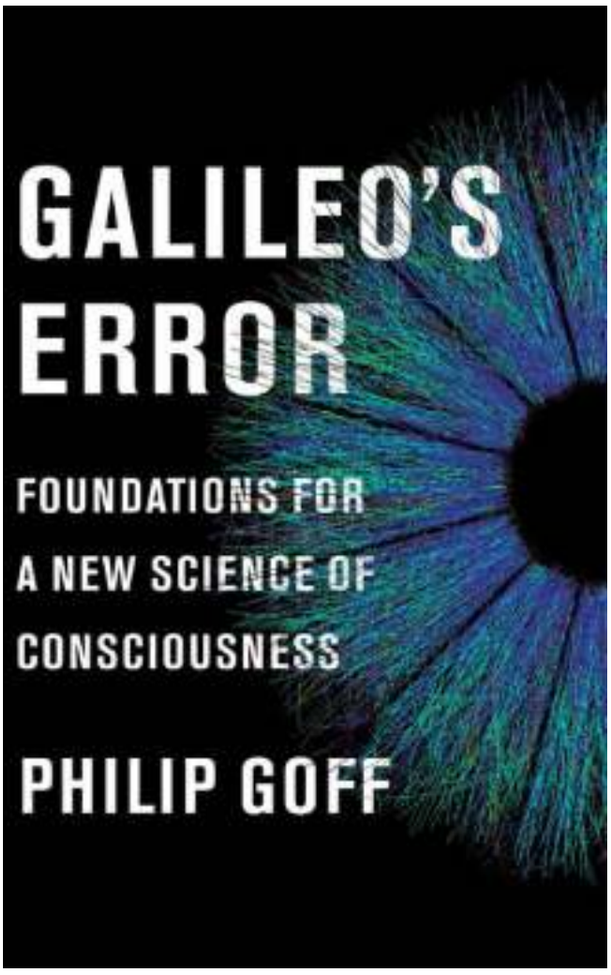
A Post-Galilean Paradigm | Edge.org

A Conversation with

A POST-GALILEAN PARADIGM



It's broadly agreed these days that consciousness poses a very serious challenge for contemporary science. What I'm trying to work out at the moment is why science has such difficulty with consciousness. We can trace this problem back to its root, at the start of the scientific revolution.

A crucial moment in the scientific revolution is when Galileo declares that mathematics is to be the language of the new science. The new science is to have a purely quantitative vocabulary. This is a much-discussed moment, but what is less reflected on is the philosophical work Galileo had to do to get to that point.

Before Galileo, people thought the physical world was full of qualities—the colors on the surfaces of objects, tastes in food, smells floating through the air. The trouble is, you can't capture these qualities in the purely quantitative vocabulary of mathematics. You can't capture the redness of a red experience or the spiciness of paprika in an equation. This was a challenge for Galileo's aspiration to describe the physical world in mathematics. Galileo's solution to this was to propose a radically new philosophical theory of reality. According to this

theory, the qualities aren't really out there in the physical world, they're in the soul, which Galileo took to be outside of the domain of science. The redness isn't on the surface of the tomato, it's in the soul of the person perceiving the tomato. The spiciness of the paprika isn't in the paprika, it's in the soul of the person eating it. Galileo stripped the physical world of its qualities, and after he'd done that, all that remained were the purely quantitative features of matter—size, shape, location, motion—that can be captured in a purely mathematical vocabulary in mathematical geometry. This is the start of mathematical physics.

It's crucial to realize that in Galileo's worldview, this radical division between the physical world—with its purely quantitative properties—is the domain of science, and the soul—with its qualities—is outside the domain of science. Mathematical physics has obviously gone very well, but the problem is that you can't deal with consciousness if you're not going to deal with qualities because conscious experience is essentially defined by the qualities that characterize every second of waking life—the colors, the smells, the sounds, the tastes. Effectively, by excluding qualities from the domain of science, Galileo excluded consciousness from the domain of science. To be fair to Galileo, he was completely clear about this. He only ever intended physical science as a partial description of reality. If Galileo were to time travel to the present day and hear about this problem of explaining consciousness in physical science terms, he'd say, "Of course you can't do that. I designed physical science to deal with quantities, not qualities."

