

≡ STEPHEN WOLFRAM Writings

RECENT | CATEGORIES | 

Contents

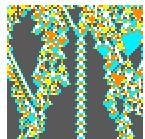
Observer Theory

PODCAST VIDEO

Observer Theory

December 11, 2023

The Concept of the Observer



We call it perception. We call it measurement. We call it analysis. But in the end it's about how we take the world as it is, and derive from it the impression of it that we have in our minds.

We might have thought that we could do science “purely objectively” without any reference to observers or their nature. But what we’ve [discovered particularly dramatically](#) in our [Physics Project](#) is that the nature of us as observers is critical even in determining the most fundamental laws we attribute to the universe.

But what ultimately does an observer—say like us—do? And how can we make a theoretical framework for it? Much as we have a [general model for the process of computation](#)—instantiated by something like a [Turing machine](#)—we’d like to have a general model for the process of observation: a general “observer theory”.

Central to what we think of as an observer is the notion that the observer will take the raw complexity of the world and extract from it some reduced representation suitable for a finite mind. There might be zillions of photons impinging on our eyes, but all we extract is the arrangement of objects in a visual scene. Or there might be zillions of gas molecules impinging on a piston, yet all we extract is the overall pressure of the gas.

In the end, we can think of it fundamentally as being about equivalencing. There are immense numbers of different individual configurations for the photons or the gas molecules—that are all treated as equivalent by an observer who's just picking out the particular features needed for some reduced representation.

There's in a sense a certain [duality between computation and observation](#). In computation one's generating new states of a system. In observation, one's equivalencing together different states.

That equivalencing must in the end be implemented “underneath” by computation. But in observer theory what we want to do is just characterize the equivalencing that's achieved. For us as observers it might in practice be all about how our senses work, what our biological or cultural nature is—or what technological devices or structures we've built. But what makes a coherent concept of observer theory possible is that there seem to be general, abstract characterizations that capture the essence of different kinds of observers.

It's not immediately obvious that anything suitable for a finite mind could ever be extracted from the complexity of the world. And indeed the [Principle of Computational Equivalence](#) implies that [computational irreducibility](#) (and its [multicomputational generalization](#)) will be ubiquitous. But within computational irreducibility there must always be slices of computational reducibility. And it's these slices of reducibility that an observer must try to pick out—and that ultimately make it possible for a finite mind to develop a “useful narrative” about what happens in the world, that allows it to make decisions, predictions, and so on.

How “special” is what an observer does? At its core it's just about taking a large set of possible inputs, and returning a much smaller set of possible outputs. And certainly that's a conceptual idea that's [appeared in many fields](#) under many different names: a contractive mapping, reduction to canonical form, a classifier, an acceptor, a forgetful functor, evolving to an attractor, extracting statistics, model fitting, lossy compression, projection, phase transitions, renormalization group transformations, coarse graining and so on. But here we want to think not about what's “mathematically describable”, but instead about what [in general is](#)

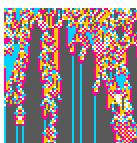
actually implemented—say by our senses, our measuring devices, or our ways of analyzing things.

At an ultimate level, everything that happens can be thought of as being captured by **the ruliad**—the unique object that emerges as the entangled limit of all possible computations. And in a vast generalization of ideas like that our brains—like any other material thing—are made of atoms, so too any observer must be embedded as some kind of structure within the ruliad. But a key concept of observer theory is that it's possible to make conclusions about an observer's impression of the world just by knowing about the capabilities—and assumptions—of the observer, without knowing in detail what the observer is “like inside”.

And so it is, for example, that in our Physics Project we seem to be able to derive—essentially from the structure of the ruliad—the core laws of twentieth-century physics (general relativity, quantum mechanics and the Second Law) just on the basis of two features of us as observers: that we're computationally bounded, and that we believe we're persistent in time (even though “underneath” we're made of different atoms of space at every successive moment). And we can expect that if we were to **include other features of us as observers** (for example, that we believe there are persistent objects in the world, or that we believe we have free will) then we'd be able to derive more aspects of the universe as we experience it—or of natural laws we attribute to it.

But the notion of observers—and observer theory—isn't **limited purely to “physical observers”**. It applies whenever we try to “get an impression” of something. And so, for example, we can also operate as “mathematical observers”, sampling the ruliad to build up conclusions about mathematical laws. Some features of us as physical observers—like the computational boundedness associated with the finiteness of our minds—inevitably carry over to us as mathematical observers. But other features do not. But the point of observer theory is to provide a general framework in which we can characterize observers—and then see the consequences of those characterizations for the impressions or conclusions observers will form.

The Operation of Observers



As humans we have senses like sight, hearing, touch, taste, smell and balance. And through our technology we also have access to a **few thousand other kinds of measurements**. So how basically do all these work?

The vast majority in effect aggregate a large number of small inputs to generate some kind of “average” output—which in the case of measurements is often specified as a (real) number. In a few cases, however, there’s instead a discrete choice between outputs that’s made on the basis of whether the total input exceeds a threshold (think: **distributed consensus** schemes, weighing balances, etc.)

But in all cases what’s fundamentally happening is that lots of different input configurations are all being equivalenced—or, more operationally, the dynamics of the system essentially make all equivalenced states evolve to the same “attractor state”.

As an example, let’s consider measuring the pressure of a gas. There are various ways to do this. But a very direct one is just to have a piston, and see how much force is exerted by the gas on this piston. So where does this force come from? At the lowest level it’s the result of lots of individual molecules bouncing off the surface of the piston, each transferring a tiny amount of momentum to it. If we looked at the piston at an atomic scale, we’d see it temporarily deform from each molecular impact. But the crucial point is that at a large scale the piston moves together, as a single rigid object—aggregating the effects of all those individual molecular impacts.

But why does it work this way? Essentially it’s because the intermolecular forces inside the piston are much stronger than the forces associated with molecules in the gas. Or, put more abstractly, there’s more coupling and coherence “inside the observer” than between the observer and what it’s observing.

We see the same basic pattern over and over again. There’s some form of transduction that couples the individual elements of what’s being observed to the observer. Then “within the observer” there’s something that in essence aggregates all these small effects. Sometimes that aggregation is “directly numerical”, as in the addition of lots of small momentum transfers. But

sometimes it's instead more explicitly like evolution to one attractor rather than another.

Consider, for example, the case of vision. An array of photons fall on the photoreceptor cells on our retinas, generating electrical signals transmitted through nerve fibers to our brains. Within the brain there's then effectively a **neural net that evolves to different attractors** depending on what one's looking at. Most of the time a small change in input image won't affect what attractor one evolves to. But—much like with a weighing balance—there's an “edge” at which even a small change can lead to a different output.

One can go through lots of different types of sensory systems and measuring devices. But the basic outline seems to always be the same. First, there's a coupling between what is being sensed or measured and the thing that's doing the sensing or measuring. Quite often that coupling involves transducing from one physical form to another—say from light to electricity, or from force to position. Sometimes then the crucial step of equivalencing different detailed inputs is achieved by simple “numerical aggregation”, most often by accumulation of objects (atoms, raindrops, etc.) or physical effects (forces, currents, etc.). But sometimes the equivalencing is instead achieved by a more obviously dynamical process.

