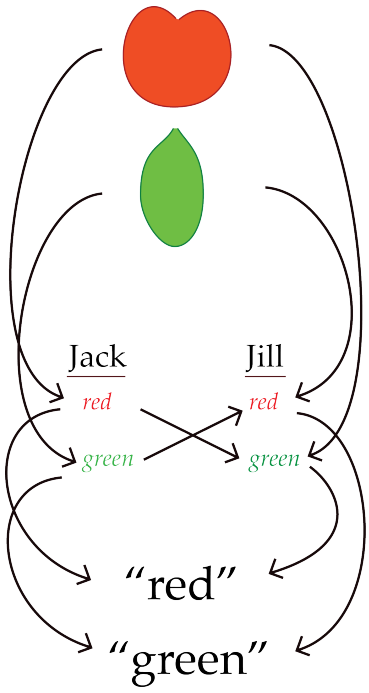

Figure 6. How Jack and Jill can have differing color experiences and yet be functionally identical.

This is a simple example, but the scrambling theorem proves that no matter how complex the example gets, no matter how many conscious experiences are involved, and no matter how much scrambling there is between the conscious experiences of Jack and Jill, it is always possible to arrange the arrows so that they are functionally identical in every experiment that could be performed, including any psychophysical, brain imaging, and neural recording experiments. The scrambling theorem holds regardless of the geometry or symmetries of the space of conscious experiences, contrary to prior proposals.[48]

The scrambling theorem applies to a theory of consciousness called integrated information theory (IIT), developed by Giulio Tononi and Gerald Edelman.[47,49–53] One intuition driving IIT is that each conscious state is highly informative, in the sense that it is but one of a large repertoire of potential conscious states. A second intuition is that, "Phenomenologically, every experience is an integrated whole, one that means what it means by virtue of being one, and that is experienced from a single point of view. For instance, the experience of a red square cannot be decomposed into the separate experience of red and the separate experience of a square."[47]

These intuitions are formalized in a definition of *integrated information*, denoted Φ, which quantifies the amount of information a system generates as a whole beyond what is generated independently by its minimal parts. Specific qualia (i.e., specific conscious experiences) are represented by particular shapes in an information-theoretic qualia space denoted *Q*. They then propose, "According to the IIT, consciousness is one and the same thing as integrated information."[47]

This is a reductive functionalist proposal. They propose to *identify* consciousness with the functional property Φ together with the structure of *Q*. This proposal contradicts the scrambling theorem and is thus false.

This does not mean that Φ and *Q* are useless in the study of consciousness. To the contrary, once one drops the false claim that Φ and *Q* are *identical* to consciousness, one can then explore the interesting empirical claim that Φ and Q *correlate well* with the amounts and kinds of consciousness in a variety of systems. If they do, then one can try to develop a scientific theory of consciousness that explains why and how this is so. In the process, one must account for empirical phenomena that appear to violate the claim that consciousness is an integrated whole. For instance, in some experiments, observers demonstrate illusory conjunctions, in which

they incorrectly bind visual features such as color and form in their conscious experiences.[54,55] In other experiments, they exhibit change blindness, in which observers fail to integrate into their conscious experience visual features that are right before their eyes.[56] Although casual examination of conscious phenomenology suggests that it is an integrated whole, perhaps this reveals little about the true nature of consciousness and more about our inability to be aware of our own blindnesses. These are the kinds of empirical and theoretical challenges that Φ and *Q* face once we give up the false claim that they are identical to consciousness and begin the serious work of building a genuine theory.

So IIT, properly understood, proposes *correlates* of consciousness but offers no *explanatory theory* of consciousness. The mystery that puzzled Huxley in 1866 is no less puzzling to IIT today.

9. Let's Abandon False Assumptions

So far, we have questioned two key assumptions of most current attempts to construct a scientific theory of consciousness. The first assumption, that natural selection favors true perceptions, finds little support in empirical studies using evolutionary games and genetic algorithms. The second assumption, that the mind is what the brain does, is provably false.

It's not easy to abandon the first assumption, to let go of a realist interpretation of our perceptual experiences and instead adopt an interface interpretation. As Thomas Nagel put it, "[S]cientific realism would be undermined if we abandoned a realistic interpretation of the perceptual experiences on which science is based."[27] In particular, scientific realism about neurons and neural activity would be undermined, and this, in turn, would undermine the quest for a biological basis for consciousness. More generally, scientific realism about space-time and physical objects would be undermined, and this, in turn, would undermine the quest for a physicalist theory of consciousness. Not a happy idea for most current researchers. But evolutionary game theory, applied to perceptual evolution, sends a clear message. It tells us not to reify our perceptions.

It's also not easy to abandon the second assumption. The mind-body problem is so mysterious that claims of identity between consciousness and brain function seem to be the only way out. But the scrambling theorem clearly shows that such identity claims are simply giving up, throwing in the towel. It tells us that we cannot shirk the job of developing a scientific theory, by instead trying to pawn off a claim of identity. It tells us not to reify our descriptions.

Since most current research is predicated on the two assumptions we've just rejected, the obvious question is: How shall we now proceed in our quest for a scientific theory of consciousness? What different assumptions shall we try?

At points like this, there are no formulas for how to proceed. Even principles like Occam's Razor are fallible guides (I once heard Francis Crick, at a meeting of the Helmholtz Club, wryly remark, "Many men have slit their throats with Occam's Razor.") These are points of creativity, of revolution, of risk. We strike out in a direction, knowing full well we are likely to be wrong.

