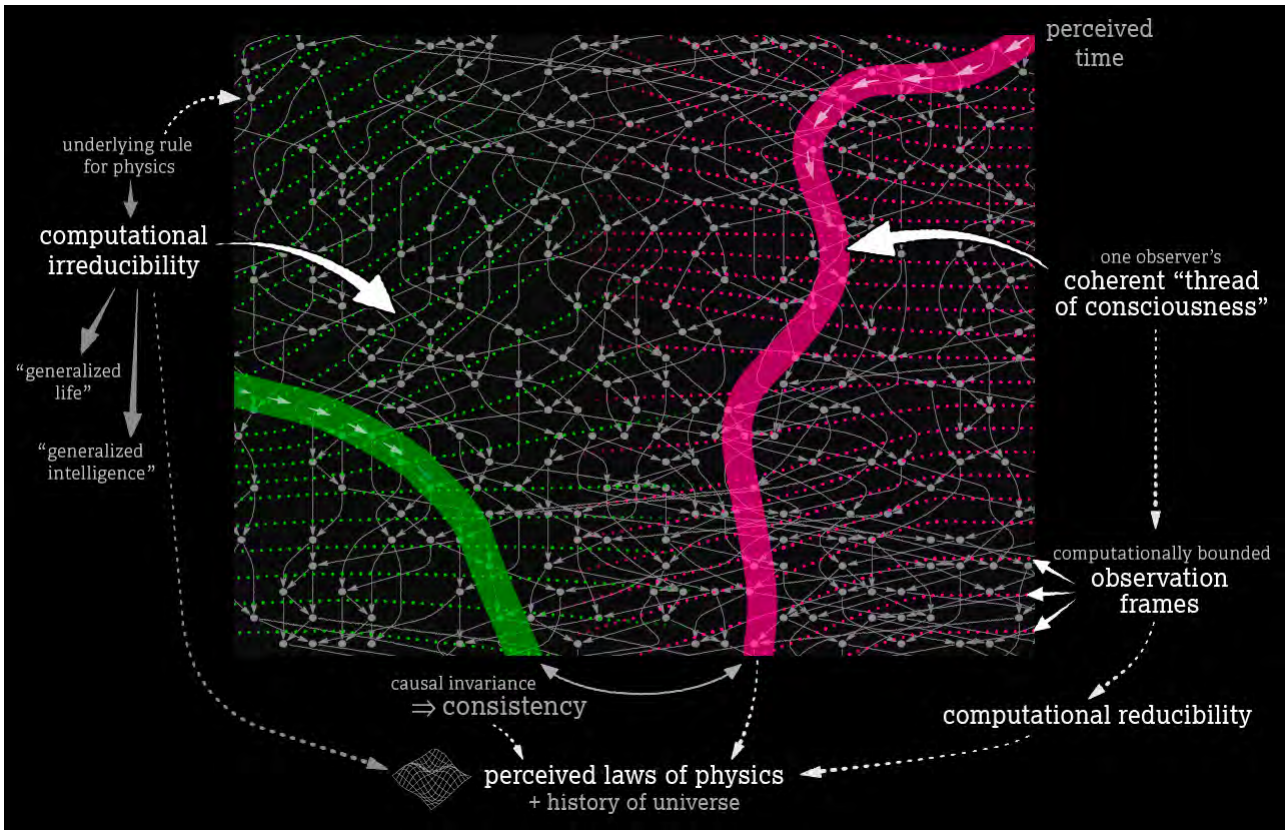
What Is Consciousness? Some New Perspectives from Our Physics Project

March 22, 2021



"What about Consciousness?"

For years I've batted it away. I'll be talking about my discoveries in the computational universe, and computational irreducibility, and my Principle of Computational Equivalence, and people will ask "So what does this mean about consciousness?" And I'll say "that's a slippery topic". And I'll start talking about the sequence: life, intelligence, consciousness.

I'll ask "What is the abstract definition of life?" We know about the case of life on Earth, with all its RNA and proteins and other implementation details. But how do we generalize?

Contents

the Principle of Computational Equivalence says happens all over the place. Then I'll talk about intelligence. And I'll argue it's the same kind of thing. We know the case of human intelligence. But if we generalize, it's just computational sophistication—and it's ubiquitous. And so it's perfectly reasonable to say that "the weather has a mind of its own"; it just happens to be a mind whose details and "purposes" aren't aligned with our existing human experience.

I've always implicitly assumed that consciousness is just a continuation of the same story: something that, if thought about in enough generality, is just a feature of computational sophistication, and therefore quite ubiquitous. But from our Physics Project—and particularly from thinking about its implications for the foundations of quantum mechanics—I've begun to realize that at its core consciousness is actually something rather different. Yes, its implementation involves computational sophistication. But its essence is not so much about what can happen as about having ways to integrate what's happening to make it somehow coherent and to allow what we might see as "definite thoughts" to be formed about it.

And rather than consciousness being somehow beyond "generalized intelligence" or general computational sophistication, I now instead see it as a kind of "step down"—as something associated with simplified descriptions of the universe based on using only bounded amounts of computation. At the outset, it's not obvious that a notion of consciousness defined in this way could consistently exist in our universe. And indeed the possibility of it seems to be related to deep features of the formal system that underlies physics.