It could amount to simple amplification, in which, say, the presence of a small element of input (say an individual particle) “tips over” some metastable system so that it goes into a certain final state. Or it could be more like a neural net where there's a more complicated translation defined by hard-to-describe borders between basins of attraction leading to different attractors.

But, OK, so what's the endpoint of a process of observation? Ultimately for us humans it's an impression created in our minds. Of course that gets into lots of slippery philosophical issues. Yes, each of us has an “inner experience” of what's going on in our mind. But anything else is ultimately an extrapolation. We make the assumption that other human minds also “see what we see”, but we can never “feel it from the inside”.

We can of course make increasingly detailed measurements—say of neural activity—to see how similar what's going on is between one brain and another.

But as soon as there's the slightest structural—or situational—difference between the brains, we really can't say exactly how their “impressions” will compare.

But for our purposes in constructing a general “observer theory” we're basically going to make the assumption (or, in effect, “philosophical approximation”) that whenever a system does enough equivalencing, that's tantamount to it “acting like an observer”, because it can then act as a “front end” that takes the “incoherent complexity of the world” and “collimates it” to the point where a mind will derive a definite impression from it.

Of course, there's still a lot of subtlety here. There has to be “just enough equivalencing” and not too much. For example, if all inputs were always equivalenced to the same output, there'd be nothing useful observed. And in the end there's somehow got to be some kind of match between the compression of input achieved by equivalencing, and the “capacity” of the mind that's ultimately deriving an impression from it.

A crucial feature of anything that can reasonably be called a mind is that “something's got to be going on in there”. It can't be, for example, that the internal state of the system is fixed. There has to be some internal dynamics—some computational process that we can identify as the ongoing operation of the mind.

At an informational level we might say that there has to be more information processing going on inside than there is flow of information from the outside. Or, in other words, if we're going to be meaningful “observers like us” we can't just be bombarded by input we don't process; we have to have some capability to “think about what we're seeing”.

All of this comes back to the idea that a crucial feature of us as observers is that we are computationally bounded. We do computation; that's why we can have an “inner sense of things going on”. But the amount of computation we do is tiny compared to the computation going on in the world around us. Our experience represents a heavily filtered version of “what's happening outside”. And the essence of “being an observer like us” is that we're effectively doing lots of equivalencing to get to that filtered version.

But can we imagine a future in which we “expand our minds”? Or perhaps encounter some alien intelligence with a fundamentally “less constrained mind”? Well, at some point there’s an issue with this. Because in a sense the idea that we have a coherent existence relies on us having “limited minds”. For without such constraints there wouldn’t be a coherent “self” that we could identify—with coherent inner experience.

Let’s say we’re shown some system—say in nature—“from the outside”. Can we tell if “there’s an observer in there”? Ultimately not, because in a sense we’d have to be “inside that observer” and be able to experience the impression of the world that it’s getting. But in much the same way as we extrapolate to believing that, say, other human minds are experiencing things like we’re experiencing, so also we can potentially extrapolate to say what we might think of as an observer.

And the core idea seems to be that an “observer” should be a subsystem whose “internal states” are affected by the rest of the system, but where many “external states” lead to the same internal state—and where there is rich dynamics “within the observer” that in effect operates only on its internal states. Ultimately—following the [Principle of Computational Equivalence](#)—both the outside and the inside of the “observer subsystem” can be expected to be equivalent in the computations they’re performing. But the point is that the coupling from outside the subsystem to inside effectively “coarse grains” what’s outside, so that the “inner computation” is operating on a much-reduced set of elements.

Why should any such “observer subsystems” exist? Presumably at some level it’s inevitable from the presence of pockets of computational reducibility within arbitrary computationally irreducible systems. But more important for us is that our very existence—and the possibility of our coherent inner experience—depends on us “operating as observers”. And—almost as a “self-fulfilling prophecy”—our behavior tends to perpetuate our ability to successfully do this. For example, we can think of us as choosing to put ourselves in situations and environments where we can “predict what’s going to happen” well enough to “survive as observers”. (At a mundane practical level we might do this by not living in places subject to unpredictable natural forces—or by doing things like building ourselves structures that shelter us from those forces.)

We've talked about observers operating by [compressing the complexities](#) of the world to "inner impressions" suitable for finite minds. And in typical situations that we describe as perception and measurement, the main way this happens is by fairly direct equivalencing of different states. But in a sense there's a higher-level story that relies on formalization—and in essence computation—and that's what we usually call "analysis".

Let's say we have some intricate structure—perhaps some nested, fractal pattern. A direct rendering of all the pixels in this pattern ultimately won't be something well suited for a "finite mind". But if we gave rules—or a program—for generating the pattern we'd have a much more succinct representation of it.

But now there's a problem with computational irreducibility. Yes, the rules determine the pattern. But to get from these rules to the actual pattern can require an irreducible amount of computation. And to "reverse engineer the pattern" to find the rules can require even more computation.

Yes, there are particular cases—like repetitive and simple nested patterns—where there's enough immediate computational reducibility that a computationally bounded system (or observer) can fairly easily "do the analysis" and "get the compression". But in general it's hard. And indeed in a sense it's the whole mission of science to pick away at the problem, and try to find more ways to "reduce the complexities of the world" to "human-level narratives".

Computational irreducibility limits the extent to which this can be successful. But the inevitable existence of pockets of reducibility even within computational irreducibility guarantees that progress can always in principle be made. As we invent more kinds of measuring devices we can extend our domain as observers. And the same is true when we invent more methods of analysis, or identify more principles in science.

But the overall picture remains the same: what's crucial to "being an observer" is equivalencing many "states of the world", either through perceiving or measuring only specific aspects of them, or through identifying "simplified narratives" that capture them. (In effect, perception and measurement tend to do "lossy compression"; analysis is more about "lossless compression" where the

equivalencing is effectively not between possible inputs but between possible generative rules.)

How Observers Construct Their Perceived Reality



Our view of the world is ultimately determined by what we observe of it. We take what's "out there in the world" and in effect "construct our perceived reality" by our operation as observers. Or, in other words, insofar as we have a narrative about "what's going on in the world", that's something that comes from our operation as observers.

And in fact from our Physics Project we're led to an extreme version of this—in which what's "out there in the world" is just the whole ruliad, and in effect everything specific about our perceived reality must come from how we operate as observers and thus [how we sample the ruliad](#).

But long before we get to this ultimate level of abstraction, there are lots of ways in which our nature as observers "builds" our perceived reality. Think about any material substance—like a fluid. Ultimately it's made up of lots of individual molecules "doing their thing". But observers like us aren't seeing those molecules. Instead, we're aggregating things to the point where we can just describe the system as a fluid, that operates according to the "narrative" defined by the laws of fluid mechanics.