10. Let's Assume That Consciousness Is Fundamental

It's in this spirit that I suggest we try to develop a scientific theory of consciousness that takes consciousness as fundamental, not as derivative on neural activity or functional complexity. I call this approach *conscious realism*. Abandoning a physicalist ontology is, of course, not ipso facto renouncing scientific methodology. To the contrary, it is scientific methodology, and the spectacular failure of physicalist theories, that prompts the proposal of conscious realism.

A conscious realist theory of consciousness owes us a mathematically precise theory of consciousness *qua* consciousness. What structures and dynamics does consciousness *itself* have? How are these related to the structures and dynamics in well-established theories of physics such as relativity and quantum theory? For ideas and constraints on such a theory of consciousness, we can consider, inter alia, NCCs, psychophysical experiments, brain imaging, and mathematical correlates such as Φ and Q, but our goal is not a theory of reduction or emergence. It is a mathematical theory of consciousness on its own terms.

Conscious realism is not the transcendental idealism of Kant. For Kant, the noumenal world, the thing in itself, was beyond description and, thus, beyond the ken of science. Conscious realism proposes that consciousness is the thing in itself and is within the purview of scientific.

The goal of conscious realism differs from the de facto prior history of subjective and objective idealism, which have never produced a mathematically precise scientific theory of consciousness (and indeed have sometimes been promoted as adversarial to science). The goal of conscious realism is a rigorous and falsifiable theory of consciousness, that takes consciousness as fundamental but makes

full contact with current theories in physics (i.e., explains how these theories fit within the framework of conscious realism).

There has been some progress toward a mathematical theory of consciousness qua consciousness, and of integrating this theory with quantum theory.[10,12,19,20] This is not the place for mathematical details, but the flavor of the approach can be appreciated if one knows just a bit about so-called Markovian kernels, as illustrated in Figure 7.

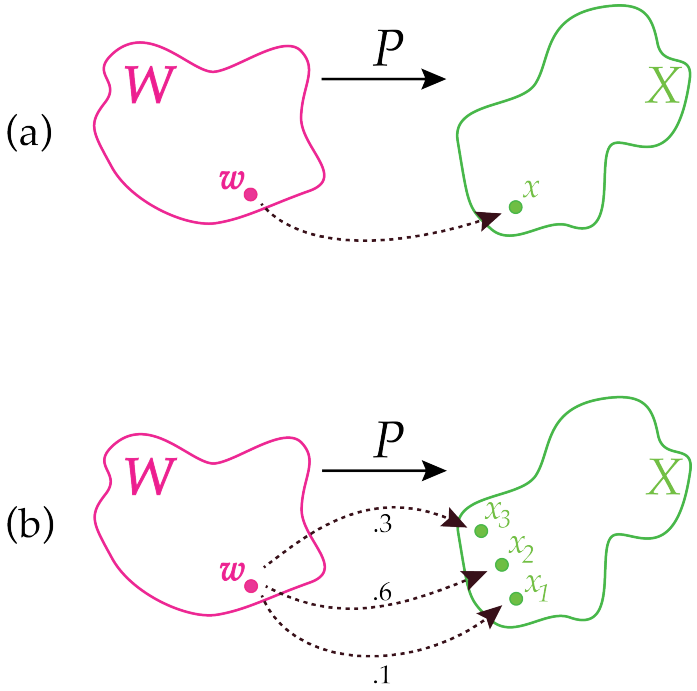

Figure 7. Markovian kernels. In (a) is shown a perceptual strategy that is a function. In this case, a given world state *w* triggers only *one* perceptual state *x*. In (b) is shown a perceptual strategy that is a Markovian kernel. A given world state *w* can, in each act of observation, trigger one of *several* perceptual states x_1, x_2, x_3 and so on. In this example, the probability of triggering perceptual state x_1 is .1, the probability of triggering x_2 is .6, and the probability of x_3 is .3.

In section 3, we defined a perceptual strategy to be a function $P : W \to X$, where *W* denotes possible states of the objective world (whatever it might be) and *X* is a set of possible perceptual states of some organism. This models the situation where a specific state of the world, say state *w*, triggers a specific perceptual response, say

x_1. But what if things aren't so simple? What if sometimes *w* triggers x_1, but other times it instead triggers x_2 or x_3? In this case, we can no longer use a function to describe the perceptual strategy. But all is not lost. We can use probabilities instead. We can say that if *w* obtains then the probability that we will see x_1 is such and such, the probability that we will see x_2 is such and such, and so on for all the relevant possible perceptions. This is similar to saying that if we roll a fair die, then the probability of rolling a 1 is such and such, the probability of rolling a 2 is such and such, and so on. But if the die is not fair, then we will need to assign different probabilities for these outcomes. Thus, for each different state of the world, we get a different set of probabilities for the perceptions that might be triggered by that state of the world. The mathematical object that does this, that for each possible state of the world gives the probabilities of the various possible perceptions, is called a *Markovian kernel*. For each state *w* of the world, it gives a probability distribution on the possible perceptions that might occur.

Now that we know a bit about Markovian kernels, we can use them not just to describe perceptions but also decisions and actions. Suppose that an organism has a repertoire of possible behaviors, say *G*. A particular action g_i might be, say, to take one step forward, another action g_j might be to turn 90 degrees to the right, and so on. Then we can think of a decision as choosing a behavior based on one's current perceptions and goals. If my current perceptions are x_i, then I might, with a certain probability, choose behavior g_j or g_k, and so on. Thus, we can model decisions by a Markovian kernel, call it *D*, that describes for each of our possible perceptions the probabilities of various behaviors we might choose to perform.