In the end, there's a lot going on in the universe that's in a sense "beyond consciousness". But the core notion of consciousness is crucial to our whole way of seeing and describing the universe—and at a very fundamental level it's what makes the universe seem to us to have the kinds of laws and behavior it does.

Consciousness is a topic that's been discussed and debated for centuries. But the surprise to me is that with what we've learned from exploring the computational universe and especially from our recent Physics Project it seems there may be new perspectives to be had, which most significantly seem to have the potential to connect questions about consciousness to concrete, formal scientific ideas.

Contents

foundations of physics—is quite conceptually complex, and all I'll try to do here is sketch some preliminary ideas. No doubt quite a bit of what I say can be connected to existing philosophical and other thinking, but so far I've only had a chance to explore the ideas themselves, and haven't yet tried to study their historical context.

Observers and Their Physics

The universe in our models is full of sophisticated computation, all the way down. At the lowest level it's just a giant collection of "atoms of space", whose relationships are continually being updated according to a computational rule. And inevitably much of that process is computationally irreducible, in the sense that there's no general way to "figure out what's going to happen" except, in effect, by just running each step.

But given that, how come the universe doesn't just seem to us arbitrarily complex and unpredictable? How come there's order and regularity that we can perceive in it? There's still plenty of computational irreducibility. But somehow there are also pockets of reducibility that we manage to leverage to form a simpler description of the world, that we can successfully and coherently make use of. And a fundamental discovery of our Physics Project is that the two great pillars of twentieth-century physics—general relativity and quantum mechanics—correspond precisely to two such pockets of reducibility.

There's an immediate analog—that actually ends up being an example of the same fundamental computational phenomenon. Consider a gas, like air. Ultimately the gas consists of lots of molecules bouncing around in a complicated way that's full of computational irreducibility. But it's a central fact of statistical mechanics that if we look at the gas on a large scale, we can get a useful description of what it does just in terms of properties like temperature and pressure. And in effect this reflects a pocket of computational reducibility, that allows us to operate without engaging with all the computational irreducibility underneath.

How should we think about this? An idea that will generalize is that as "observers" of the gas, we're conflating lots of different microscopic configurations of molecules, and just paying attention to overall aggregate properties. In the language of statistical mechanics, it's effectively a story of "coarse graining". But within our computational approach, there's now a clear, computational way to characterize this. At the level of individual molecules

Contents

observer is doing a computation. But the crucial point is that if there's a certain boundedness to that computation then this has immediate consequences for the effective behavior the observer will perceive. And in the case of something like a gas, it turns out to directly imply the Second Law of Thermodynamics.

In the past there's been a certain amount of mystery around the origin and validity of the Second Law. But now we can see it as a consequence of the interplay between underlying computational irreducibility and the computational boundedness of observers. If the observer kept track of all the computationally irreducible motions of individual molecules, they wouldn't see Second Law behavior. The Second Law depends on a pocket of computational reducibility that in effect emerges only when there's a constraint on the observer that amounts to the requirement that the observer has a "coherent view" of what's going on.

So what about physical space? The traditional view had been that space was something that could to a large extent just be described as a coherent mathematical object. But in our models of physics, space is actually made of an immense number of discrete elements whose pattern of interconnections evolves in a complex and computationally irreducible way. But it's much like with the gas molecules. If an observer is going to form a coherent view of what's going on, and if they have bounded computational capabilities, then this puts definite constraints on what behavior they will perceive. And it turns out that those constraints yield exactly relativity.

In other words, for the "atoms of space", relativity is the result of the interplay between underlying computational irreducibility and the requirement that the observer has a coherent view of what's going on.

It may be helpful to fill in a little more of the technical details. Our underlying theory basically says that each elementary element of space follows computational rules that will yield computationally irreducible behavior. But if that was all there was to it, the universe would seem like a completely incoherent place, with every part of it doing irreducibly unpredictable things.

But imagine there's an observer who perceives coherence in the universe. And who, for example, views there as being a definite coherent notion of "space". What can we say about such an observer? The first thing is that since our model is supposed to describe everything

Contents

embedded part of the system—made up of the same atoms of space, and following the same rules, as everything else.

And there's an immediate consequence to this. From "inside" the system there are only certain things about the system that the observer can perceive. Let's say, for example, that in the whole universe there's only one point at which anything is updated at any given time, but that that "update point" zips around the universe (in "Turing machine style"), sometimes updating a piece of the observer, and sometimes updating something they were observing. If one traces through scenarios like this, one realizes that from "inside the system" the only thing the observer can ever perceive is causal relationships between events.

They can't tell "specifically when" any given event happens; all they can tell is what event has to happen before what other one, or in other words, what the causal relationships between events are. And this is the beginning of what makes relativity inevitable in our models.

But there are two other pieces. If the observer is going to have a coherent description of "space" they can't in effect be tracking each atom separately; they'll have to fit them into some overall framework, say by assigning each of them particular "coordinates", or, in the language of relativity, defining a "reference frame" that conflates many different points in space. But if the observer is computationally bounded, then this puts constraints on the structure of the reference frame: it can't for example be so wild that it separately traces the computationally irreducible behavior of individual atoms of space.

