Quanta magazine

The Evolutionary Argument Against Reality

The cognitive scientist Donald Hoffman uses evolutionary game theory to show that our perceptions of an independent reality must be illusions.

By Amanda Gefter
20160421

As we go about our daily lives, we tend to assume that our perceptions — sights, sounds, textures, tastes — are an accurate portrayal of the real world. Sure, when we stop and think about it — or when we find ourselves

fooled by a perceptual illusion — we realize with a jolt that what we perceive is never the world directly, but rather our brain's best guess at what that world is like, a kind of internal simulation of an external reality. Still, we bank on the fact that our simulation is a reasonably decent one. If it wasn't, wouldn't evolution have weeded us out by now? The true reality might be forever beyond our reach, but surely our senses give us at least an inkling of what it's really like.

Not so, says Donald D. Hoffman, a professor of cognitive science at the University of California, Irvine. Hoffman has spent the past three decades studying perception, artificial intelligence, evolutionary game theory and the brain, and his conclusion is a dramatic one: The world presented to us by our perceptions is

nothing like reality. What's more, he says, we have evolution itself to thank for this magnificent illusion, as it maximizes evolutionary fitness by driving truth to extinction.

Getting at questions about the nature of reality, and disentangling the observer from the observed, is an endeavor that straddles the boundaries of neuroscience and fundamental physics. On one side you'll find researchers scratching their chins raw trying to understand how a three-pound lump of gray matter obeying nothing more than the ordinary laws of physics can give rise to first-person conscious experience. This is the aptly named "hard problem." On the other side are quantum physicists, marveling at the strange fact that quantum systems don't seem to be definite objects localized in space until we come along to observe them — whether we are conscious humans or inanimate measuring

devices. Experiment after experiment has shown — defying common sense — that if we assume that the particles that make up ordinary objects have an objective, observer-independent existence, we get the wrong answers. The central lesson of quantum physics is clear: There are no public objects sitting out there in some preexisting space. As the physicist John Wheeler put it, "Useful as it is under ordinary circumstances to say that the world exists 'out there' independent of us, that view can no longer be upheld."

So while neuroscientists struggle to understand how there can be such a thing as a first-person reality, quantum physicists have to grapple with the mystery of how there can be anything *but* a first-person reality. In short, all roads lead back to the observer. And that's where you can find Hoffman — straddling the boundaries, attempting a mathematical model of the

observer, trying to get at the reality behind the illusion. *Quanta Magazine* caught up with him to find out more. An edited and condensed version of the conversation follows.

QUANTA MAGAZINE: People often use Darwinian evolution as an argument that our perceptions accurately reflect reality. They say, "Obviously we must be latching onto reality in some way because otherwise we would have been wiped out a long time ago. If I think I'm seeing a palm tree but it's really a tiger, I'm in trouble."

> Evolution has shaped us with perceptions that allow us to survive. But part of that involves hiding from us the stuff we don't need to know.

And that's pretty much all of reality, whatever reality might be.

DONALD HOFFMAN: Right. The classic argument is that those of our ancestors who saw more accurately had a competitive advantage over those who saw less accurately and thus were more likely to pass on their genes that coded for those more accurate perceptions, so after thousands of generations we can be quite confident that we're the offspring of those who saw accurately, and so we see accurately. That sounds very plausible. But I think it is utterly false. It misunderstands the fundamental fact about evolution, which is that it's about fitness functions — mathematical functions that describe how well a given strategy achieves the goals of survival and

reproduction. The mathematical physicist Chetan Prakash proved a theorem that I devised that says: According to evolution by natural selection, an organism that sees reality as it is will never be more fit than an organism of equal complexity that sees none of reality but is just tuned to fitness. Never.

You've done computer simulations to show this. Can you give an example?

Suppose in reality there's a resource, like water, and you can quantify how much of it there is in an objective order — very little water, medium amount of water, a lot of water. Now suppose your fitness function is linear, so a little water gives you a little fitness, medium water gives you medium fitness, and lots of water gives

you lots of fitness — in that case, the organism that sees the truth about the water in the world can win, but only because the fitness function happens to align with the true structure in reality. Generically, in the real world, that will never be the case. Something much more natural is a bell curve — say, too little water you die of thirst, but too much water you drown, and only somewhere in between is good for survival. Now the fitness function doesn't match the structure in the real world. And that's enough to send truth to extinction. For example, an organism tuned to fitness might see small and large quantities of some resource as, say, red, to indicate low fitness, whereas they might see intermediate quantities as green, to indicate high fitness. Its perceptions will

be tuned to fitness, but not to truth. It won't see any distinction between small and large — it only sees red — even though such a distinction exists in reality.