Once we have decided on a behavior, we then act on the world and change the state of the world. If we act using behavior g_i, then we can assume that there is some probability that the new state of the world *W* will state w_j, another probability that the new state will be w_k, and so on. Thus, once again we can use a Markovian kernel to model our actions on the world.

We can represent these ideas in a simple diagram, the PDA (Perception-Decision-Action) loop, as shown in Figure 8. The perceptual kernel *P* maps the world *W* to the organism's perceptions *X*; the decision kernel *D* maps the organism's perceptions to behaviors; the action kernel *A* maps the organism's behaviors onto changed states of the world.

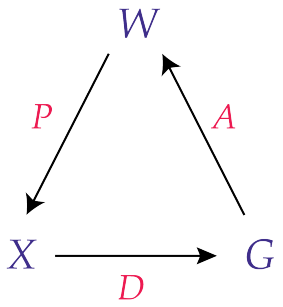

Figure 8. The Perception-Decision-Action (PDA) loop.

We can then, as a first step toward a mathematically precise theory of consciousness qua consciousness, propose a definition of the technical term *conscious agent*. A *conscious agent* is a 5-tuple (*X*, *G*, *P*, *D*, *A*), where *X* is a set of perceptions, *G* a set of behaviors, and *P*, *D*, and *A* are Markovian kernels as shown in Figure 8.

This definition of conscious agent is not intended to be a reductive functionalist theory of consciousness. To the contrary, the term conscious agent is here treated as a *technical* term that has a precise mathematical definition. It is then an empirical question as to how well conscious agents perform as a descriptive and predictive model of consciousness. If empirical research turns up shortcomings of conscious agents, we can revise the definition or abandon it altogether in favor of a better theory.

11. Conscious Agents Are a Promising Model of Consciousness

But of course, I propose this definition of conscious agents because I think it might do well as a theory of consciousness. There are several reasons why.

Conscious Agents and Bayesian Perception
First, researchers have had striking success in modeling perception, multimodal integration, and perceptually guided behavior as so-called Bayesian inference.[57–59] Consider, for instance, the conscious experience of apparent motion, which you can see here: Sphere Applet. The applet shows you a sequence of movie frames in which dots appear in each frame. If the frames are shown slowly enough, then you see discrete frames with dots that are unrelated from one frame to the next. But if the frames are shown more quickly, your conscious experience suddenly transforms. You see dots moving smoothly, as you can check for yourself in the applet.

This transformation of conscious experience is a remarkable feat. Figure 9 shows why. For simplicity, let's just consider a movie in which each frame has only two dots, and let's just focus on two successive frames of this movie. Figure 9a shows this situation, in which the two dots in the first frame are colored black, and in the second frame, red. Now if we are to experience smooth movement of the dots, then the visual system must decide either to move the dots as shown in Figure 9b or as in Figure 9c. This is known as the "correspondence problem," deciding for each dot in one frame where it moves to in the next frame. If there are many dots, then there are many possible correspondences. But we only see one correspondence from frame to frame, and, thus, one smooth motion.

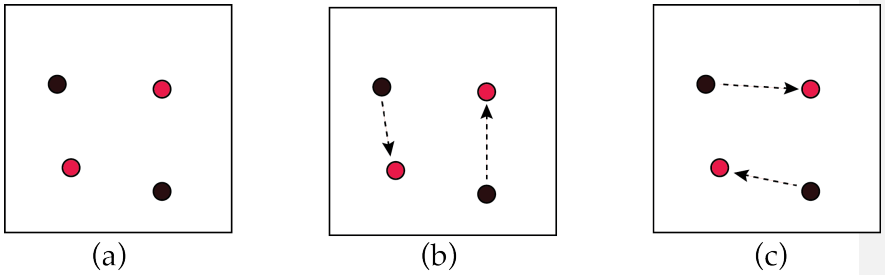

Figure 9. The correspondence problem in apparent motion.

We can model this perception as an inference. The premises of the inference are the positions of the dots in the two frames. Given these premises, the visual system tries to infer the "best" correspondence. For instance, the visual system seems to prefer correspondences in which all the dots move as little as possible from one frame to the next. It turns out that one can model these preferences, and the choice of correspondence, using Bayesian inference.[60] Briefly, if the positions of dots in the two frames is *D* and the possible correspondences are *C*, then the visual system is effectively computing the conditional probability $p(C \mid D)$ and then choosing the particular correspondence, which, say, maximizes this conditional probability.

But $p(C \mid D)$ is properly understood as a Markovian kernel: for each set of dots *D,* it gives a probability measure on the possible correspondences *C*. Thus, our conscious experience of smooth motion can be properly modeled using Markovian kernels.

Indeed, all perceptual experiences can be modeled using Bayesian inference, and, thus, by Markovian kernels. For this reason, using Markovian kernels to model perceptions and decisions (i.e., the

maps *P* and *D* of Figure 8) in the definition of conscious agents allows this definition to immediately inherit substantial support, both empirical and theoretical, from current research on conscious perceptual experiences.

The Sphere Applet also demonstrates a second transformation of conscious experience. As you can check for yourself, not only do the dots appear to move smoothly but they also appear to pop out in 3D, forming a sphere. The applet lets you play with this. If you click the button labeled “More Slant” six times, the sphere will disappear, and the dots will appear to move only in a plane. By clicking the “More Slant” and “Less Slant” buttons, you can make the sphere appear and disappear.