But let's say an observer has successfully picked some reference frame. What's to say that as the universe evolves it's still possible to consistently maintain that reference frame? Well, this relies on a fundamental property that we believe either directly or effectively defines the operation of our universe: what we call "causal invariance". The underlying rules just describe possible ways that the connections between atoms of space can be updated. But causal invariance implies that whatever actual sequence of updatings is used, there must always be the same graph of causal relationships.

And it's this that gives observers the ability to pick different reference frames, and still have the same consistent and coherent perception of the behavior of the universe. And in the end, we have a definite result: that if there's underlying computational irreducibility—plus

Contents

computationally bounded way must inevitably perceive the universe to follow the laws of general relativity.

But—much like with the Second Law—this conclusion relies on having an observer who forms a coherent perception of the universe. If the observer could separately track every atom of space they won't "see general relativity"; that only emerges for an observer who forms a coherent perception of the universe.

The Quantum Observer

OK, so what about quantum mechanics? How does that relate to observers? The story is actually surprisingly similar to both the Second Law and general relativity: quantum mechanics is again something that emerges as a result of trying to form a coherent perception of the universe.

In ordinary classical physics one considers everything that happens in the universe to happen in a definite way, in effect defining a single thread of history. But the essence of quantum mechanics is that actually there are many threads of history that are followed. And an important feature of our models is that this is inevitable.

The underlying rules define how local patterns of connections between atoms of space should be updated. But in the hypergraph of connections that represents the universe there will in general be many different places where the rules can be applied. And if we trace all the possibilities we get a multiway graph that includes many possible threads of history, sometimes branching and sometimes merging.

So how will an observer perceive all this? The crucial point is that the observer is themselves part of this multiway system. So in other words, if the universe is branching, so is the observer. And in essence the question becomes how a "branching brain" will perceive a branching universe.

It's fairly easy to imagine how an observer who is "spatially large" compared to individual molecules in a gas—or atoms of space—could conflate their view of these elements so as to perceive only some aggregate property. Well, it seems like very much the same kind of thing is going on with observers in quantum mechanics. It's just that instead of being extended in physical space, they're extended in what we call branchial space.

Contents

slicing through this graph at a particular level that in effect corresponds to a particular time. In that slice there will be a certain set of nodes of the multiway graph, representing possible states of the system. And the structure of the multiway graph then defines relationships between these states (say through common ancestry). And in a large-scale limit we can say that the states are laid out in branchial space.

In the language of quantum mechanics, the geometry of branchial space in effect defines a map of entanglements between quantum states, and coordinates in branchial space are like phases of quantum amplitudes. In the evolution of a quantum system, one might start from a certain bundle of quantum states, then follow their threads of history, looking at where in branchial space they go.

But what would a quantum observer perceive about this? Even if they didn't start that way, over time a quantum observer will inevitably become spread out in branchial space. And so they'll always end up sampling a whole region in branchial space, or a whole bundle of "threads of history" in the multiway graph.

What will they make of them? If they considered each of them separately no coherent picture would emerge, not least since the underlying evolution of individual threads of history can be expected to be computationally irreducible. But what if the observer just defines their way of viewing things to be one that systematically organizes different threads of history, say by conflating "computationally nearby" ones? It's similar to setting up a reference frame in relativity, except that now the coherent representation that this "quantum frame" defines is of branchial space rather than physical space.

But what will this coherent representation be like? Well, it seems to be exactly quantum mechanics as it was developed over the past century. In other words, just like general relativity emerges as an aggregate description of physical space formed by a computationally bounded observer, so quantum mechanics emerges as an aggregate description of branchial space.

Does the observer "create" the quantum mechanics? In some sense, yes. Just as in the spacetime case, the multiway graph has all sorts of computationally irreducible things going on. But if there's an observer with a coherent description of what's going on, then their description must follow the laws of quantum mechanics. Of course, there are lots of other things going on too—but they don't fit into this coherent description.

Contents

different threads of history to get a coherent description of what's going on. How will their description correlate with what another observer—with a different quantum frame—would perceive? In the traditional formalism of quantum mechanics it's always been difficult to explain why different observers—making different measurements—still fundamentally perceive the universe to be working the same.

In our model, there's a clear answer: just like in the spacetime case, if the underlying rules show causal invariance, then regardless of the frame one uses, the basic perceived behavior will always be the same. Or, in other words, causal invariance guarantees the consistency of the behavior deduced by different observers.

There are many technical details to this. The traditional formalism of quantum mechanics has two separate parts. First, the time evolution of quantum amplitudes, and second, the process of measurement. In our models, there's a very beautiful correspondence between the phenomenon of motion in space and the evolution of quantum amplitudes. In essence, both are associated with the deflection of (geodesic) paths by the presence of energy-momentum. But in the case of motion this deflection (that we identify as the effect of gravity) happens in physical space, while in the quantum case the deflection (that we identify as the phase change specified by the path integral) happens in branchial space. (In other words, the Feynman path integral is basically just the direct analog in branchial space of the Einstein equations in physical space.)