But [why do things work this way](#)? Ultimately it's the result of the repeated story of the interplay between underlying computational irreducibility, and the computational boundedness of us as observers. At the lowest level the motion of the molecules is governed by simple rules of mechanics. But the phenomenon of computational irreducibility implies that to work out the detailed consequences of "running these rules" involves an irreducible amount of computational work—which is something that we as computationally bounded observers can't do. And the result of this is that we'll end up describing the detailed behavior of the molecules as just "random". As I've [discussed at length elsewhere](#), this is the fundamental origin of the Second Law of thermodynamics. But for our purposes here the important point is that it's what makes observers like us "construct the reality" of things like fluids. Our computational boundedness as observers makes

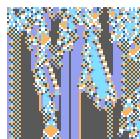
us unable to trace all the detailed behavior of molecules, and leaves us “content” to describe fluids in terms of the “narrative” defined by the laws of fluid mechanics.

Our Physics Project implies that it’s the **same kind of story with physical space**. For in our Physics Project, space is ultimately “made” of a network of relations (or connections) between discrete “atoms of space”—that’s progressively being updated in what ends up being a computationally irreducible way. But we as computationally bounded observers can’t “decode” all the details of what’s happening, and instead we end up with a simple “aggregate” narrative, that turns out to correspond to continuum space operating according to the laws of general relativity.

The way both coherent notions of “matter” (or fluids) and spacetime emerge for us as observers can be thought of as a consequence of the equivalencing we do as observers. In both cases, there’s immense and computationally irreducible complexity “underneath”. But we’re ignoring most of that—by effectively treating different detailed behaviors as equivalent—so that in the end we get to a (comparatively) “simple narrative” more suitable for our finite minds. But we should emphasize that what’s “really going on in the system” is something much more complicated; it’s just that we as observers aren’t paying attention to that, so our perceived reality is much simpler.

OK, but what about quantum mechanics? In a sense that’s an extreme test of our description of how observers work, and the extent to which the operation of observers “constructs their perceived reality”.

The Case of Quantum Mechanics



In our Physics Project the underlying structure (hypergraph) that represents space and everything in it is progressively being rewritten according to definite rules. But the crucial point is that at any given stage there can be lots of ways this rewriting can happen. And the result is that there’s a **whole tree of possible “states of the universe”** that can be generated. So given this, why do we ever think that definite things happen in the universe? Why

don't we just think that there's an infinite tree of branching histories for the universe?

Well, it all has to do with our nature as observers, and the equivalencing we do. At an immediate level, we can imagine looking at all those different possible branching paths for the evolution of the universe. And the key point is that even though they come from different paths of history, two states can just be the same. Sometimes it'll be obvious that they're same; sometimes one might have to determine, say, whether two hypergraphs are isomorphic. But the point is that to any observer (at least one that isn't managing to look at arbitrary "implementation details"), the states will inevitably be considered equivalent.

But now there's a bigger point. Even though "from the outside" there might be a whole branching and merging **multiway graph of histories** for the universe, observers like us can't trace that. And in fact all we perceive is a single thread of history. Or, said another way, we believe that we have a **single thread of experience**—something closely related to our belief that (despite the changing "underlying elements" from which we are made) we are somehow persistent in time (at least during the span of our existence).

But operationally, how do we go from all those underlying branches of history to our perceived single thread of history? We can think of the states on different threads of history as being related by what we call a **branchial graph**, that joins states that have immediate common ancestors. And in the limit of many threads, we can think of these different states as being laid out "branchial space". (In traditional quantum mechanics terms, this layout defines a "map of quantum entanglements"—with each piece of common ancestry representing an entanglement between states.)

In physical space—whether we're looking at molecules in a fluid or atoms of space—we can think of us operating as observers who are physically large enough to span many underlying discrete elements, so that what we end up observing is just some kind of aggregate, averaged result. And it's very much the same kind of thing in branchial space: we as observers tend to be large enough in branchial space to be spread across an immense number of branches of history, so that what we observe is just aggregate, averaged results across all those branches.

There's lots of detailed complexity in what happens on different branches, just like there is in what happens to different molecules, or different atoms of space. And the reason is that there's inevitably computational irreducibility, or, in this case, more accurately, multicomputational irreducibility. But as computationally bounded observers we just perceive aggregate results that "average out" the "underlying apparent randomness" to give a consistent single thread of experience.

And effectively this is what happens in the transition from quantum to classical behavior. Even though there are many possible detailed ("quantum") threads of history that an object can follow, what we perceive corresponds to a single consistent "aggregate" ("classical") sequence of behavior.

And this is typically true even at the level of our typical observation of molecules and chemical processes. Yes, there are many possible threads of history for, say, a water molecule. But most of our observations aggregate things to the point where we can talk about a definite shape for the molecule, with definite "chemical bonds", etc.

But there is a special situation that actually looms large in typical discussions of quantum mechanics. We can think of it as the result of doing measurements that aren't "aggregating threads of history to get an average", but are instead doing something more like a weighing balance, always "tipping" one way or the other. In the language of [quantum computing](#), we might say that we're arranging things to be able to "measure a single qubit". In terms of the equivalencing of states, we might say that we're equivalencing lots of underlying states to specific canonical states (like "spin up" and "spin down").

Why do we get one outcome rather than another? Ultimately we can think of it as all depending on the details of us as observers. To see this, let's start from the corresponding question in physical space. We might ask why we observe some particular thing happening. Well, in our Physics Project everything about "what happens" is deterministic. But there's still the "arbitrariness" of where we are in physical space. We'll always basically see the same laws of physics, but the particulars of what we'll observe depend on where we are, say on the surface of the Earth versus in interstellar space, etc.

Is there a “theory” for “where we are”? In some sense, yes, because we can go back and see why the molecules that make us up landed up in the particular place where they did. But what we can’t have an “external theory” for is just which molecules end up making up “us”, as we experience ourselves “from inside”. In our view of physics and the universe, it’s in some sense the only “ultimately subjective” thing: where our internal experience is “situated”.

And the point is that basically—even though it’s much less familiar—the same thing is going on at the level of quantum mechanics. Just as we “happen” to be at a certain place in physical space, so we’re at a certain place in branchial space. Looking back we can trace how we got here. But there’s no *a priori* way to determine “where our particular experience will be situated”. And that means we can’t know what the “local branchial environment” will be—and so, for example, what the outcome of “balance-like” measurements will be.

Just as in traditional discussions of quantum mechanics, the mechanics of doing the measurement—which we can think of as effectively equivalencing many underlying branches of history—will have an effect on subsequent behavior, and subsequent measurements.

But let’s say we look just at the level of the underlying multiway graph—or, more specifically, the multiway causal graph that records causal connections between different updating events. Then we can identify a complicated web of interdependence between events that are timelike, spacelike and branchlike separated. And this interdependence seems to correspond precisely to what’s expected from quantum mechanics.

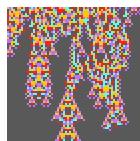
In other words, even though the multiway graph is completely determined, the arbitrariness of “where the observer is” (particularly in branchial space), combined with the inevitable interdependence of different aspects of the multiway (causal) graph, seems sufficient to reproduce the not-quite-purely-probabilistic features of quantum mechanics.

In making observations in physical space, it’s common to make a measurement at one place or time, then make another measurement at another place or time, and, for example, see how they’re related. But in actually doing this, the observer will have to move from one place to the other, and persist from one time to another.