We're now going through a phase of history where people are so blown away at the success of physical science and the wonderful technology it's produced that they've forgotten its philosophical underpinnings. They've forgotten its inherent limitations. If we want a science of consciousness, we need to move beyond Galileo. We need to move to what I call a post-Galilean paradigm. We need to rethink what science is. That doesn't mean we stop doing physical science or we do physical science differently—I'm not here to tell physical scientists how to do their jobs. It does, however, mean that it's not the full story. We need physical science to encompass a more expansive conception of the scientific method. We need to adopt a worldview that can accommodate both the quantitative data of physical science and the qualitative reality of consciousness. That's essentially the problem.

Fortunately, there is a way forward. There is a framework that could allow us to make progress on this. It's inspired by certain writings from the 1920s of the philosopher Bertrand Russell and the scientist Arthur Eddington, who is incidentally the first scientist to confirm general relativity after the First World War. I'm inclined to think that these guys did in the 1920s for the science of consciousness what Darwin did in the 19th century for the science of life. It's a tragedy of history that this was completely forgotten about for a long time for various historical reasons we could talk about. But, it's recently been rediscovered in the last five or ten years in academic philosophy, and it's causing a lot of excitement and interest.

There are two dominant positions on consciousness. On the one hand, people think that it's so magical and mysterious that we're never going to be able to give a scientific account of it. In a way, that's Nick Humphrey's position. He wants to say in some sense that it's an illusion, that we can't give a scientific account of it. But it's also the view of the dualist who thinks it's just outside of the domain of science. I was interviewed recently by a very radical dualist. I'm usually prepared to defend that my view is scientific, but this guy was saying, "Why are you bothering with science? We all know science is a load of rubbish." On one hand, it's so magical and mysterious that we'll never get a scientific account of it. The other view is that we just need to keep doing neuroscience in standard ways and we'll eventually crack it.

My view is in the middle. We hope, at least, that one day we will have a science of consciousness. But we need to rethink what science is because I don't think physical science was ever designed to deal with consciousness. It was designed to give mathematical models that can accurately predict the behavior of matter, and that's gone really well, but it was never designed to deal with the subjective qualities of consciousness.

Russell is a famous philosopher. People know about his logical linguistic work and his pacifism, but his views from *The Analysis of Matter*, in 1927, have been almost completely airbrushed out of history. Eddington followed it up in his Gifford Lectures, also in 1927. It's a wonderful interaction between science and philosophy. The starting point of Russell and Eddington is that physical science tells you a lot less than you think about the nature of matter. In the public mind, physics is on its way to giving us this complete story of the nature of space, time, and matter. But Eddington and Russell realized that, on reflection, physical science is confined to telling you about the behavior of matter, about what it does. Think about what physics tells us about an electron: An electron has, for example, mass and negative charge. What is mass? Physics tells us that things with mass attract other things with mass and resist acceleration. The more mass they have, the more they resist acceleration. What is negative charge? Things with negative charge attract things with positive charge and repel other things with negative charge. This all concerns the behavior of the electron—what it does. Physics is confined to telling us about the behavior. We find a similar story in the higher-level sciences of chemistry and neurophysiology. As a whole, physical science tells us about behavior.

This is incredibly useful information. If you have rich information about the behavior of matter, you can manipulate the physical world in all sorts of extraordinary ways wielding the incredible technology that's transformed our planet. But if you're only focused on the behavior of, say, an electron, then you can only talk about the relationships the electron bears to other particles or fields. You can't say anything about what philosophers like to call the *intrinsic nature* of the electron, how the electron is in and of itself.