OK, so what about quantum measurement? Doing a quantum measurement involves somehow taking many threads of history (corresponding to a superposition of many quantum states) and effectively reducing them to a single thread that coherently represents the "outcome". A quantum frame defines a way to do this—in effect specifying the pattern of threads of history that should be conflated. In and of itself, a quantum frame—like a relativistic reference frame—isn't a physical thing; it just defines a way of describing what's going on.

But as a way of probing possible coherent representations that an observer can form, one can consider what happens if one formally conflates things according to a particular quantum frame. In an analogy where the multiway graph defines inferences between propositions in a formal system, conflating things is like "performing certain completions". And each completion is then like an elementary step in the act of measurement. And by

Contents

Quantum Mechanics" suggested by Jonathan Gorard.

Assuming that the underlying rule for the universe ultimately shows causal invariance, doing these completions is never fundamentally necessary, because different threads of history will always eventually give the same results for what can be perceived within the system. But if we want to get a "possible snapshot" of what the system is doing, we can pick a quantum frame and formally do the completions it defines.

Doing this doesn't actually "change the system" in a way that we would "see from outside". It's only that we're in effect "doing a formal projection" to see how things would be perceived by an observer who's picked a particular quantum frame. And if the observer is going to have a coherent perception of what's going on, they in effect have to have picked some specific quantum frame. But then from the "point of view of the observer" the completions associated with that frame in some sense "seem real" because they're the way the observer is accessing what's going on.

Or, in other words, the way a computationally bounded "branching brain" can have a coherent perception of a "branching universe" is by looking at things in terms of quantum frames and completions, and effectively picking off a computationally reducible slice of the whole computationally irreducible evolution of the universe—where it then turns out that the slice must necessarily follow the laws of quantum mechanics.

So, once again, for a computationally bounded observer to get a coherent perception of the universe—with all its underlying computational irreducibility—there's a strong constraint on what that perception can be. And what we've discovered is that it turns out to basically have to follow the two great core theories of twentieth-century physics: general relativity and quantum mechanics.

It's not immediately obvious that there has to be any way to get a coherent perception of the universe. But what we now know is that if there is, it essentially forces specific major results about physics. And, of course, if there wasn't any way to get a coherent perception of the universe there wouldn't really be systematic overall laws, or, for that matter, anything like physics, or science as we know it.

So, What Is Consciousness?

Contents

that we even have a notion of "experiencing" it at all is special. The world is doing what it does, with all sorts of computational irreducibility. But somehow even with the computationally bounded resources of our brains (or minds) we're able to form some kind of coherent model of what's going on, so that, in a sense, we're able to meaningfully "form coherent thoughts" about the universe. And just as we can form coherent thoughts about the universe, so also we can form coherent thoughts about that small part of the universe that corresponds to our brains—or to the computations that represent the operation of our minds.

But what does it mean to say that we "form coherent thoughts"? There's a general notion of computation, which the Principle of Computational Equivalence tells us is quite ubiquitous. But it seems that what it means to "form coherent thoughts" is that computations are being "concentrated down" to the point where a coherent stream of "definite thoughts" can be identified in them.

At the outset it's certainly not obvious that our brains—with their billions of neurons operating in parallel—should achieve anything like this. But in fact it seems that our brains have a quite specific neural architecture—presumably produced by biological evolution—that in effect attempts to "integrate and sequentialize" everything. In our cortex we bring together sensory data we collect, then process it with a definite thread of attention. And indeed in medical settings observed deficits in this are what are normally used to identify absence of levels of consciousness. There may still be neurons firing but without integration and sequentialization there doesn't really seem to be what we normally consider consciousness.

These are biological details. But they seem to point to a fundamental feature of consciousness. Consciousness is not about the general computation that brains—or, for that matter, many other things—can do. It's about the particular feature of our brains that causes us to have a coherent thread of experience.

But what we have now realized is that the notion of having a coherent thread of experience has deep consequences that far transcend the details of brains or biology. Because in particular what we've seen is that it defines the laws of physics, or at least what we consider the laws of physics to be.

Contents

single case of humans. But just as we've seen that the notion of intelligence can be generalized to the notion of arbitrary sophisticated computation, so now it seems that the notion of consciousness can be generalized to the notion of forming a coherent thread of representation for computations.

Operationally, there's potentially a rather straightforward way to think about this, though it depends on our recent understanding of the concept of time. In the past, time in fundamental physics was usually viewed as being another dimension, much like space. But in our models of fundamental physics, time is something quite different from space. Space corresponds to the hypergraph of connections between the elements that we can consider as "atoms of space". But time is instead associated with the inexorable and irreducible computational process of repeatedly updating these connections in all possible ways.

There are definite causal relationships between these updating events (ultimately defined by the multiway causal graph), but one can think of many of the events as happening "in parallel" in different parts of space or on different threads of history. But this kind of parallelism is in a sense antithetical to the concept of a coherent thread of experience.

And as we've discussed above, the formalism of physics—whether reference frames in relativity or quantum mechanics—is specifically set up to conflate things to the point where there is a single thread of evolution in time.

So one way to think about this is that we're setting things up so we only have to do sequential computation, like a Turing machine. We don't have multiple elements getting updated in parallel like in a cellular automaton, and we don't have multiple threads of history like in a multiway (or nondeterministic) Turing machine.