This conscious experience is called “structure from motion” and can also be modeled as Bayesian inference.[61] Here the visual system starts with the correspondences C and infers 3D objects T. In the process, the visual system is effectively computing the conditional probability $p(T \mid C)$, which can also be represented by a Markovian kernel. Thus, we see that conscious agents can build on each other to create new and more complex conscious experiences. In the sphere applet, one conscious agent is using the kernel $p(C \mid D)$ to create the conscious experience of smooth motion in 2D, and a second builds on this, using the kernel $p(T \mid C)$ to create the conscious experience of a 3D object.

Conscious Agents and the Combination Problem

Those who take consciousness as fundamental face what is known as the “combination problem.”[62–65] William Seager defines this as “the problem of explaining how the myriad elements of ‘atomic consciousness’ can be combined into a new, complex and rich consciousness such as that we possess.”[62] William James understood this problem back in 1890: “Where the elemental units are supposed to be feelings, the case is in no wise altered. Take a hundred of them, shuffle them and pack them as close together as you can (whatever that may mean); still each remains the same feeling it always was, shut in its own skin, windowless, ignorant of what the other feelings are and mean. There would be a hundred-and-first feeling there, if, when a group or series of such feelings were set up, a consciousness belonging to the group as such should emerge. And this 101st feeling would be a totally new fact; the 100 original feelings might, by a curious physical law, be a signal for its creation, when they came together; but they would have no substantial identity with it, nor it with them, and one could never deduce the one from the others, or (in any intelligible sense) say that

they evolved it…. The private minds do not agglomerate into a higher compound mind."

Conscious agents provide a natural solution to the combination problem. We just saw an example of this in the case of motion perception. One group of conscious agents starts with discrete frames of static dots and creates conscious experiences of dots that move smoothly in 2D. These conscious experiences are combined as the input to a higher conscious agent that creates a literally new *dimension* of conscious experience, namely, a 3D experience.

Formally, conscious agents can model such combinations of consciousness by so-called kernel "tensor products," "direct sums," and "composition." This is not the place to delve into mathematical details. But intuitively, the tensor products and sums of kernels can be used to take the output experiences of one group of conscious agents and arrange them to be the proper input for a higher conscious agent that creates a new kind of conscious experience, (e.g., a 3D experience out of input experiences that are only 2D). The hierarchy relationship between conscious agents can be modeled formally by kernel composition.[19]

More intuitively, conscious agents can be mathematically combined together to create new conscious agents. That is, when conscious agents are properly combined together, the new composite mathematical structure satisfies the definition of a conscious agent, and, thus, is a conscious agent. The sphere applet, as we just discussed, illustrates the corresponding phenomenology. There are many examples in visual perception of similar phenomena, in which conscious visual experiences of one type are combined together to form the inputs for a new conscious visual experience with literally new phenomenological features that cannot be reduced to or identified with the component experience.[15]

Conscious Agents and Quantum Bayesianism

Conscious agents provide a promising link with quantum theory. In standard formulations of quantum theory, observers play a key but controversial role; the field of quantum measurement tries to understand this role.[67–69] Although there is no consensus among experts in quantum theory about the relationship between consciousness and quantum mechanics, some theories of consciousness build on aspects of quantum theory.[70–71]

One interpretation of quantum theory that has arisen from recent work in quantum information and computation is called "quantum Bayesianism" or QBism for short.[72–73] According to QBism, the state

of a quantum system is not a description of an objective reality independent of any observer. Instead, the quantum state depends on the observer and, indeed, "a quantum state is a state of belief about what will come about as a consequence of his actions upon the system." [70] Just as the interface theory of perception claims that our perceptions do not faithfully represent the true nature of reality, the QBist claims "there is no sense in which the quantum state itself represents (pictures, copies, corresponds to, correlates with) a part or whole of the external world, much less of a world that *just is*. In fact, the very character of the theory seems to point to the inadequacy of the representationalist program when attempted on the particular world we live in." [70] In consequence, quantum measurements are not reports of objective reality: "At the instigation of a quantum measurement, something new comes into the world that was not there before; and that is about as clear an instance of *creation* as one can imagine." [70]

Why should it be that quantum states are not reports of objective reality? When a quantum state describes a quantum object in terms of position, momentum, and so forth, it is using predicates grounded in our perceptions, (e.g., of space and time and physical objects). Now physics doesn't use our perceptual predicates just as they are in our untutored perceptions. In our untutored perception of space, for instance, the moon looks about as far away from us as the stars. Physics takes our untutored perceptual predicates and extends them (e.g., using symmetry groups) to new predicates. But the basic predicates of space, time, and physical objects are simply adaptations that have been shaped by natural selection into the perceptual systems of *Homo sapiens* and, as we have seen from evolutionary game theory, natural selection does not, in general, favor true perceptions. Our perceptions were shaped to guide adaptive behavior, not to report truth.

So evolution by natural selection is the reason why quantum states are not reports of objective reality. Instead, as QBism says, the information in an observer's quantum state gives "The consequences (for *me*) of *my* actions upon the physical system." [70] Of course, natural selection has shaped our perceptions exactly for the purpose of informing us about fitness consequences of our behaviors. Fitness, not truth, is the coin of the perceptual realm. My perceptions have been shaped by natural selection to tell *me* about the fitness consequences for *me* of my actions.