And in the abstract it's not obvious that that's possible. For example, it could be that an observer won't be able to move without changing—or, in other words, that "pure motion" won't be possible for an observer. But in effect this is something we as observers assume about ourselves. And indeed, as I've [discussed elsewhere](#), this is a crucial part of why we perceive spacetime to operate according to the laws of physics we know.

But what about in branchial space? We have much less intuition for this than for physical space. But we still effectively believe that **pure motion is possible for us as observers in branchial space**. It could be—like an observer in physical space, say, near a spacetime singularity—that an observer would get "shredded" when trying to "move" in branchial space. But our belief is that typically nothing like that happens. At some level being at different locations in branchial space presumably corresponds to picking different bases for our quantum states, or effectively to defining our experiments differently. And somehow our belief in the possibility of pure motion in branchial space seems related to our belief in the possibility of making arbitrary sequences choices in sets of experiments we do.

Observers of Abstract Worlds



We might have thought that the only thing ultimately "out there" for us to observe would be our physical universe. But actually there are important situations where we're essentially operating not as observers of our familiar physical universe, but instead of what amount to abstract universes. And what we'll see is that the ideas of observer theory seem to apply there too—except that now what we're picking out and reducing to "internal impressions" are features not of the physical world but of abstract worlds.

Our [Physics Project](#) in a sense brings ideas about the physical and abstract worlds closer—and the [concept of the ruliad](#) ultimately leads to a deep unification between them. For what we now imagine is that the physical universe as we perceive it is just the result of the particular kind of sampling of the ruliad made by us as certain kinds of observers. And the point is that we as observers can make other kinds of samplings, leading to what we can describe as abstract

universes. And one particularly prominent example of this is [mathematics](#), or rather, [metamathematics](#).

Imagine starting from all possible axioms for mathematics, then constructing the network of all possible theorems that can be derived from them. We can consider this as forming a kind of “[metamathematical universe](#)”. And the particular mathematics that some mathematician might study we can then think of as the result of a “mathematical observer” observing that metamathematical universe.

There are both close analogies and differences between this and the experience of a physical observer in the physical universe. Both ultimately correspond to samplings of the ruliad, but somewhat different ones.

In our Physics Project we imagine that physical space and everything in it is ultimately made up of discrete elements that we identify as “atoms of space”. But in the ruliad in general we can think of everything being made up of “pure atoms of existence” that we call emes. In the particular case of physics we interpret these emes as atoms of space. But in metamathematics we can think of emes as corresponding to (“subaxiomatic”) elements of symbolic structures—from which things like axioms or theorems can be constructed.

A central feature of our interaction with the ruliad for physics is that observers like us don’t track the detailed behavior of all the various atoms of space. Instead, we equivalence things to the point where we get descriptions that are reduced enough to “fit in our minds”. And something similar is going on in mathematics.

We don’t track all the individual [subaxiomatic emes](#)—or usually in practice even the details of fully formalized axioms and theorems. Instead, mathematics typically operates at a much higher and “more human” level, dealing not with questions like how real numbers can be built from emes—or even axioms—but rather with what can be deduced about the properties of mathematical objects like real numbers. In a physics analogy to the behavior of a gas, typical human mathematics operates not at the “molecular” level of individual emes (or even axioms) but rather at the “fluid dynamics” level of “human-accessible” mathematical concepts.

In effect, therefore, a mathematician is operating as an observer who equivalences many detailed configurations—ultimately of emes—in order to form higher-level mathematical constructs suitable for our computationally bounded minds. And while at the outset one might have imagined that anything in the ruliad could serve as a “possible mathematics”, the point is that observers like us can only sample the ruliad in particular ways—leading to only particular possible forms for “human-accessible” mathematics.

It’s a very similar story to the one we’ve encountered many times in thinking about physics. In studying gases, for example, we could imagine all sorts of theories based on tracking detailed molecular motions. But for observers like us—with our computational boundedness—we inevitably end up with things like the **Second Law of thermodynamics**, and the laws of fluid mechanics. And in mathematics the main thing we end up with is “higher-level mathematics”—mathematics that we can do directly in terms of typical textbook concepts, rather than constantly having to “drill down” to the level of axioms, or emes.

In physics we’re usually particularly concerned with issues like predicting how things will evolve through time. In mathematics it’s more about **accumulating what can be considered true**. And indeed we can think of an idealized mathematician as going through the ruliad and collecting in their minds a “bag” of theorems (or axioms) that they “consider to be true”. And given such a collection, they can essentially follow the “entailment paths” defined by computations in the ruliad to find more theorems to “add to their bag”. (And, yes, if they put in a false theorem then—because a false premise in the standard setup of logic implies everything—they’ll end up with an “**infinite explosion of theorems**”, that won’t fit in a finite mind.)

In observing the physical universe, we talk about our different possible senses (like vision, hearing, etc.) or different kinds of measuring devices. In observing the metamathematical universe the analogy is basically different possible kinds of theories or abstractions—say, algebraic vs. geometrical vs. topological vs. categorical, etc. (with new approaches being like new kinds of measuring devices).

Particularly when we think in terms of the ruliad we can expect a certain kind of **ultimate unity in the metamathematical universe**—but different theories and different abstractions will pick up different aspects of it, just as vision and hearing pick up different aspects of the physical universe. But in a sense observer theory gives us a global way to talk about this, and to characterize what kinds of observations observers like us can make—whether of the physical universe or the metamathematical one.

In physics we've then seen in our Physics Project how this allows us to find general laws that describe our perception of the physical world—and that turn out to reproduce the core known laws of physics. In mathematics we're not as familiar with the concept of general laws, though the very fact that higher-level mathematics is possible is presumably in essence such a law, and perhaps the kinds of regularities seen in areas like category theory are others—as are the inevitable dualities we expect to be able to identify between different fields of mathematics. All these laws ultimately rely on the structure of the ruliad. But the crucial point is that they're not talking about the “raw ruliad”; instead they're talking about just certain samplings of the ruliad that can be done by observers like us, and that lead to certain kinds of “internal impressions” in terms of which these laws can be stated.

Mathematics represents a certain kind of abstract setup that's been studied in a particularly detailed way over the centuries. But it's not the only kind of “abstract setup” we can imagine. And indeed there's even a much more familiar one: the use of concepts—and words—in human thinking and language.

We might imagine that at some time in the distant past our forebears could signify, say, rocks only by pointing at individual ones. But then there emerged the general notion of “rock”, captured by a word for “rock”. And once again this is a story of observers and equivalences. When we look at a rock, it presumably produces all sorts of detailed patterns of neuron firings in our brains, different for each particular rock. But somehow—presumably essentially through evolution to an attractor in the neural net in our brains—we equivalence all these patterns to extract our “inner impression” of the “concept of a rock”.

In the typical tradition of quantitative science we tend to be interested in doing measurements that lead to things like numerical results. But in representing the world using language we tend to be interested instead in creating symbolic structures that involve collections of discrete words embedded in a grammatical framework. Such linguistic descriptions don't capture every detail; in a typical observer kind of way they broadly equivalence many things—and in a sense reduce the complexity of the world to a description in terms of a limited number of discrete words and linguistic forms.