The operation of the universe may be fundamentally parallel, but our "parsing" and "experience" of it is somehow sequential. As we've discussed above, it's not obvious that such a "sequentialization" would be consistent. But if it's done with frames and so on, the interplay between causal invariance and underlying computational irreducibility ensures that it will be—and that the behavior of the universe that we'll perceive will follow the core features of twentieth-century physics, namely general relativity and quantum mechanics.

But do we really "sequentialize" everything? Experience with artificial neural networks seems to give us a fairly good sense of the basic operation of brains. And, yes, something

Contents

to things we might realistically describe as "thoughts" the more sequential things seem to get. And a notable feature is that what seems to be our richest way to communicate thoughts, namely language, is decidedly sequential.

When people talk about consciousness, something often mentioned is "self-awareness" or the ability to "think about one's own processes of thinking". Without the conceptual framework of computation, this might seem quite mysterious. But the idea of universal computation instead makes it seem almost inevitable. The whole point of a universal computer is that it can be made to emulate any computational system—even itself. And that is why, for example, we can write the evaluator for Wolfram Language in Wolfram Language itself.

The Principle of Computational Equivalence implies that universal computation is ubiquitous, and that both brains and minds, as well as the universe at large, have it. Yes, the emulated version of something will usually take more time to execute than the original. But the point is that the emulation is possible.

But consider a mind in effect thinking about itself. When a mind thinks about the world at large, its process of perception involves essentially making a model of what's out there (and, as we've discussed, typically a sequentialized one). So when the mind thinks about itself, it will again make a model. Our experiences may start by making models of the "outside world". But then we'll recursively make models of the models we make, perhaps barely distinguishing between "raw material" that comes from "inside" and "outside".

The connection between sequentialization and consciousness gives one a way to understand why there can be different consciousnesses, say associated with different people, that have different "experiences". Essentially it's just that one can pick different frames and so on that lead to different "sequentialized" accounts of what's going on.

Why should they end up eventually being consistent, and eventually agreeing on an objective reality? Essentially for the same reason that relativity works, namely that causal invariance implies that whatever frame one picks, the causal graph that's eventually traced out is always the same.

If it wasn't for all the interactions continually going on in the universe, there'd be no reason for the experience of different consciousnesses to get aligned. But the interactions—with

Contents

consistency that's needed, and, as we've discussed, something else too: particular effective laws of physics, that turn out to be just the relativity and quantum mechanics we know.

Other Consciousnesses

The view of consciousness that we've discussed is in a sense focused on the primacy of time: it's about reducing the "parallelism" associated with space—and branchial space—to allow the formation of a coherent thread of experience, that in effect occurs sequentially in time.

And it's undoubtedly no coincidence that we humans are in effect well placed in the universe to be able to do this. In large part this has to do with the physical sizes of things—and with the (undoubtedly not coincidental) fact that human scales are intermediate between those at which the effects of either relativity or quantum mechanics become extreme.

Why can we "ignore space" to the point where we can just discuss things happening "wherever" at a sequence of moments in time? Basically it's because the speed of light is large compared to human scales. In our everyday lives the important parts of our visual environment tend to be at most tens of meters away—so it takes light only tens of nanoseconds to reach us. Yet our brains process information on timescales measured in milliseconds. And this means that as far as our experience is concerned, we can just "combine together" things at different places in space, and consider a sequence of instantaneous states in time.

If we were the size of planets, though, this would no longer work. Because—assuming our brains still ran at the same speed—we'd inevitably end up with a fragmented visual experience, that we wouldn't be able to think about as a single thread about which we can say "this happened, then that happened".

Even at standard human scale, we'd have somewhat the same experience if we used for example smell as our source of information about the world (as, say, dogs to a large extent do). Because in effect the "speed of smell" is quite slow compared to brain processing. And this would make it much less useful to identify our usual notion of "space" as a coherent concept. So instead we might invent some "other physics", perhaps labeling things in terms of the paths of air currents that deliver smells to us, then inventing some elaborate gauge-field-like construct to talk about the relations between different paths.

Contents

brains are small and slow enough that they're not limited by the speed of light, which is why it's possible for them to "form coherent thoughts" in the first place. If our brains were the size of planets, it would necessarily take far longer than milliseconds to "come to equilibrium", so if we insisted on operating on those timescales there'd be no way—at least "from the outside"—to ensure a consistent thread of experience.

From "inside", though, a planet-size brain might simply assume that it has a consistent thread of experience. And in doing this it would in a sense try to force a different physics on the universe. Would it work? Based on what we currently know, not without at least significantly changing the notions of space and time that we use.

By the way, the situation would be even more extreme if different parts of a brain were separated by permanent event horizons. And it seems as if the only way to maintain a consistent thread of experience in this case would be in effect to "freeze experience" before the event horizons formed.

What if we and our brains were much smaller than they actually are? As it is, our brains may contain perhaps 10^{300} atoms of space. But what if they contained, say, only a few hundred? Probably it would be hard to avoid computational irreducibility—and we'd never even be able to imagine that there were overall laws, or generally predictable features of the universe, and we'd never be able to build up the kind of coherent experience needed for our view of consciousness.