The concrete technical challenge here is to connect the formal definition of a conscious agent, and the formalism of quantum theory as it is interpreted by QBism. For instance, referring again to Figure

8, the act of measurement within the formalism of conscious agents would be modeled by the kernel composition *AP*. The conscious agent can know the kernel *AP*, since this kernel is a map from its possible behaviors *G* to its possible perceptions *X*, and these are clearly known by the conscious agent. But the conscious agent cannot know *A* and *P* separately, because each kernel involves the unknown world *W*. So, according to the theory of conscious agents, every quantum measurement must be modeled as a composition of two kernels, *AP,* factoring through an unknown world *W*. What constraint does this place on models of measurement? How does it relate to the unusual calculus of probabilities that arises in the Born rule for quantum measurement, in which probabilities are given by squares of complex amplitudes? QBists have shown that the appearance of complex amplitudes in the measurement process is merely a computational convenience and not a fundamentally more powerful calculus of probabilities. One could, in principle, dispense with complex numbers and do quantum theory entirely with standard probabilities. The Born rule then turns out to be simply a quantum law of total probability, relating actual measurements to counterfactual measurements.[70] How is this related to the kernel *AP* of conscious agents, which always factors real measurements through an unknown world *W*?

12. Objections and Replies

Questioning fundamental and widely believed assumptions is no easy task. Such assumptions are widely held for good reason, and it is natural and healthy that new proposals, such as are offered here, should be met with skepticism. In this last section, I canvas a few objections and offer responses.

> *Your interface theory of perception is clearly false. It says that physical objects are just icons of a species-specific interface, and, thus, are not real. But if a bus hurtles down a road at high speed, would you step in front of it? If you did, you would find out that it is not just an icon, it is real, and your theory is nonsense.*

The interface theory of perception does indeed assert that physical objects are simply icons of a species-specific perceptual interface. Still, I would not step in front of the bus for the same reason I wouldn't carelessly drag a file icon on my desktop to the trashcan. Why? I don't take the icon *literally*, but I do take it *seriously*. The color, shape, and position of the icon are not literally true descriptions of the file. Indeed, color and shape are even the wrong language to attempt a true description. But the interface is designed

to guide useful behaviors, and those behaviors have consequences even if the interface does not literally resemble the truth. Natural selection shaped our perceptions, in part, to keep us alive long enough to reproduce. We had better take our perceptions seriously. If you see a tiger, keep away. If you see a cliff, don't step over. Natural selection ensures that we must take our perceptions seriously. But it is a logical error to conclude that we must, therefore, take our perceptions literally.

As discussed before, the interface theory of perception fits well with QBist interpretations of quantum theory, which say that we should not take quantum states literally as descriptions of an objective reality independent of the observer. Thus, the interface theory is not falsified by current physics but instead fits well with and even offers evolutionary explanations for puzzling aspects of quantum physics.

The objection uses the world *real*. This word is used with two very different meanings. In the objection, it is used to mean that something exists even if it is not observed. So, the bus is argued to be real in the sense that it would exist even if no one observed it. But there is another sense of real, as when I say I have a real headache. The headache would not exist if no one (e.g., me) observed it. But if you claimed on those grounds that my headache wasn't real, I would be cross with you. So the interface theory says that physical objects such as a bus are real in the headache sense of real. But it denies that they are real in the sense of existing whether or not they are observed.

> *Doesn't the interface theory say that the moon is only there when you look? That's clearly absurd.*

Yes, the interface theory says that the moon is only there when I look. However, the interface theory does not deny that, when I see the moon, *something* exists whether I observe it or not. But that something is not the moon, and it is probably not anything in space and time. Space, time, and the moon are just the best that I, as a humble member of the species *H. sapiens*, can come up with. There is a reality that exists independent of my perceptions; the interface theory does not endorse metaphysical solipsism. But it is an elementary mistake to assume that what exists in any way resembles what I perceive.

The moon is my perceptual experience. When you see the moon, you have your own perceptual experience that is distinct from (not numerically identical to) my perceptual experience. So when we both look up at "the moon" there are actually two moons, one of your

experience and one of mine. There is something that exists that triggers each of us to create an experience of the moon, but that something, in all probability, does not resemble the moon.

> *Actually, the interface theory is nothing new. Physicists have been telling us for decades that objects are mostly empty space. That desk looks solid, but it is really just particles whizzing through empty space at high speeds.*

Indeed, physicists have been telling us this for some time. But the claim of the interface theory is different, and more radical. It says that the particles themselves, and the empty space through which they travel, are not the objective reality. They are still part of the interface. Suppose I admit that the icon on my desktop is not the reality of the file, but then I whip out a magnifying glass, look closely at the icon, and conclude that the pixels I see are the reality. I've made a fundamental mistake. The pixels are still part of the desktop interface and they don't resemble the real file any more than the icon does. The same is true of the particles whizzing through empty space.

> *The interface theory of perception means science is not possible. If our senses don't deliver the truth, then how can science possibly proceed?*

The interface theory poses no problem to science. It simply says that one particular theory is incorrect, viz., the theory that objective reality consists in part of space, time, and physical objects. Discarding false theories is genuine scientific progress. Now that we know not to take our perceptions at face value, we can be more sophisticated in their interpretation. We now understand that our perceptions are shaped by natural selection to inform us about fitness, not truth. We can still construct theories about the nature of objective reality and about how that reality relates to our perceptions. We can then make empirical predictions that can be tested. The methodology of science is not called into question by the interface theory.

> *You use evolutionary game theory to conclude that our perceptions do not report the truth. But how about our logic and mathematics? Does evolution also shape them to be incorrect? And if so, isn't this a defeater for your whole program? You use the logic and mathematics of evolution to conclude that logic and mathematics are unreliable.*

I agree that if evolutionary games show that natural selection favors incorrect logic and mathematics, then I have a real problem. It would be self-refuting. This is clearly an important research area.