Contrast an electron with a chess piece. What might you want to know about a chess piece? You might want to know what it does (if it's a king, it moves one space in any direction). But you might also want to know what it's like in and of itself (is it made of wood or plastic?). What is its intrinsic nature independently of its behavior? Similarly, you might be very interested to know what physicists have to say about the behavior of the electron, but you might also want to know what the electron is in and of itself. What is its intrinsic nature independent of its behavior? It turns out there's this huge hole in the center of our scientific worldview. Physics—and physical science more generally—tells us lots of stuff about the behavior of matter, but it's completely silent on its intrinsic nature. So what does this have to do with consciousness? The genius of Russell and Eddington was to bring together two problems that, on the face of it, have nothing to do with each other—the problem of consciousness and the problem of intrinsic natures.

The problem of consciousness is this challenge of finding a place for consciousness in our scientific worldview. The problem of intrinsic natures is that we have this huge hole in our scientific worldview. The solution is to put consciousness in the hole. The resulting theory is that there's just matter. This is not dualism, there's nothing spiritual or supernatural. Matter can be described from two perspectives. Physical science describes matter from the outside, in terms of its behavior. But from the inside, in terms of its intrinsic nature, matter is constituted of forms of consciousness. This is a form of panpsychism, the ancient view that consciousness is a fundamental and ubiquitous feature of matter. This has new agey connotations that some people feel a bit uncomfortable with, but we should judge a view not by its cultural associations

but by its explanatory power. What this Russell-Eddington panpsychism offers us is a way of integrating consciousness into our scientific worldview. We know that consciousness exists. Nothing is more evident than the reality of our feelings and experiences. We have to fit it into the scientific story somehow. The Russell-Eddington panpsychist view offers us a beautifully simple, elegant, unified way of integrating consciousness into our scientific worldview, and in a way that, unlike dualism, is completely consistent with everything we know about the brain scientifically.

Apart from it feeling a bit funny, this is a wonderful way of bringing consciousness into science. But of course, it's just a first step. The Russell-Eddington panpsychism is not a final theory of consciousness; it's a framework for making theoretical progress, just as Darwin's principle of natural selection was a framework for making theoretical progress. This is a theory in which we can make progress. It's going to take decades or centuries of interdisciplinary labor to try and fill in some of those details. I'd like to try to get this out to a broader audience, and to scientists as well. It's becoming more widely known in philosophy, but it's still pretty much unknown outside the ivory tower of academic philosophy. I want to get the idea out there more generally so we can work on it as a scientific community.

Philosophy is crucial when the science is not fully formed or when we haven't worked out how to make the problem tractable. At this stage, at least, it's important to distinguish the more empirical observational aspects of the science of consciousness and the more theoretical philosophical aspects of the science of consciousness.

Just focusing on the empirical aspect, neuroscience is absolutely crucial for a science of consciousness, but in my view, you can't get a science of consciousness just by doing neuroscience. Neuroscience essentially gives you correlations. You can scan people's brains and ask them what they're feeling and experiencing. And you can discover that the feeling of hunger is correlated with a certain kind of activity in the hypothalamus. Or you can think, like Giulio Tononi proposed, that consciousness in general is correlated with maximal integrated information. This is what neuroscience gives us—this wonderful body of correlations. But that in itself isn't a science of consciousness because we then want to know *why* you get a feeling of hunger when you have this activity in the hypothalamus. Why should that be? That's where you get to the more theoretical aspect. As soon as you start trying to explain those correlations, you're moving beyond what can be, in any straightforward sense, settled empirically. You're essentially doing philosophy. That's true whether you're a materialist, or a dualist, or a panpsychist. Some people think that materialists are just doing neuroscience and will solve the problem that way. You can't get a science of consciousness just by doing neuroscience, though neuroscience is an absolutely crucial preliminary to a theory of consciousness.

How do we do the more theoretical aspect? Some people think the materialists are just doing the neuroscience, and the philosophers are doing all this other weird stuff, like panpsychism or dualism. If you're a materialist, you get your correlations, but then you've got a big theoretical philosophical problem. How do you bridge the gap from the purely quantitative properties of neuroscience to the qualities of consciousness? No one's ever found a way of making any progress, in my view, on bridging that gap. It's not a gap you can just do more neuroscience to solve. You've got to do some philosophy. The truth is, every theory of consciousness has deep theoretical problems that we need to philosophize and to try and solve. That's true for the materialists as much as anyone else.