What about our extent in branchial space? In effect, our perception that "definite things happen even despite quantum mechanics" implies a conflation of the different threads of history that exist in the region of branchial space that we occupy. But how much effect does this have on the rest of the universe? It's much like the story with the speed of light, except now what's relevant is a new quantity that appears in our models: the maximum entanglement speed. And somehow this is large enough that over "everyday scales" in branchial space it's adequate for us just to pick a quantum frame and treat it as something that can be considered to have a definite state at any given instant in time—so that we can indeed consistently maintain a "single thread of experience".

OK, so now we have a sense of why with our particular human scale and characteristics our view of consciousness might be possible. But where else might consciousness be possible?

Contents

be able to build up something that "viewed from the inside" represents a coherent thread of experience. But the issue is that we're in effect "on the outside". We know about our human thread of experience. And we know about the physics that effectively follows from it. And we can ask how we might experience that if, for example, our sensory systems were different. But to truly "get inside" we have to be able to imagine something very alien. Not only different sensory data and different "patterns of thinking", but also different implied physics.

An obvious place to start in thinking about "other consciousnesses" is with animals and other organisms. But immediately we have the issue of communication. And it's a fundamental one. Perhaps one day there'll be ways for various animals to fluidly express themselves through something like human-relatable videogames. But as of now we have surprisingly little idea how animals "think about things", and, for example, what their experience of the world is.

We can guess that there will be many differences from ours. At the simplest level, there are organisms that use different sensory modalities to probe the world, whether those be smell, sound, electrical, thermal, pressure, or other. There are "hive mind" organisms, where whatever integrated experience of the world there may be is built up through slow communication between different individuals. There are organisms like plants, which are (quite literally) rooted to one place in space. There are also things like viruses where anything akin to an "integrated thread of experience" can presumably only emerge at the level of something like the progress of an epidemic.

Meanwhile, even in us, there are things like the immune system, which in effect have some kind of "thread of experience" though with rather different input and output than our brains. Even if it seems bizarre to attribute something like consciousness to the immune system, it is interesting to try to imagine what its "implied physics" would be.

One can go even further afield, and think about things like the complete tree of life on Earth, or, for that matter, the geological history of the Earth, or the weather. But how can these have anything like consciousness? The Principle of Computational Equivalence implies that all of them have just the same fundamental computational sophistication as our brains. But, as we have discussed, consciousness seems to require something else as well: a kind of coherent integration and sequentialization.

Contents

patterns of fluid flow in the atmosphere. But—like fundamental processes in physics—it seems to be happening all over the place, with nothing, it seems, to define anything like a coherent thread of experience.

Coming a little closer to home, we can consider software and AI systems. One might expect that to "achieve consciousness" one would have to go further than ever before and inject some special "human-like spark". But I suspect that the true story is rather different. If one wants the systems to make the richest use of what the computational universe has to offer, then they should behave a bit like fundamental physics (or nature in general), with all sorts of components and all sorts of computationally irreducible behavior.

But to have something like our view of consciousness requires taking a step down, and effectively forcing simpler behavior in which things are integrated to produce a "sequentialized" experience. And in the end, it may not be that different from picking out of the computational universe of possibilities just what can be expressed in a definite computational language of the kind the Wolfram Language provides.

Again we can ask about the "implied physics" of such a setup. But since the Wolfram Language is modeled on picking out the computational essence of human thinking it's basically inevitable that its implied physics will be largely the same as the ordinary physics that is derived from ordinary human thinking.

One feature of having a fundamental model for physics is that it "reduces physics to mathematics", in the sense that it provides a purely formal system that describes the universe. So this raises the question of whether one can think about consciousness in a formal system, like mathematics.

For example, imagine a formal analog of the universe constructed by applying axioms of mathematics. One would build up an elaborate network of theorems, that in effect populate "metamathematical space". This setup leads to some fascinating analogies between physics and metamathematics. The notion of time effectively remains as always, but here represents the progressive proving of new mathematical theorems.

The analog of our spatial hypergraph is a structure that represents all theorems proved up to a given time. (And there's also an analog of the multiway graph that yields quantum

Contents

a theorem.) So what about things like reference frames?

Well, just as in physics, a reference frame is something associated with an observer. But here the observer is observing not physical space, but metamathematical space. And in a sense any given observer is "discovering mathematics in a particular order". It could be that all the different "points in metamathematical space" (i.e. theorems) are behaving in completely incoherent—and computationally irreducible—ways. But just as in physics, it seems that there's a certain computational reducibility: causal invariance implies that different reference frames will in a sense ultimately always "see the same mathematics".

There's an analog of the speed of light: the speed at which a new theorem can affect theorems that are progressively further away in metamathematical space. And relativistic invariance then becomes the statement that "there's only one mathematics"—but it can just be explored in different ways.

How does this relate to "mathematical consciousness"? The whole idea of setting up reference frames in effect relies on the notion that one can "sequentialize metamathematical space". And this in turn relies on a notion of "mathematical perception". The situation is a bit like in physics. But now one has a formalized mathematician whose mind stretches over a certain region of metamathematical space.