I think, however, that it will turn out that the same evolutionary games which demonstrate that natural selection does not favor true perceptions will also demonstrate that natural selection favors true logic and mathematics. Suppose, for instance, that the objective world contains two resources and that the *fitness payoff* of these resources, for a specific organism, depends on the *sum* of the resource quantities. Then an organism whose perceptual system performs the sum correctly will be better able to reap the fitness benefits of those resources than one that does not. More generally, if the fitness payoffs are some function *f* of structures in the objective world, then selection pressures will shape organisms to correctly compute *f*.

There are a couple provisos. First, the selection pressures will only shape organisms to correctly compute the portions of *f* that are, in fact, relevant to fitness. If, for instance, the payoff function rewards only one element of the range of *f* and gives no rewards for any other elements of its range, then an organism that only correctly computes the pullback of that single element will be able to reap all the fitness rewards. However, as the behavioral repertoire of the organism increases and other elements of the range of *f* are rewarded for different behaviors, then the organism will need to correctly compute the pullbacks of these elements as well. Thus, the selection pressures are toward truth, even if, in practice, they don't get all the way there.

A second proviso is that it is not clear that selection pressures will uniquely determine the range of a function. It appears that, as long as all the pullbacks are computed correctly, they can be randomly assigned (even incorrectly) to different elements of the range, and the organism can still reap all the fitness benefits. Thus, it might turn out that selection pressures are toward the truth, but only up to automorphisms of the range of functions.

Now I have been speaking of logic and math as they apply in the normal functioning of our perceptual processing, not as they are used in our deliberate reasoning. It is quite possible that our deliberate reasoning has evolved not as a guide to truth but simply to serve some other useful function. Dan Sperber and his colleagues, for instance, argue that reasoning evolved to allow us to devise and evaluate arguments designed to persuade others about what we want.[75] The goal of our reasoning is successful argument,

not truth. And this, they suggest, is one reason for the notorious confirmation bias in human reasoning.

The ideas discussed here have implications for long-standing debates about whether evolution is compatible with the claim that our cognitive faculties are reliable. Plantinga, for instance, argues that evolution and naturalism together make it improbable or inscrutable whether our cognitive faculties are reliable; this, he says, is a defeater for all our beliefs, including beliefs in evolution and naturalism.[35] But the ideas discussed here suggest that the question must be refined if we are to make real progress. Asking whether evolution is likely to produce reliable cognitive faculties is too broad a question. Perhaps evolution produces untrue perceptions but reliable logic and mathematics. We shall have to look at each aspect of human cognition separately and ask, using tools such as evolutionary games and genetic algorithms, what natural selection is likely to do with that aspect.

> *When you dismissed the integrated information theory (IIT) of consciousness, you dismissed the measure Φ of integrated information, which may turn out to be useful in the study of consciousness. This is a serious mistake.*

I did not dismiss IIT tout court. I dismissed Tononi's claim of *identity* between consciousness and Φ. That claim is false, as is established by the scrambling theorem. But I am certainly open to the possibility that Φ will turn out to be a useful measure in the study of consciousness. If so, it can be applied within the formalism of conscious agents. The Markovian kernels within that formalism are amenable to IIT analyses such as effective information and Φ.

> *Your interface theory of perception and conscious-agent theory of consciousness make no predictions and are thus not genuine scientific theories.*

Here are some predictions. No physical object has real values of dynamical physical properties (such as position, momentum, spin) when it is not observed. If we find definitive evidence otherwise, my theories would be in ruins. The experimental evidence so far is that quantum objects violate Bell's inequalities, which is often interpreted as a refutation of local realism;[67] such an interpretation is exactly what is predicted by the interface theory of perception. However, other interpretations such as Bohm's, which keeps realism at the expense of locality, and Everett's, which keeps realism at the expense of counterfactual definiteness, are not ruled out.

Another prediction: No physical object has any causal powers. I call this doctrine *epiphysicalism*: Consciousness creates physical objects and their properties, but physical objects themselves have no causal powers. This is the converse of *epiphenomenalism*, which claims that physical objects, such as brains, create conscious experiences, but conscious experiences themselves have no causal powers. If any physical object were shown to have causal powers, my theories would be in ruins.

Another prediction: Every perceptual capacity can be represented by the conscious-agent formalism. If there were some perceptual capacity whose formal statement could not be represented within the formalism of conscious agents, then the conscious-agent formalism would be falsified. This claim about conscious agents and perceptual capacities is analogous to the claim that is made about Turing machines and effective procedures. The Church-Turing thesis states that every algorithm can be instantiated by some Turing machine. Were someone to produce an algorithm that could not be so instantiated, then the Church-Turing thesis would be falsified, and Turing machines would be an inadequate representation of algorithms. Similarly, the Conscious-Agent thesis states that every perceptual capacity can be instantiated by some conscious agent. Were someone to produce a perceptual capacity that could not be so instantiated, then the Conscious-Agent thesis would be falsified. The Conscious-Agent thesis is effectively the claim that conscious agents are an adequate formalism to represent all conscious perceptual experiences.

Acknowledgements
For helpful discussions, I thank P. Foley, B. Marion, J. Mark, C. Prakash, M. Singh, G. Souza, and K. Stephens. Any errors are, of course, mine, not theirs. I also thank Elaine Ku for the rabbit images used in Figure 3.