Within any given person's brain there'll be "thoughts" defined by patterns of neuron firings. And the crucial role of language is to provide a way to robustly "package up" those thoughts, and for example represent them with discrete words, so they can be communicated to another person—and unpacked in that person's brain to produce neuron firings that reproduce what amount to those same thoughts.

When we're dealing with something like a numerical measurement we might imagine that it could have some kind of absolute interpretation. But words are much more obviously an "arbitrary basis" for communication. We could pick a different specific word (say from a different human language) but still "communicate the same thing". All that's required is that everyone who's using the word agrees on its meaning. And presumably that normally happens because of shared "social" history between people who use a given word.

It's worth pointing out that for this to work there has to be a certain separation of scales. The collective impression of the meaning of a word may change over time, but that change has to be slow compared to the rate at which the word is used in actual communication. In effect, the meaning of a word—as we humans might understand it—emerges from the aggregation of many individual uses.

In the abstract, there might not be any reason to think that there'd be a way to "understand words consistently". But it's a story very much like what we've encountered in both physics and mathematics. Even though there are lots of complicated individual details "underneath", we as observers manage to pick out features that are "simple enough for us to understand". In the case of molecules

in a gas that might be the overall pressure of the gas. And in the case of words it's a **stable notion of "meaning"**.

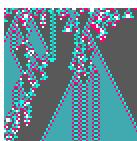
Put another way, the possibility of language is another example of observer theory at work. Inside our brains there are all sorts of complicated neuron firings. But somehow these can be "packaged up" into things like words that form "human-level narratives".

There's a certain complicated feedback loop between the world as we experience it and the words we use to describe it. We invent words for things that we commonly encounter ("chair", "table", ...). Yet once we have a word for something we're more able to form thoughts about it, or communicate about it. And that in turn makes us more likely to put instances of it in our environment. In other words, we tend to build our environment so that the way we have of making narratives about it works well—or, in effect, so our inner description of it can be as simple as possible, and it can be as predictable to us as possible.

We can view our experience of physics and of mathematics as being the result of us acting as physical observers and mathematical observers. Now we're viewing our experience of the "conceptual universe" as being the result of us acting as "conceptual observers". But what's crucial is that in all these cases, we have the same intrinsic features as observers: computational boundedness and a belief in persistence. The computational boundedness is what makes us equivalence things to the point where we can have symbolic descriptions of the world, for example in terms of words. And the belief in persistence is what lets those words have persistent meanings.

And actually these ideas extend beyond just language—to paradigms, and general ways of thinking about things. When we define a word we're in effect defining an abstraction for a class of things. And paradigms are somehow a generalization of this: ways of taking lots of specifics and coming up with a uniform framework for them. And when we do this, we're in effect making a classic observer theory move—and equivalencing lots of different things to produce an "internal impression" that's "simple enough" to fit in our finite minds.

In the End It's All Just the Ruliad



Our tendency as observers is always to believe that we can separate our “inner experience” from what’s going on in the “outside world”. But in the end everything is just part of the ruliad. And at the level of the ruliad we as observers are ultimately “made of the same stuff” as everything else.

But can we imagine that we can point at one part of the ruliad and say “that’s an observer”, and at another part and say “that’s not”? At least to some extent the answer is presumably yes—at least if we restrict ourselves to “observers like us”. But it’s a somewhat subtle—and seemingly circular—story.

For example, one core feature of observers like us is that we have a certain persistence, or at least we believe we have a certain persistence. But, inevitably, at the level of the “raw ruliad”, we’re continually being made from different atoms of existence, i.e. different emes. So in what sense are we persistent? Well, the point is that an observer can equivalence those successive patterns of emes, so that what they observe is persistent. And, yes, this is at least on the face of it circular. And ultimately to identify what parts of the ruliad might be “persistent enough to be observers”, we’ll have to ground this circularity in some kind of further assumption.

What about the computational boundedness of observers like us, which forces us to do lots of equivalencing? At some level that equivalencing must be implemented by lots of different states evolving to the same states. But once again there’s circularity, because even to define what we mean by “the same states” (“Are isomorphic graphs the same?”, etc.) we have to be imagining certain equivalencing.

So how do we break out of the circularity? The key is presumably the presence of additional features that define “observers like us”. And one important class of such features has to do with scale.

We’re neither tiny nor huge. We involve enough emes that consistent averages can emerge. Yet we don’t involve so many emes that we span anything but an absolutely tiny part of the whole ruliad.

And actually a lot of our experience is determined by “our size as observers”. We’re large enough that certain equivalencing is inevitable. Yet we’re small

enough that we can reasonably think of there being many choices for “where we are”.

The overall structure of the ruliad is a [matter of formal necessity](#); there’s only one possible way for it to be. But there’s contingency in our character as observers.

And for example in a sense there’s a fundamental constant of nature as we perceive it, which is our extent in the ruliad, say measured in emes (and appropriately projected into physical space, branchial space, etc.).

And the fact that this extent is small compared to the whole ruliad means that there are “many possible observers”—who we can think of as existing at different positions in the ruliad. And those different observers will look at the ruliad from different “points of view”, and thus develop different “internal impressions” of “perceived reality”.

But a crucial fact central to our Physics Project is that there are certain aspects of that perceived reality that are inevitable for observers like us—and that correspond to core laws of physics. But when it gets to more specific questions (“What does the night sky look like from where you are?”, etc.) different observers will inevitably have different versions of perceived reality.

So is there a way to translate from one observer to another? Essentially that’s a [story of motion](#). What happens when an observer at one place in the ruliad “moves” to another place? Inevitably, the observer will be “made of different emes” if it’s at a different place. But will it somehow still “be the same”? Well, that’s a subtle question, that depends both on the background structure of the ruliad, and the nature of the observer.

If the ruliad is “too wild” (think: spacetime near a singularity) then the observer will inevitably be “shredded” as it “moves”. But computational irreducibility implies a certain overall regularity to most of the ruliad, making “pure motion” at least conceivable. But to achieve “pure motion” the observer still has to be “made of” something that is somehow robust—essentially some “lump of computational reducibility” that can “predictably survive” the underlying background of computational irreducibility.

In spacetime we can identify such “lumps” with things like black holes, and particles like electrons, photons, etc. (and, yes, in our models there’s probably considerable commonality between black holes and particles). It’s not yet clear quite what the analog is in branchial space, though a very simple example might involve persistence of qubits. And in rulial space, one kind of analog is the very notion of concepts. For in effect concepts (as represented for example by words) are the analog of particles in rulial space: they are the robust structures that can move across rulial space and “maintain their identity”, carrying “the same thoughts” to different minds.

So what does all this mean for what can constitute an observer in the ruliad? Observers in effect leverage computational reducibility to extract simplified features that can “fit in finite minds”. But observers themselves must also embody computational reducibility in order to maintain their own persistence and the persistence of the features they extract. Or in other words, observers must in a sense always correspond to “patches of regularity” in the ruliad.

But can any patch of regularity in the ruliad be thought of as an observer? Probably not usefully so. Because another feature of observers like us is that we are connected in some kind of collective “social” framework. Not only do we individually form internal impressions in our minds, but we also communicate these impressions. And indeed without such communication we wouldn’t, for example, be able to set up things like coherent languages with which to describe things.