To my mind, the problems panpsychism faces just look to be more tractable than the problems

materialism faces. Compare it to physics. In physics, we're used to distinguishing between empirical observational physicists and theoretical physicists. We see that both have a role. In quantum mechanics, for example, we've got the equations that are empirically confirmed, but then no one knows what the hell they imply about reality. And so we've got a more theoretical task of trying to assess these different speculative models of quantum mechanics.

For some reason when it comes to consciousness, a lot of people think that we should just be doing neuroscience, that there's no room for anything theoretical. But if we ever want to move beyond the correlations that neuroscience gives us, we better sit in an armchair and do some theorizing. That's what philosophers are good at.

There's been a lot of change in the last forty years. Consciousness used to be a taboo topic on which you couldn't do serious science. It's now broadly accepted as a problem we need to address. The next stage is thinking of consciousness as a datum in its own right, consciousness as we're immediately aware of it.

People talk about the grand unified theory that physics is aiming for. If we one day have a theory that can account for all the data of observation and experiments, but it can't account for consciousness, then it can't be true because it's incomplete. What I admire about Dan Dennett is he understands this and he just denies it as a datum, which is completely consistent. Humphrey as well.

There are three categories of people. You can think of David Chalmers and me on the one hand who think that it is a real datum. It's in some sense extra to empirical data. So, we've got the empirical data of third-person observation experiment, and we've got this other thing to account for—consciousness—that we're immediately aware of. We need to rethink and expand science. That's one view. Dennett, at the other extreme, says that we need to account for the behavior and the empirical observable facts of the mind, but there's no extra datum. That's consistent as well. I would say that most people are in the middle, arguing that we do need to give a theory of consciousness, but we just need to carry on doing neuroscience and it will happen. That middle position doesn't make any sense. If you think there is this extra datum, then you're going to end up with correlations from doing neuroscience and you need to somehow explain those correlations.

Another way of putting it is that consciousness is unobservable. You can't look inside someone's head and see their feelings and experiences. Science deals with unobservable things, but it postulates unobservable things in order to explain what we can observe. What's unique about consciousness is that the thing we're trying to explain is unobservable. That's one way of seeing that a radically new approach to science is called for when the datum we're trying to account for is itself unobservable.

One underexplored question in the science of consciousness is the relationship between thought and consciousness. You can see this because the dominant theories of thought from the 20th century by Donald Davidson or Jerry Fodor have absolutely nothing to say about consciousness. People thought you could give a theory of thought, or what's more broadly called "intentionality," which just means mental representation, without mentioning consciousness at all. And you see someone like, for example, Andy Clark, who's done interesting stuff on mental representation, what we call "content," without discussing consciousness. Andy Clark once said to me, "Just deal with content, and consciousness will look after itself." It's a nice line. I call this view *separatism*, that thought and consciousness are

completely different. We can deal with thought without thinking about consciousness at all. But there's now a growing minority of philosophers—I happen to be amongst them—who think that thought is a kind of consciousness. These are people who believe in cognitive consciousness, often called *cognitive phenomenology* (phenomenology just being a synonym for consciousness). These people think that, as well as familiar sensory consciousness—colors, sounds, smells—there's also cognitive experience, cognitive consciousness, the experience of worrying that climate change is irrevocable. They think that's a kind of experience. So when you're sitting there wondering if climate change is irrevocable, you're having a certain kind of cognitive experience, and your having of that thought is constituted by your having this kind of cognitive experience.

So, which of these views is true has dramatic implications for AI. Think about Commander Data from *Star Trek*. Suppose, for the sake of discussion, you can't make something conscious from silicon. Suppose, for the sake of discussion, you need warm, wet, fleshy stuff to get consciousness. Commander Data is made of silicon, he's not conscious, but he's a behavioral functional duplicate of a human being. He talks as though he has thoughts, you might find him articulating in great detail about the problems of a globalized economy, advocating a Keynesian solution. The question is, does Commander Data really understand economics? Does he really have opinions on economics? Or is he just parroting words? Is he just a complicated mechanism set up to behave as though he has thoughts? If you're on the side of the cognitive consciousness people, you're going to say, "No, he's not conscious. You need cognitive consciousness to have thought and understanding. He's a faker. He's acting as though he has thoughts, but he doesn't really." Whereas, if you're a separatist, if you think thought has nothing to do with consciousness, then you're probably going to think Commander Data does have thought. He's not conscious, but you don't need consciousness to have thoughts. And you're probably going to think thoughts have something to do with complex behavioral functioning. This question is so little discussed.