References

[1] G. Miller, “What is the Biological Basis of Consciousness?” *Science, 309* (2005), 79.
[2] C. Koch, *The Quest for Consciousness: A Neurobiological Approach* (Englewood, CO: Roberts & Company, 2004).
[3] T. J. Huxley, “Lessons in Elementary Psychology,” 8 (1866), 210.
[4] C. Koch, *Consciousness: Confessions of a romantic reductionist* (Cambridge, MA: MIT Press, 2012).
[5] F. Crick, *The Astonishing Hypothesis: The Scientific Search for the Soul* (New York: Scribners, 1994).

[6] C. McGinn, “Can We Solve the Mind-body Problem?” *Mind, 98,* (1989), 349–366.
[7] A. Peres, *Quantum Theory: Concepts and Methods* (Boston: Kluwer, 1995).
[8] W. V. O. Quine, “Two Dogmas of Empiricism,” *The Philosophical Review*, 60 (1951), 20-43.
[9] J. Mark, B. Marion, and D. D. Hoffman, “Natural Selection and Veridical Perceptions,” *Journal of Theoretical Biology*, 266 (2010), 504–515.
[10] D. D. Hoffman, M. Singh, “Computational Evolutionary Perception,” *Perception*, 41 (2012), 1073-1091.
[11] D. D. Hoffman, M. Singh, J. Mark, “Does Evolution Favor True Perceptions?” *Proceedings of the SPIE, Human Vision and Electronic Imaging XVIII* (2013), Doi: 10.1117/12.2011609.
[12] M. Singh, D. D. Hoffman, “Natural Selection and Shape Perception,” *Shape Perception in Human and Computer Vision: An Interdisciplinary Perspective*, edited by S. Dickinson, S. and Z. Pizlo (New York: Springer, 2013).
[13] D. D. Hoffman, “The Scrambling Theorem: A Simple Proof of the Logical Possibility of Spectrum Inversion,” *Consciousness and Cognition*, 15 (2006), 31-45.
[14] D. D. Hoffman, “The Scrambling Theorem Unscrambled: A Response to Commentaries,” *Consciousness and Cognition*, 15 (2006), 51-43.
[15] D. D. Hoffman, *Visual intelligence: How we create what we see* (New York: W.W. Norton, 1998).
[16] D. D.Hoffman, “The Interface Theory of Perception,” *Object Categorization: Computer and Human Vision Perspectives,* edited by S. Dickinson, M. Tarr, A. Leonardis, B. Schiele (Cambridge: Cambridge University Press, 2009), 148-165.
[17] J. J. Koenderink, “Vision and information,” *Perception Beyond Inference. The Information Content of Visual Processes* edited by L. Albertazzi, G. V. Tonder, and D. Vishnawath, D. (Cambridge, MA: MIT Press, 2011).
[18] J. J. Koenderink, “World, Environment, Umwelt, and Innerworld: A Biological Perspective on Visual Awareness,” *Proceedings of the SPIE, Human Vision and Electronic Imaging XVIII* (2013) Doi: 10.1117/12.2011874.
[19] B. M. Bennett, D. D. Hoffman, and C. Prakash, *Observer mechanics: A formal theory of perception* (San Diego: Academic Press, 1989).
[20] D. D. Hoffman, “Conscious Realism and the Mind-body Problem,” *Mind & Matter*, 6 (2008), 87-121.
[21] R. Briggs, “Distorted Reflection,” *Philosophical Review*, 118 (2009), 59-85.

[22] S. Palmer, *Vision science: Photons to phenomenology* (Cambridge, MA: MIT Press, 1999).
[23] A. Noë, and J. K. Regan, “On the Brain-basis of Visual Consciousness: A Sensorimotor Account,” *Vision and Mind: Selected Readings in the Philosophy of Perception*, edited by A. Noë and E. Thompson (MIT Press, Cambridge, MA, 2002).
[24] W. S. Geisler, and R. L. Diehl, “A Bayesian Approach to the Evolution of Perceptual and Cognitive Systems,” *Cognitive Science*, 27 (2003), 379-402.
[25] D. Marr, *Vision*: *A Computational Investigation into the Human Representation and Processing of Visual Information* (San Francisco: Freeman, 1982).
[26] A. Shimony, “Perception from an Evolutionary Point of View,” *Journal of Philosophy*, 68, 19 (1971), 571–583.
[27] T. Nagel, *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False* (Oxford University Press, 2012).
[28] M. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life* (Cambridge, MA: Belknap Press of Harvard University Press, 2006).
[29] M. Mitchell, *An Introduction to Genetic Algorithms* (Cambridge, MA: MIT Press Bradford, 1998).
[30] W. Fish, *Perception, Hallucination, and Illusion* (Oxford Press, 2009).
[31] W. Fish, *Philosophy of Perception: A Contemporary Introduction* (New York: Routledge, 2010).
[32] K. Lorenz, *Behind the Mirror: A Search for a Natural History of Human Knowledge* (New York: Harcourt Brace Jovanovich, 1973).
[33] H. Kornblith, *Naturalizing Epistemology* (Cambridge, MA: MIT Bradford, 1987).
[34] G. Radnitzky, and W. W. Bartley, *Evolutionary Epistemology, Rationality, and the Sociology of Knowledge* (La Salle, IL: Open Court, 1993).
[35] J. Belby, *Naturalism Defeated?* (Ithaca, NY: Cornell University Press, 2002).
[36] B. Skyrms, *Signals: Evolution, Learning, and Information* (Oxford University Press, 2010).
[37] J. Hofbauer, and K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge University Press, 1998).
[38] J. Maynard Smith, *Evolution and the Theory of Games* (Cambridge University Press, 1982).
[39] M. A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life* (Cambridge, MA: Belknap Press, 2006).
[40] L. Samuelson, *Evolutionary Games and Equilibrium Selection* (Cambridge, MA: MIT Press, 1997)