What We Assume about Ourselves



A key implication of our Physics Project and the concept of the ruliad is that we perceive the universe to be the way we do because we are the way we are as observers. And the most fundamental aspect of observers like us is that we’re doing lots of equivalencing to reduce the “complexity of the world” to “internal impressions” that “fit into our minds”. But just what kinds of equivalencing are we actually doing? At some level a lot of that is defined by the things we believe—or assume—about ourselves and the way we interact with the world.

A very central assumption we make is that we're somehow "stable observers" of a changing "outside world". Of course, at some level we're actually not "stable" at all: we're built up from emes whose configuration is changing all the time. But our belief in our own stability—and, in effect, our belief in our "persistence in time"—makes us equivalence those configurations. And having done that equivalencing we perceive the universe to operate in a certain way, that turns out to align with the laws of physics we know.

But actually there's more than just our assumption of persistence in time. For example, we also have an assumption of persistence in space: we assume that—at least on reasonably short timescales—we're consistently "observing the universe from the same place", and not, say, "continually darting around". The network that represents space is continually changing "around us". But we equivalence things so that we can assume that—in a first approximation—we are "staying in the same place".

Of course, we don't believe that we have to stay in exactly the same place all the time; we believe we're able to move. And here we make what amounts to another "assumption of stability": we assume that **pure motion is possible** for us as observers. In other words, we assume that we can "go to different places" and still be "the same us", with the same properties as observers.

At the level of the "raw ruliad" it's not at all obvious that such assumptions can be consistently made. But as we discussed above, the fact that for observers like us they can (at least to a good approximation) is a reflection of certain properties of us as observers—in particular of our physical scale, being large in terms of atoms of space but small in terms of the whole universe.

Related to our assumption about motion is our assumption that "space exists"—or that we can treat space as something coherent. Underneath, there's all sorts of complicated dynamics of changing patterns of emes. But on the timescales at which we experience things we can equivalence these patterns to allow us to think of space as having a "coherent structure". And, once again, the fact that we can do this is a consequence of physical scales associated with us as observers. In particular, the speed of light is "fast enough" that it brings information to us from the local region around us in much less time than it takes our brain to process it.

And this means that we can equivalence all the different ways in which different pieces of information reach us, and we can consistently just talk about the state of a region of space at a given time.

Part of our assumption that we're "persistent in time" is that our thread of experience is—at least locally—continuous, with no breaks. Yes, we're born and we die—and we also sleep. But we assume that at least on scales relevant for our ongoing perception of the world, we experience time as something continuous.

More than that, we assume that we have just a single thread of experience. Or, in other words, that there's always just "one us" going through time. Of course, even at the level of neurons in our brains all sorts of activity goes on in parallel. But somehow in our normal psychological state we seem to concentrate everything so that our "inner experience" follows just one "thread of history", on which we can operate in a computationally bounded way, and form definite memories and have definite sequences of thoughts.

We're not as familiar with branchial space as with physical space. But presumably our "fundamental assumption of stability" extends there as well. And when combined with our basic computational boundedness it then becomes inevitable that (as we discussed above) we'll conflate different "quantum paths of history" to give us as observers a definite "classical thread of inner experience".

Beyond "stability", another very important assumption we implicitly make about ourselves is what amounts to an assumption of "independence". We imagine that we can somehow separate ourselves off from "everything else". And one aspect of this is that we assume we're localized—and that most of the ruliad "doesn't matter to us", so that we can equivalence all the different states of the "rest of the ruliad".

But there's also another aspect of "independence": that in effect we can choose to do "whatever we want" independent of the rest of the universe. And this means that we assume we can, for example, essentially "do any possible experiment", make any possible measurement—or "go anywhere we want" in physical or branchial space, or indeed rulial space. We assume that we effectively have "free

“will” about these things—determined only by our “inner choices”, and independent of the state of the rest of the universe.

Ultimately, of course, we’re just part of the ruliad, and everything we do is determined by the structure of the ruliad and our history within it. But we can view our “belief of freedom” as a reflection of the fact that we don’t know *a priori* where we’ll be located in the ruliad—and even if we did, computational irreducibility would [prevent us from making predictions about what we will do](#).

Beyond our assumptions about our own “independence from the rest of the universe”, there’s also the question of independence between different parts of what we observe. And quite central to our way of “parsing the world” is our typical assumption that we can “think about different things separately”. In other words, we assume it’s possible to “factor” what we see happening in the universe into independent parts.

In science, this manifests itself in the idea that we can do “controlled experiments” in which we study how something behaves in isolation from everything else. It’s not self-evident that this will be possible (and indeed in areas like ethics it might fundamentally not be), but we as observers tend to implicitly assume it.

And actually, we normally go much further. Because we typically assume that we can describe—and think about—the world “symbolically”. In other words, we assume that we can take all the complexity of the world and represent at least the parts of it that we care about in terms of discrete symbolic concepts, of the kind that appear in human [\(or computational\) language](#). There’s lots of detail in the world that our limited collection of symbolic concepts doesn’t capture, and effectively “equivalences out”. But the point is that it’s this symbolic description that normally seems to form the backbone of the “inner narrative” we have about the world.

There’s another implicit assumption that’s being made here, however. And that’s that there’s some kind of stability in the symbolic concepts we’re using. Yes, any particular mind might parse the world using a particular set of symbolic concepts. But we make the implicit assumption that there are other minds out there that work like ours. And this makes us imagine that there can be some form

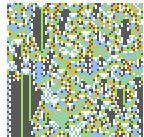
of “objective reality” that’s just “always out there”, to be sampled by whatever mind might happen to come along.

Not only, therefore, do we assume our own stability as observers; we also assume a certain stability to what we perceive of “everything that’s out there”.

Underneath, there’s all the wildness and complexity of the ruliad. But we assume that we can successfully equivalence things to the point where all we perceive is something quite stable—and something that we can describe as ultimately governed by consistent laws.

It could be that every part of the universe just “does its own thing”, with no overall laws tying everything together. But we make the implicit assumption that, no, the universe—at least as far as we perceive it—is a more organized and consistent place. And indeed it’s that assumption that makes it feasible for us to operate as observers like us at all, and to even imagine that we can usefully reduce the complexity of the world to something that “fits in our finite minds”.

The Cost of Observation



What resources does it take for an observer to make an observation? In most of traditional science, observation is at best added as an afterthought, and no account is taken of the process by which it occurs. And indeed, for example, in the traditional formalism of quantum mechanics, while “measurement” can have an effect on a system, it’s still assumed to be an “indivisible act” without any “internal process”.

But in observer theory, we’re centrally talking about the process of observation. And so it makes sense to try asking questions about the resources involved in this process.

We might start with our own everyday experience. Something happens out in the world. What resources—and, for example, how much time—does it take us to “form an impression of it”? Let’s say that out in the world a cat either comes into view or it doesn’t. There are signals that come to our brain from our eyes, effectively carrying data on each pixel in our visual field. Then, inside our brain,

these signals are processed by a succession of layers of neurons, with us in the end concluding either “there’s a cat there”, or “there’s not”.