But how can seeing a false reality be beneficial to an organism's survival?

There's a metaphor that's only been available to us in the past 30 or 40 years, and that's the desktop interface. Suppose there's a blue rectangular icon on the lower right corner of your computer's desktop — does that mean that the file itself is blue and rectangular and lives in the lower right corner of your computer? Of course not. But those are the only things that can be asserted about anything on the desktop — it has color, position and shape. Those are the only categories available to you, and

yet none of them are true about the file itself or anything in the computer. They couldn't possibly be true. That's an interesting thing. You could not form a true description of the innards of the computer if your entire view of reality was confined to the desktop. And yet the desktop is useful. That blue rectangular icon guides my behavior, and it hides a complex reality that I don't need to know. That's the key idea. Evolution has shaped us with perceptions that allow us to survive. They guide adaptive behaviors. But part of that involves hiding from us the stuff we don't need to know.

And that's pretty much all of reality, whatever reality might be. If you had to spend all that time figuring it out, the tiger would eat you.

So everything we see is one big illusion?

We've been shaped to have perceptions that keep us alive, so we have to take them seriously. If I see something that I think of as a snake, I don't pick it up. If I see a train, I don't step in front of it. I've evolved these symbols to keep me alive, so I have to take them seriously. But it's a logical flaw to think that if we have to take it seriously, we also have to take it literally.

If snakes aren't snakes and trains aren't trains, what are they?

Snakes and trains, like the particles of physics, have no objective, observer-independent features. The snake I see is a description created by my sensory system to inform me of the fitness consequences

---

https://www.quantamagazine.org/the-evolutionary-argument-against-reality p/.

of my actions. Evolution shapes acceptable solutions, not optimal ones. A snake is an acceptable solution to the problem of telling me how to act in a situation. My snakes and trains are my mental representations; your snakes and trains are your mental representations.

How did you first become interested in these ideas?

As a teenager, I was very interested in the question "Are we machines?" My reading of the science suggested that we are. But my dad was a minister, and at church they were saying we're not. So I

decided I needed to figure it out for myself. It's sort of an important personal question — if I'm a machine, I would like to find that out! And if I'm not, I'd like to know, what is that special magic beyond the machine? So eventually in the 1980s I went to the artificial intelligence lab at MIT and worked on machine perception. The field of vision research was enjoying a newfound success in developing mathematical models for specific visual abilities. I noticed that they seemed to share a common mathematical structure, so I thought it might be possible to write down a formal structure for observation that

encompassed all of them, perhaps all possible modes of observation. I was inspired in part by Alan Turing. When he invented the Turing machine, he was trying to come up with a notion of computation, and instead of putting bells and whistles on it, he said, Let's get the simplest, most pared down mathematical description that could possibly work. And that simple formalism is the foundation for the science of computation. So I wondered, could I provide a similarly simple formal foundation for the science of observation?

A mathematical model of consciousness.

That's right. My intuition was, there are conscious experiences. I have pains, tastes, smells, all my sensory experiences, moods, emotions and so forth. So I'm just going to say: One part of this consciousness structure is a set of all possible experiences. When I'm having an experience, based on that experience I may want to change what I'm doing. So I need to have a collection of possible actions I can take and a decision strategy that, given my experiences, allows me to change how I'm acting. That's the basic idea of the whole thing. I have a space X of experiences, a space G of actions, and an algorithm D that lets me choose a new action given my

experiences. Then I posited a W for a world, which is also a probability space. Somehow the world affects my perceptions, so there's a perception map P from the world to my experiences, and when I act, I change the world, so there's a map A from the space of actions to the world. That's the entire structure. Six elements. The claim is: This is the structure of consciousness. I put that out there so people have something to shoot at.

But if there's a W, are you saying there is an external world?