[41] J. Locke, *An Essay Concerning Human Understanding* (Oxford University Press, 1690/1979).
[42] N. Block, J. Fodor, “What Psychological States Are Not, *Philosophical Review*, 81 (1972), 159–181.
[43] J. Bickle, *Philosophy and Neuroscience: A Ruthlessly Reductive Account* (Dordrecht: Kluwer Academic Publishers, 2003).
[44] D. Chalmers, *The Conscious Mind* (Oxford University Press, 1996).
[45] P. S. Churchland, *Brain-wise: Studies in Neurophilosophy* (Cambridge, MA: MIT Press, 2002).
[46] D. Dennett, *Consciousness Explained* (Boston: Back Bay Books, 1992).
[47] G. Tononi, “Consciousness as Integrated Information: A Provisional Manifesto” *Biological Bulletin*, 215 (2008), 216-242.
[48] S. Palmer, “Color, Consciousness, and the Isomorphism Constraint,” *Behavioral and Brain Sciences*, 22 (1999), 923–989.
[49] G. Tononi, and G.Edelman, “Consciousness and Complexity,” *Science*, 282 (1998), 1846-1851.
[50] G. Tononi, and O. Spoorns, “Measuring Information Integration,” *BMC Neuroscience*, 4 (2003), 31.
[51] G. Tononi, “An Information Integration Theory of Consciousness,” *BMC Neuroscience*, 5 (2004), 42.
[52] G. Tononi, and C. Koch, “The Neural Correlates of Consciousness: An Update,” *Annals of the New York Academy of Sciences*, 1124 (2008), 239-261.
[53] A. B. Barrett, A. K. Seth, “Practical Measures of Integrated Information for Time-series Data,” *PLOS Computational Biology*, 7 (2011), 1.
[54] A. Treisman, and H. Schmidt, “Illusory Conjunctions in the Perception of Objects,” *Cognitive Psychology*, 14(1), (1982), 107-141.
[55] P. T. Quinlan, “Visual Feature Integration Theory: Past, Present, and Future.” *Psychological Bulletin*, 5 (2003), 643-673.
[56] D. J. Simons, and M. S. Ambinder, “Change Blindness: Theory and Consequences,” *Current Directions in Psychological Science*, 14 (2005), 44-48.
[57] D. Knill, and W. Richards, *Perception as Bayesian Inference* (Cambridge University Press, 1996).
[58] D. Kersten, P. Mamassian, and A. Yuille, “Object Perception as Bayesian Inference,” *Annual Review of Psychology*, 55 (2004), 271-304.
[59] T. E. Hudson, L. T. Maloney, and M. S. Landy, “Optimal Compensation for Temporal Uncertainty in Movement Planning,” *PLOS Computational Biology*, 4(7), (2008), e1000130. doi: 10.1371/journal.pcbi.1000130.

[60] J. C. Read, “A Bayesian Model of Stereopsis Depth and Motion Direction Discrimination,” *Biological Cybernetics*, 86 (2002), 117-136.
[61] D. A. Forsythe, S. Ioffe, and J. Haddon, “Bayesian Structure from Motion,” *Proceedings of the 7th IEEE International Conference on Computer Vision*, 1 (1999), 660-665.
[62] W. Seager, “Consciousness, Information, and Panpsychism,” *Journal of Consciousness Studies*, 2 (1995), 272-288.
[63] P. Goff, “Why Panpsychism Doesn’t Help Us Explain Consciousness,” *Dialectica*, 63 (2009), 289-311.
[64] M. Blamauer, “Is the Panpsychist Better Off as an Idealist? Some Leibnizian Remarks on Consciousness and Composition,” *Eidos*, 15 (2011), 48-75.
[65] S. Coleman, “The real Combination Problem: Panpsychism, Micro-subjects, and Emergence,” *Erkenntnis.* (2013), DOI 10.1007/s10670-013-9431-x.
[66] W. James, *The Principles of Psychology* (Vol. 1) (New York: Cosimo Inc., 1890/2007).
[67] D. Albert, *Quantum Mechanics and Experience* (Cambridge, MA: Harvard University Press, 1992).
[68] J. A. Wheeler, and W. H. Zurek, *Quantum Theory and Measurement* (Princeton University Press, 1983).
[69] G. Greenstein, and A. G. Zajonc, *The Quantum Challenge* (Sudbury, MA: Jones and Bartlett, 2005).
[70] N. Herbert, *Elemental Mind: Human Consciousness and the New Physics* (New York: Plume, 1993).
[71] S. Kak, R. Schild, R. Penrose, and S. Hameroff, *Cosmology of Consciousness: Quantum Physics and Neuroscience of Mind* (Cambridge, MA: Cosmology Science Publishers, 2011).
[72] S. Kak, R. Penrose, and S. Hameroff, *Quantum Physics of Consciousness* (Cambridge, MA: Cosmology Science Publishers, 2011).
[73] C. A. Fuchs, “QBism, the Perimeter of Quantum Bayesianism, arXiv:1003.5209 (2010).
[74] C. A. Fuchs, R. Schack, “A Quantum-Bayesian Route to Quantum-state Space,” *Foundations of Physics*, 41 (2011), 345-356.
[75] H. Mercier, D. Sperber, “Why Do Humans Reason? Arguments for an Argumentative Theory,” *Behavioral and Brain Sciences*, 34 (2011), 57-111.