And from artificial neural nets we can get a pretty good idea of how this likely works. And the key to it—as we discussed above—is that there’s an attractor. Lots of different detailed configurations of pixels all evolve either to the “cat” or “no cat” final state. The different configurations have been equivalenced, so that only a “final conclusion” survives.

The story is a bit trickier though. Because “cat” or “no cat” really isn’t the final state of our brain; hopefully it’s not the “last thought we have”. Instead, our brain will continue to “think more thoughts”. So “cat”/“no cat” is at best some kind of intermediate waypoint in our process of thinking; an instantaneous conclusion that we’ll continue to “build on”.

And indeed when we consider measuring devices (like a piston measuring the pressure of a gas) we similarly usually imagine that they will “come to an instantaneous conclusion”, but “continue operating” and “producing more data”. But how long should we wait for each intermediate conclusion? How long, for example, will it take for the stresses generated by a particular pattern of molecules hitting a piston to “dissipate out”, and for the piston to be “ready to produce more data”?

There are lots of specific questions of physics here. But if our purpose is to build a formal observer theory, how should we think about such things? There is something of an analogy in the formal theory of computation. An actual computational system—say in the physical world—will just “keep computing”. But in formal computation theory it’s useful to talk about computations that halt, and about functions that can be “evaluated” and give a “definite answer”. So what’s the analog of this in observer theory?

Instead of general computations, we’re interested in computations that effectively “implement equivalences”. Or, put another way, we want computations that “destroy information”—and that have many incoming states but few outgoing ones. As a practical matter, we can either have the outgoing states explicitly represent whole equivalence classes, or they can just be “canonical

representatives”—like in a network where at each step each element takes on whatever the “**majority**” or “**consensus**” value of its neighbors was.

But however it works, we can still ask questions about what computational resources were involved. How many steps did it take? How many elements were involved?

And with the idea that observers like us are “computationally bounded”, we expect limitations on these resources. But with this formal setup we can start asking just how far an observer like us can get, say in “coming to a conclusion” about the results of some computationally irreducible process.

An interesting case arises in **putative quantum computers**. In the model implied by our Physics Project, such a “quantum computer” effectively “performs many computations in parallel” on the separate branches of a multiway system representing the various threads of history of the universe. But if the observer tries to “come to a conclusion” about what actually happened, they have to “knit together” all those threads of history, in effect by implementing equivalences between them.

One could in principle imagine an observer who’d just follow all the quantum branches. But it wouldn’t be an observer like us. Because what seems to be a core feature of observers like us is that we believe we have just a single thread of experience. And to maintain that belief, our “process of observation” must equivalence all the different quantum branches.

How much “effort” will that be? Well, inevitably if a thread of history branched, our equivalencing has to “undo that branching”. And that suggests that the number of “elementary equivalencings” will have to be at least comparable to the number of “elementary branchings”—making it seem that the “effort of observation” will tend to be at least comparable to reduction of effort associated with parallelism in the “underlying quantum process”.

In general it’s interesting to compare the “effort of observation” with the “effort of computation”. With our concept of “elementary equivalencings” we have a way to measure both in terms of computational operations. And, yes, both could in principle be implemented by something like a Turing machine, though in practice

the equivalencings might be most conveniently modeled by something like string rewriting.

And indeed one can often go much further, talking not directly in terms of equivalencings, but rather about processes that show attractors. There are different kinds of attractors. Sometimes—as in [class 1 cellular automata](#)—there are just a limited number of static, global fixed points (say, either all cells black or all cells white). But in other cases—such as [class 3 cellular automata](#)—the number of “output states” may be smaller than the number of “input states” but there may be no computationally simple characterization of them.

“Observers like us”, though, mostly seem to make use of the fixed points. We try to “symbolize the world”, taking all the complexities “out there”, and reducing them to “discrete conclusions”, that we might for example describe using the discrete words in a language.

There’s an immediate subtlety associated with attractors of any kind, though. [Typical physics is reversible](#), in the sense that any process (say two molecules scattering from each other) can run equally well forwards and backwards. But in an attractor one goes from lots of possible initial states to a smaller number of “attractor” final states. And there are [two basic ways this can happen](#), even when there’s underlying reversibility. First, the system one’s studying can be “open”, in the sense that effects can “radiate” out of the region that one’s studying. And second, the states the system gets into can be “complicated enough” that, say, a computationally bounded observer will inevitably equivalence them. And indeed that’s the main thing that’s happening, for example, when a system “reaches thermodynamic equilibrium”, as described by the Second Law.

And actually, once again, there’s often a certain circularity. One is trying to determine whether an observer has “finished observing” and “come to a conclusion”. But one needs an observer to make that determination. Can we tell if we’ve finished “forming a thought”? Well, we have to “think about it”—in effect by forming another thought.

Put another way: imagine we are trying to determine whether a piston has “come to a conclusion” about pressure in a gas. Particularly if there’s microscopic reversibility, the piston and things around it will “continue wiggling around”, and

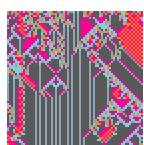
it'll "take an observer" to determine whether the "heat is dissipated" to the point where one can "read out the result".

But how do we break out of what seems like an infinite regress? The point is that whatever mind is ultimately forming the impression that is "the observation" is inevitably the final arbiter. And, yes, this could mean that we'd always have to start discussing all sorts of details about photoreceptors and neurons and so on. But—as we've discussed at length—the key point that makes a general observer theory possible is that there are many conclusions that can be drawn for large classes of observers, quite independent of these details.

But, OK, what happens if we think about the raw ruliad? Now all we have are emes and elementary events updating the configuration of them. And in a sense we're "[fishing out of this](#)" pieces that represent observers, and pieces that represent things they're observing. Can we "assess the cost of observation" here? It really depends on the fundamental scale of what we consider to be observers. And in fact we might even think of our scale as observers (say [measured in emes](#) or elementary events) as defining a "fundamental constant of nature"—at least for the universe as we perceive it. But given this scale, we can for example ask for there to develop "consensus across it", or at least for "every eme in it to have had time to communicate with every other".

In an attempt to formalize the "cost of observation" we'll inevitably have to make what seem like arbitrary choices, just as we would in setting up a scheme to determine when an ongoing computational process has "generated an answer". But if we assume a certain boundedness to our choices, we can expect that we'll be able to draw definite conclusions, and in effect be able to construct an analog of computational complexity theory for processes of observation.

The Future of Observer Theory



My goal here has been to explore some of the key concepts and principles needed to create a framework that we can call observer theory. But what I've done is just the beginning, and there is much still to be done in fleshing out the theory and investigating its implications.

One important place to start is in making more explicit models of the “mechanics of observation”. At the level of the general theory, it’s all about equivalencing. But how specifically is that equivalencing achieved in particular cases? There are many thousands of kinds of sensors, measuring devices, analysis methods, etc. All of these should be systematically inventoried and classified. And in each case there’s a metamodel to be made, that clarifies just how equivalencing is achieved, and, for example, what separation of physical (or other) scales make it possible.