In current formalized approaches to mathematics, a typical "human-scale mathematical theorem" might correspond to perhaps 10^5 lowest-level mathematical propositions. Meanwhile, the "mathematician" might "integrate into their experience" some small fraction of the metamathematical universe (which, for human mathematics, is currently perhaps 3×10^6 theorems). And it's this setup—which amounts to defining a "sequentialized mathematical consciousness"—that means it makes sense to do analysis using reference frames, etc.

So, just as in physics, it's ultimately the characteristics of our consciousness that lead to the physics we attribute to the universe, so something similar seems to happen in mathematics.

Clearly we've now reached a quite high level of abstraction, so perhaps it's worth mentioning one more wrinkle that involves an even higher level of abstraction.

We've talked about applying a rule to update the abstract structure that represents the universe. And we've discussed the fact that the rule can be applied at different places, and

Contents

specific rule; we can consider all possible rules.

The result is a rulial multiway graph of possible states of the universe. On different paths, different specific rules are followed. And if you slice across the graph you can get a map of states laid out in rulial space, with different positions corresponding to the outcomes of applying different rules to the universe.

An important fact is then that at the level of the rulial multiway graph there is always causal invariance. So this means that different "rulial reference frames" must always ultimately give equivalent results. Or, in other words, even if one attributes the evolution of the universe to different rules, there is always fundamental equivalence in the results.

In a sense, this can be viewed as a reflection of the Principle of Computational Equivalence and the fundamental idea that the universe is computational. In essence it is saying that since whatever rules one uses to "construct the universe" are almost inevitably computation universal, one can always use them to emulate any other rules.

How does this relate to consciousness? Well, one feature of different rulial reference frames is that they can lead to utterly and incoherently different basic descriptions of the universe.

One of them could be our hypergraph-rewriting-based setup, with a representation of space that corresponds well with what emerged in twentieth-century physics. But another could be a Turing machine, in which one views the updating of the universe as being done by a single head zipping around to different places.

We've talked about some possible systems in which consciousness could occur. But one we haven't yet mentioned—but which has often been considered—is "extraterrestrial intelligences". Before our Physics Project one might reasonably have assumed that even if there was little else in common with such "alien intelligences", at least they would be "experiencing the same physics".

But it's now clear that this absolutely does not need to be the case. An alien intelligence could perfectly well be experiencing the universe in a different rulial reference frame, utterly incoherent with the one we use.

Is there anything "sequentializable" in a different rulial reference frame? Presumably it's possible to find at least something sequentializable in any rulial reference frame. But the

Contents

different one.

Does there need to be a "sequentializable consciousness" to imply "meaningful laws of physics"? Presumably meaningful laws have to somehow be associated with computational reducibility; certainly that would be true if they were going to be useful to a "computationally bounded" alien intelligence.

But it's undoubtedly the case that "sequentializability" is not the only way to access computational reducibility. In a mathematical analogy, using sequentializability is a bit like using ordinary mathematical induction. But there are other axiomatic setups (like transfinite induction) that define other ways to do things like prove theorems.

Yes, human-like consciousness might involve sequentializability. But if the general idea of consciousness is to have a way of "experiencing the universe" that accesses computational reducibility then there are no doubt other ways. It's a kind of "second-order alienness": in addition to using a different rulial reference frame, it's using a different scheme for accessing reducibility. And the implied physics of such a setup is likely to be very different from anything we currently think of as physics.

Could we ever expect to identify what some of these "alien possibilities" are? The Principle of Computational Equivalence at least implies that we can in principle expect to be able to set up any possible computational rule. But if we start doing experiments we can't have an expectation that scientific induction will work, and it is potentially arbitrarily difficult to identify computational reducibility. Yes, we might recognize some form of prediction or regularity that we are familiar with. But to recognize an arbitrary form of computational reducibility in effect relies on some analog of a definition of consciousness, which is what we were looking for in the first place.

What Now?

Consciousness is a difficult topic, that has vexed philosophers and others for centuries. But with what we know now from our Physics Project it at least seems possible to cast it in a new light much more closely connected to the traditions of formal science. And although I haven't done it here, I fully anticipate that it'll be possible to take the ideas I've discussed and use them to create formal models that can answer questions about consciousness and capture its connections, particularly to physics.

Contents

useful. Perhaps one will already be able to get worthwhile information about how branching brains perceive a branching universe by looking at some simple case of a multiway Turing machine. Perhaps some combinator system will already reveal something about how different versions of physics could be set up.

In a sense what's important is that it seems we may have a realistic way to formalize issues about consciousness, and to turn questions about consciousness into what amount to concrete questions about mathematics, computation, logic or whatever that can be formally and rigorously explored.

But ultimately the way to tether the discussion—and to have it not for example devolve into debates about the meaning of words—is to connect it to actionable issues and applications.

As a first example, let's discuss distributed computing. How should we think about computations that—like those in our model of physics—take place in parallel across many different elements? Well—except in very simple or structured cases—it's hard, at least for us humans. And from what we've discussed about consciousness, perhaps we can now understand why.

The basic issue is that consciousness seems to be all about forming a definite "sequentialized" thread of experience of the world, which is directly at odds with the idea of parallelism.