This is another role that the philosophers have to contribute—pointing out, for example, this question about the relationship between thought and consciousness that is glossed over. People make assumptions one way or another without realizing that there's a point of controversy. It also has implications in general for a theory of consciousness, because whether you think consciousness is just to do with sensory experience or whether you think it's involved in cognition, that's going to make a real difference to which bits of the brain you're looking for for the neural correlates of consciousness. Jesse Prinz's very interesting theory of consciousness, for example, is completely dependent on a presupposition that cognitive consciousness doesn't exist.

I was obsessed with the problem of consciousness from day one as a philosophy undergraduate at eighteen. When I was an undergraduate, we were told that the only two options on consciousness were materialism on the one hand, dualism on the other. So, I tried to find out everything I could about these two options. I initially decided I was a materialist and defended that with great vigor. But I slowly came to worry about the clash between the purely quantitative language of physical science and the qualities that seem to essentially characterize conscious experience. When I finished my undergraduate degree, I thought the problem was irresolvable. I wrote my third-year dissertation on how the problem is irresolvable, and I went off and did something else. While I was doing something else, trying not to think about consciousness, I came across the paper by Thomas Nagel from the 1970s, "Panpsychism," which was not something I'd learned about as an undergraduate. I hadn't realized there was this middle way option that sounded a bit crazy but seemed to avoid the deep difficulties facing

dualism on the one hand and materialism on the other.

I decided I wanted to do graduate study. I did graduate study with Galen Strawson. There weren't many universities that had a panpsychist professor. Galen Strawson was one, in the University of Redding. Fifteen years ago, panpsychism was laughed at insofar as it was thought about at all. There has been a big change within academic philosophy, partly due to the rediscovery of these ideas by Russell and Eddington, partly because of Giulio Tononi's integrated information theory, which seems to have panpsychist implications.

Perhaps the most pressing problem or challenge for a panpsychist research program is what's become known as the combination problems. This is roughly the challenge of how to get from facts about the consciousness of particles to how to get to facts about human or animal consciousness, which is ultimately what we want to explain. There are some interesting proposals about how to make progress on this. Luke Roelofs, for example, is a research fellow at the University of Bochum in Germany whose work focuses on whether split-brain cases might help to shed light on mental combination. These are patients who've had the corpus callosum, the part of the brain that connects the two hemispheres, severed. This is a rather radical treatment for severe epilepsy. And it results in a peculiar fragmentation of consciousness. It seems as though these people end up having two conscious minds in one brain. The interest for panpsychists is that it looks like split-brain patients are the inverse of mental combination. In mental combination, we're looking for distinct conscious minds coming together to make a unified conscious mind. In split-brain patients, we've got a single conscious mind fragmenting into multiple conscious minds. Luke Roelofs thought that if we can get a grip on what's going on in split-brain cases and reverse engineer that, then maybe we could get a grip on how to think about metal combination.

One other approach is to postulate basic principles of nature to bridge the gap between facts about particle consciousness and facts about human consciousness. This is sometimes called emergentist panpsychism. One leading figure here is Hedda Hassel Mørch, who's a research fellow at the University of Oslo. She spent a year in the lab of Giulio Tononi trying to interpret the integrated information theory in an emergentist panpsychist model. The integrated information theory proposes that consciousness is correlated with maximal integrated information.