Human experience and human minds are the inspiration—and ultimate grounding—for our concept of an observer. And insofar as neural nets trained on what amounts to human experience have emerged as **somewhat faithful models for what human minds do**, we can expect to use them as a fairly detailed proxy for observers like us. So, for example, we can imagine exploring things like quantum observers by studying multiway generalizations of neural nets. (And this is something that becomes easier if instead of organizing their data into real-number weights we can “atomize” neural nets into purely discrete elements.)

Such investigations of potentially realistic models provide a useful “practical grounding” for observer theory. But to develop a general observer theory we need a more formal notion of an observer. And there is no doubt a whole abstract framework—perhaps using methods from areas like category theory—that can be developed purely on the basis of our concept of observers being about equivalencing.

But to understand the connection of observer theory to things like science as done by us humans, we need to tighten up what it means to be an “observer like us”. What exactly are all the general things we “believe about ourselves”? As we discussed above, many we so much take for granted that it’s challenging for us to identify them as actually just “beliefs” that in principle don’t have to be that way.

But I suspect that the more we can tighten up our definition of “observers like us”, the more we’ll be able to explain why we perceive the world the way we do, and attribute to it the laws and properties we do. Is there some feature of us as observers, for example, that makes us “parse” the physical world as being three-dimensional? We could represent the same data about what’s out there by assigning a one-dimensional (“space-filling”) coordinate to everything. But

somehow observers like us don't do that. And instead, in effect, we "probe the ruliad" by sampling it in what we perceive as 3D slices. (And, yes, the most obvious coarse graining just considers **progressively larger geodesic balls**, say in the spatial hypergraphs that appear in our Physics Project—but that's probably at best just an approximation to the sampling observers like us do.)

As part of our Physics Project we've discovered that the structure of the three main theories of twentieth-century physics (statistical mechanics, general relativity and quantum mechanics) can be derived from properties of the ruliad just by knowing that observers like us are computationally bounded and believe we're persistent in time. But how might we reach, say, the Standard Model of particle physics—with all its particular values of parameters, etc.? Some may be inevitable, given the underlying structure of our theory. But others, one suspects, are in effect reflections of aspects of us as observers. They are "derivable", but only given our particular character—or beliefs—as observers. And, yes, presumably things like the "constant of nature" that characterizes "our size in emes" will appear in the laws we attribute to the universe as we perceive it.

And, by the way, these considerations of "observers like us" extend beyond physical observers. Thus, for example, as we tighten up our characterization of what we're like as **mathematical observers**, we can expect that this will constrain the "possible **laws of our mathematical universe**". We might have thought that we could "pick whatever axioms we want", in effect sampling the ruliad to get any mathematics we want. But, presumably, **observers like us can't do this**—so that questions like "Is the continuum hypothesis true?" can potentially have definite answers for any observers like us, and for any coherent mathematics that we build.

But in the end, do we really have to consider observers whose characteristics are grounded in human experience? We already reflexively generalize our own personal experiences to those of other humans. But can we go further? We don't have the internal experience of being a dog, an ant colony, a computer, or an ocean. And typically at best we anthropomorphize such things, trying to reduce the behavior we perceive in them to elements that align with our own human experience.

But are we as humans just stuck with a particular kind of “internal experience”? The growth of technology—and in particular sensors and measuring devices—has certainly expanded the range of inputs that can be delivered to our brains. And the growth of our collective knowledge about the world has expanded our ways of representing and thinking about things. Right now those are basically our only ways of modifying our detailed “internal experience”. But what if we were to connect directly—and internally—into our brains?

Presumably, at least at first, we’d need the “neural user interface” to be familiar—and we’d be forced into, for example, concentrating everything into a single thread of experience. But what if we allowed “multiway experience”? Well, of course our brains are already made up of billions of neurons that each do things. But it seems to be a core feature of human experience that we concentrate those things to give a single thread of experience. And that seems to be an essential feature of being an “observer like us”.

That kind of concentration also happens in a flock of birds, an ant colony—or a human society. In all these cases, each individual organism “does their thing”. But somehow collective “decisions” get made, with many different detailed situations getting equivalenced together to leave only the “final decision”. So that means that from the outside, the system behaves as we would expect of an “observer like us”. Internally, that kind of “observer behavior” is happening “above the experience” of each single individual. But still, at the level of the “hive mind” it’s behavior typical of an observer like us.

That’s not to say, though, that we can readily imagine what it’s like to be a system like this, or even to be one of its parts. And in the effort to explore observer theory an important direction is to try to imagine ourselves having a different kind of experience than we do. And from “[within](#) that experience”, try to see what kind of laws would we attribute, say, to the physical universe.

In the early twentieth century, particularly in the context of relativity and quantum mechanics, it became clear that being “more realistic” about the observer was crucial in moving forward in science. Things like computational irreducibility—and even more so, our Physics Project—take that another step.

One used to imagine that science should somehow be “fundamentally objective”, and independent of all aspects of the observer. But what’s become clear is that it’s not. And that the nature of us as observers is actually crucial in determining what science we “experience”. But the crucial point is that there are often powerful conclusions that can be drawn even without knowing all the details of an observer. And that’s a central reason for building a general observer theory—in effect to give an objective way of formally and robustly characterizing what one might consider to be the subjective element in science.

Note

There are no doubt many precursors of varying directness that can be found to the things I discuss here; I have not attempted a [serious historical survey](#). In my own work, a notable precursor from 2002 is Chapter 10 of *A New Kind of Science*, entitled “[Processes of Perception and Analysis](#)”. I thank many people involved with our [Wolfram Physics Project](#) for related discussions, including Xerxes Arsiwalla, Hatem Elshatlawy and particularly Jonathan Gorard.

Cite this as >

Stephen Wolfram (2023), "Observer Theory," Stephen Wolfram Writings.
writings.stephenwolfram.com/2023/12/observer-theory.

Posted in: [Big Picture](#), [Language & Communication](#), [New Kind of Science](#), [Philosophy](#), [Physics](#)

Join the discussion

+ 5 comments

Related Writings



How to Think
Computationally about AI, the
Universe and Everything
October 27, 2023



The Concept of the Ruliad
November 10, 2021



Why Does the Universe Exist?
Some Perspectives from Our
Physics Project
April 28, 2021



What Is Consciousness? Some
New Perspectives from Our
Physics Project
March 22, 2021

Popular Categories

Artificial Intelligence

Historical Perspectives

Personal Analytics

Big Picture

Language & Communication

Philosophy

Companies & Business

Life & Times

Physics

Computational Science

Life Science

Ruliology

Computational Thinking

Mathematica

Software Design

Data Science

Mathematics

Wolfram|Alpha

Education

New Kind of Science

Wolfram Language

Future Perspectives

New Technology

Other

Writings by Year

[2025](#) | [2024](#) | [2023](#) | [2022](#) | [2021](#) | [2020](#) | [2019](#) | [2018](#) | [2017](#) | [2016](#) | [2015](#) |

[2014](#) | [2013](#) | [2012](#) | [2011](#) | [2010](#) | [2009](#) | [2008](#) | [2007](#) | [2006](#) | [2004](#) | [2003](#) | [All](#)