But so how can we proceed if we need to do distributed computing? Following what we believe about consciousness, I suspect a good approach will be to essentially mirror what we do in parsing the physical universe—and for example to pick reference frames in which to view and integrate the computation.

Distributed computing is difficult enough for us humans to "wrap our brains around". Multiway or nondeterministic computing tends to be even harder. And once again I suspect this is because of the "limitations imposed by consciousness". And that the way to handle it will be to use ideas that come from physics, and from the interaction of consciousness with quantum mechanics.

A few years ago at an AI ethics conference I raised the question of what would make us think AIs should have rights and responsibilities. "When they have consciousness!" said an

Contents

have consciousness. But the point is that attributing consciousness to something has potential consequences, say for ethics.

And it's interesting to see how the connection might work. Consider a system that's doing all sorts of sophisticated and irreducible computation. Already we might reasonably say that the system is showing a generalization of intelligence. But to achieve what we're viewing as consciousness the system also has to integrate this computation into some kind of single thread of experience.

And somehow it seems much more appropriate to attribute "responsibility" to that single thread that we can somehow "point to" than to a whole incoherent distributed computation. In addition, it seems much "more wrong" to imagine "killing" a single thread, probably because it feels much more unique and special. In a generic computational system there are many ways to "move forward". But if there's a single thread of experience it's more like there's only one.

And perhaps it's like the death of a human consciousness. Inevitably the history around that consciousness has affected all sorts of things in the physical universe that will survive its disappearance. But it's the thread of consciousness that ties it all together that seems significant to us, particularly as we try to make a "summary" of the universe to create our own coherent thread of experience.

And, by the way, when we talk about "explaining AI" what it tends to come down to is being able not just to say "that's the computation that ran", but being able to "tell a story" about what happened, which typically begins with making it "sequential enough" that we can relate to it like "another consciousness".

I've often noted that the Principle of Computational Equivalence has important implications for understanding our "place in the universe". We might have thought that with our life and intelligence there must be something fundamentally special about us. But what we've realized is that the essence of these is just computational sophistication—and the Principle of Computational Equivalence implies that that's actually quite ubiquitous and generic. So in a sense this promotes the importance of our human details—because that's ultimately all that's special about us.

Contents

can potentially "plug into" any pocket of reducibility of which there are inevitably infinitely many—even though we humans would not yet recognize most of them. But for our particular version of consciousness the idea of sequentialization seems to be central.

And, yes, we might have hoped that our consciousness would be something that even at an abstract level would put us "above" other parts of the physical universe. So the idea that this vaunted feature of ours is ultimately associated with what amounts to a restriction on computation might seem disappointing. But I view this as just part of the story that what's special about us are not big, abstract things, but specific things that reflect all that specific irreducible computation that has gone into creating our biology, our civilization and our lives.

In a sense the story of science is a story of struggle between computational irreducibility and computational reducibility. The richness of what we see is a reflection of computational irreducibility, but if we are to understand it we must find computational reducibility in it. And from what we have discussed here we now see how consciousness—which seems so core to our existence—might fundamentally relate to the computational reducibility we need for science, and might ultimately drive our actual scientific laws.

Notes

How does this all relate to what philosophers (and others) have said before? It will take significant work to figure that out, and I haven't done it. But it'll surely be valuable. Of course it'll be fun to know if Leibniz or Kant or Plato already figured out—or guessed—this or that, even centuries or millennia before we discovered some feature of computation or physics. But what's more important is that if there's overlap with some existing body of work then this provides the potential to make a connection with other aspects of that work, and to show, for example, how what I discuss might relate, say, to other areas of philosophy or other questions in philosophy.

My mother, Sybil Wolfram, was a longtime philosophy professor at Oxford University, and I was introduced to philosophical discourse at a very young age. I always said, though, that if there was one thing I'd never do when I was grown up, it's philosophy; it just seemed too crazy to still be arguing about the same issues after two thousand years. But after more than half a century of "detour" in science, here I am, arguably, doing philosophy after all….

Some of the early development of the ideas here were captured in the livestream: A Discussion about Physics Built by Alien Intelligences (June 25, 2020). Thanks particularly to Jeff Arle, Jonathan Gorard and Alexander Wolfram for discussions.

Contents

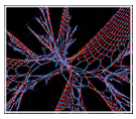
Posted in: Big Picture, New Kind of Science, Philosophy, Physics

Join the discussion

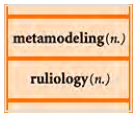
+ 42 comments

Related Writings

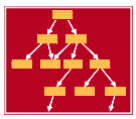
The Concept of the Ruliad
November 10, 2021



Multicomputation with Numbers: The Case of Simple Multiway Systems
October 7, 2021



Charting a Course for “Complexity”: Metamodeling, Ruliology and More
September 23, 2021



Even beyond Physics: Introducing Multicomputation as a Fourth General Paradigm for Theoretical Science
September 9, 2021

Popular Categories

Artificial Intelligence
Big Picture
Companies and Business
Computational Science
Computational Thinking
Data Science
Education
Mathematics
New Kind of Science
New Technology
Personal Analytics
Philosophy
Physics
Ruliology

Contents

Writings by Year

© Stephen Wolfram, LLC | Terms | RSS