I don't think that's a complete theory of consciousness. It's a claim about correlation. But what Hedda Hassel Mørch does is interpret that in an emergentist panpsychism framework. The result is that we postulate a basic law of nature in which you get consciousness at the level where there's most integrated information. I love the way, in that theory, that we've got the philosophical contribution and the neuroscientific contribution coming together to make a complete theory of consciousness. I've a lot of problems with it; I don't quite agree with it, but I think it's perhaps the closest we've got to a complete theory of consciousness. It seems to me the way forward: having the philosophical framework of Russell-Eddington panpsychism, trying to link that up with specific, concrete neuroscientific theories, and seeing where we end up. It might go nowhere, but we've got to try things out.

Dmitry Volkov, who's a founder of the Moscow Center for Consciousness, decided to organize a dozen philosophers and a dozen graduate students from Moscow State University to spend a week on a sailing ship in the Arctic. Most of the philosophers on board, like Dennett and Humphrey, were in some sense illusionists about consciousness, in some sense think consciousness doesn't really exist. For some official opposition, they also invited David

Chalmers, myself (the panpsychist) and Martine Nida-Rümelin (the dualist). Also onboard were Andy Clark, Patricia Churchland, Paul Churchland, Nick Humphrey. Most people on board were hardcore materialists, some even who deny the reality of consciousness. Myself, David Chalmers, and Martine Nida-Rümelin were invited along as the onboard opposition.

We had some good discussions. I even managed to persuade Daniel Dennett he was wrong about something, which is one of my proudest philosophical moments. Not about his whole worldview obviously—philosophers never change their minds. Well, that's not true, they do. About this quite specific, quite important issue of whether dualism is consistent with conservation of energy. I'm not a dualist. I think dualism is problematic for all sorts of reasons, but Dennett and Paul Churchland pushed this line that we can rule out dualism on the basis of conservation of energy. The rough thought is, if there's an immaterial mind impacting on the brain, that's going to add energy to the physical system in violation of the principle that energy is never created or destroyed in a closed system.

Dualists like David Chalmers, for example, postulate these basic psychophysical laws of nature. As well as the laws of physics, they think there are these basic psychophysical laws of nature that relate the physical world to consciousness. They could just hold that those laws respect the conservation of energy. On our current standard model of physics, there are multiple laws of nature that all work together to respect the conservation of energy. Why not the psychophysical laws as well? I raised this in Paul Churchland's talk, and I got a very fiery response. One of the Moscow graduate students said, "They turned on you like a pack of wolves!" I had a vigorous debate with Paul Churchland, but then that evening, most people went off the boat to go on an island, but me and Dennett stayed on board. I just kept saying, "I'm not saying dualism is plausible. There are all sorts of problems, but this specific problem is consistent with the conservation of energy." And in the end, though he might deny this now, he said, "Maybe that's right."

What seems to me a hugely underexplored question in the debate on artificial intelligence is the relationship between thought and consciousness; thought, or mental representation more generally, and consciousness. When I read people writing about AI, some of them seem to assume that thought has nothing to do with consciousness and we can just give an account of thought and mental representation without mentioning consciousness at all. This might be an Andy Clark position. Others on the more John Searle side think that if you don't have consciousness, you don't really have thought. You have a complex mechanism that behaves as though it has thought, but it doesn't really have thought. This is what's going on behind the Chinese room thought experiments.

People are clashing without realizing it. What I want to say is going on in the background here is a discussion of what the relationship between thought and consciousness is. Is thought a kind of consciousness, a kind of cognitive experience? Or is thought just something completely different to consciousness? This debate is on whether there is such a thing as cognitive consciousness, whether thought is a kind of experience. It ought to be possible to settle that just by reflecting on our own consciousness. We ought to be able to just introspect and see whether there's such a thing as cognitive consciousness, but for some strange reason when you ask people to do that, 50% of philosophers claim "Yeah. Obviously, there's cognitive experience when I'm wondering about whether I've left my keys at home." That's a kind of experience to wonder about where your keys are. Other philosophers, Jesse Prinz, for example, say, "No. When I introspect, I just find colors, sounds, shapes, emotions—that exhausts my consciousness. There's none of this cognitive consciousness. I just don't find that

at all." It's hard to know how to settle this issue. It's a debate about consciousness as we immediately experience it.