Here's the striking thing about that. I can pull the W out of the model and stick a conscious agent in its place and get a circuit of conscious

agents. In fact, you can have whole networks of arbitrary complexity. And that's the world.

The world is just other conscious agents?

I call it conscious realism: Objective reality is just conscious agents, just points of view. Interestingly, I can take two conscious agents and have them interact, and the mathematical structure of that interaction also satisfies the definition of a conscious agent. This mathematics is telling me something. I can take two minds, and they can generate a new, unified single mind.

Here's a concrete example. We have two hemispheres in our brain. But

when you do a split-brain operation, a complete transection of the corpus callosum, you get clear evidence of two separate consciousnesses. Before that slicing happened, it seemed there was a single unified consciousness. So it's not implausible that there is a single conscious agent. And yet it's also the case that there are two conscious agents there, and you can see that when they're split. I didn't expect that, the mathematics forced me to recognize this. It suggests that I can take separate observers, put them together and create new observers, and keep doing this ad infinitum. It's conscious agents all the way down.

If it's conscious agents all the way down, all first-person points of view, what happens to science? Science has always been a third-person description of the world.

The idea that what we're doing is measuring publicly accessible objects, the idea that objectivity results from the fact that you and I can measure the same object in the exact same situation and get the same results — it's very clear from quantum mechanics that that idea has to go. Physics tells us that there are no public physical objects. So what's going on? Here's how I think about it. I can talk to you about my headache and believe that I am

communicating effectively with you, because you've had your own headaches. The same thing is true as apples and the moon and the sun and the universe. Just like you have your own headache, you have your own moon. But I assume it's relevantly similar to mine. That's an assumption that could be false, but that's the source of my communication, and that's the best we can do in terms of public physical objects and objective science.

It doesn't seem like many people in neuroscience or philosophy of mind are thinking about fundamental physics. Do you think that's been a stumbling block for

those trying to understand consciousness?

I think it has been. Not only are they ignoring the progress in fundamental physics, they are often explicit about it. They'll say openly that quantum physics is not relevant to the aspects of brain function that are causally involved in consciousness. They are certain that it's got to be classical properties of neural activity, which exist independent of any observers — spiking rates, connection strengths at synapses, perhaps dynamical properties as well. These are all very classical notions under Newtonian physics, where time is absolute and objects exist absolutely. And then

[neuroscientists] are mystified as to why they don't make progress. They don't avail themselves of the incredible insights and breakthroughs that physics has made. Those insights are out there for us to use, and yet my field says, "We'll stick with Newton, thank you. We'll stay 300 years behind in our physics."

I suspect they're reacting to things like Roger Penrose and Stuart Hameroff's model, where you still have a physical brain, it's still sitting in space, but supposedly it's performing some quantum feat. In contrast, you're saying, "Look, quantum mechanics is telling us

that we have to question the very notions of 'physical things' sitting in 'space.'"

I think that's absolutely true. The neuroscientists are saying, "We don't need to invoke those kind of quantum processes, we don't need quantum wave functions collapsing inside neurons, we can just use classical physics to describe processes in the brain." I'm emphasizing the larger lesson of quantum mechanics: Neurons, brains, space … these are just symbols we use, they're not real. It's not that there's a classical brain that does some quantum magic. It's that there's no brain! Quantum mechanics says that classical objects

— including brains — don't exist. So this is a far more radical claim about the nature of reality and does not involve the brain pulling off some tricky quantum computation. So even Penrose hasn't taken it far enough. But most of us, you know, we're born realists. We're born physicalists. This is a really, really hard one to let go of.

To return to the question you started with as a teenager, are we machines?

The formal theory of conscious agents I've been developing is computationally universal — in that sense, it's a machine theory. And it's because the theory is

computationally universal that I can get all of cognitive science and neural networks back out of it. Nevertheless, for now I don't think we are machines — in part because I distinguish between the mathematical representation and the thing being represented. As a conscious realist, I am postulating conscious experiences as ontological primitives, the most basic ingredients of the world. I'm claiming that experiences are the real coin of the realm. The experiences of everyday life — my real feeling of a headache, my real taste of chocolate — that really is the ultimate nature of reality.

This article was reprinted on

TheAtlantic.com.

[20170903 – Formated by Wergosum]