Towards Diagonalizing Physics

Gabriel Ehrlich

Begun 30 August 2013

1 30 August 2013

Conventional treatments of physics tend to treat the physicist as an observer external to the system being examined. This pattern holds not only in classical studies but in quantum ones as well. For example, although any measuring system constructed within a quantum context (in the Copenhagen interpretation) would cause wavefunctions to collapse when it observes them, the physicist himself is somehow able to access any wavefunction and predict its behavior without disturbing it. This approach has thus far yielded results of spectacular precision, which vindicate it as an approximation. However, I see an inconsistency in the above treatment which occludes even more accurate theories of measurement.

I'll treat physics a little differently. Instead of the top-down, the-physicist-is-God approach, I'll consider the perspective of a being in a universe governed by some physical laws, with few restrictions. The chosen assumptions suggest a mathematical structure tractable to being diagonalized (in the sense of mathematical logic), with implications for impossibility results regarding the being's ability to know things in the universe.

I view the assumptions that follow as philosophically conservative. I believe them—which, from the perspective of this theory, is the important part—but I also consider them so minimal that everyone who can read this believes them. I consider Descartes's "*Je pense, donc je suis*" presumptuous by comparison (since it's not clear what he means by to exist). Here they are:

1. In this instant, I am thinking (not necessarily autonomously).
2. In this instant, I remember things as if I had existed over a period of time.
3. In this instant, my memories include observations (including observations of thoughts) and actions (not necessarily out of free will).
4. I have observed an isomorphism between my actions and patterns in my observations ("my body") a convincing number of times.

I ended up writing things a little differently from the way I've been thinking about them. I'm not sure how to make the above assumptions mathematically tractable the way the below assumptions seem to be, so I'm going to make some more controversial assumptions towards an ideal observer: the observer has a proper time, and the isomorphism always holds. I'm

also going to drop the assumptions regarding thinking and acting, although they can be reintroduced later to make a more interesting observer.

The situation now is that there is an observer/thinker whose goal is to learn the algorithm generating his observations and thoughts, if any. The observer is represented by a sequence of first-order statements in a certain language, which obeys the rules of first-order deductions augmented with rules which allow the observer to add a statement at position x representing "I thought $\langle \ldots \rangle$", where $\langle \ldots \rangle$ is a proper subsequence of the sequence up until position x. There will be additional rules which allow the observer to make predictions based on what he/she has thought so far. In lay language, the observer does nothing but generate thoughts which are logical axioms, which are consistent with thoughts made so far, which are observations of patterns in thoughts so far, or which are predictions of thoughts to come. The details of the rules are thus far fuzzy.

The question: if the observer's thoughts are generated by a recursive (in the sense of primitive or general recursive) algorithm, will the observer ever predict it?

My guess: no. You'd get the usual time-travel and halting problem paradoxes.

This is a far cry from physics, but it seems within the bounds of possibility that physics is a minor generalization of this; it just has additional observations. However, most physicists believe physics is what generates their thoughts and that physics is recursive; hence, if they also believe that their thoughts obey the restrictions above (which are pretty general), then if it turns out that the observer cannot know his/her own algorithm, then it follows that the observer cannot know all of physics.

I thought I would be writing about how diagonalization can be accomplished through an embedding of the observer's own observation mechanism in the observer's observations. I still believe something like that can be done, but it was getting complicated and I decided to simplify.

2 10 February 2014

I talked to Ben Lehmann about this briefly, leading up to a potential Society of Physics Students talk that Johan Bonilla wanted me to do, and I think I made a little breakthrough. I'm going to simplify the observer a lot, but first we need a physics. Let's call $\mathfrak{U}$ the universe of universes, where each $u \in \mathfrak{U}$ is some function of t, the proper time of the observer. (Recall that a particular universe is inextricable from the observer experiencing it, based on the discussion above.) We're going to require that any universe, which is effectively in the head of the observer, itself contain a physical model of the observer. As such, it should contain a physical model of the observer's brain, which stores the universe itself as observed by the observer.

Formally, we'll require the following for a theory on $\mathfrak{U}$, where $\mathfrak{V}$ is the union of the images of all $u \in \mathfrak{U}$ (i.e. the set of instantaneous states that universes take). Assume the domain of every u is some open interval in $\mathbb{R}$:

Ansatz: For every $u \in \mathfrak{U}$, given the domain $T = (t_1, t_2)$ of u, there exists a function $S : u(T) \to \mathfrak{V}$ such that for all t, $u(t) = S(u(t))$.

This is saying that the observer needs to be embedded in the the universe being observed, and that the extracting function S needs to find that the experience of the observer inside the universe matches the experience of the observer whose universe it is. (Note that the extracting function is independent of the proper time of the observer; I made this choice so that you can't get around this requirement by saying $S(u(t), t)$ is just $u(t)$. However, by making each universe yield ordered pair $\langle t, \ldots \rangle$ you can get around this anyway, so maybe this doesn't work.)

This seems pretty easy to diagonalize.

3 11 February 2014

I think maybe the ansatz should be changed to

Ansatz: There exists a function $S : \mathfrak{V} \to \mathfrak{V}$ (not including functions that are the identity on the physical universes) such that for every physical $u \in \mathfrak{U}$, given the domain $T = (t_1, t_2)$ of u, for all $t \in T$, $u(t) = S(u(t))$.

This deals with the ordered pair pathology, but it might be unsatisfiable. It also introduces the term "physical", which I use to distinguish the universes that are merely well-formed from the ones that actually satisfy the physical laws of the model. Note that choosing $S(v) \equiv v$ (where $v \in \mathfrak{V}$) is like choosing the mind of God, so we're interested in other choices.

4 11 February 2014, again

So one problem with quantum mechanics is that it's difficult to make the connection between what the model looks like and what observers will actually see. For a physical theory to be useful, there has to be a way of connecting the model to predicted observations so that it can be applied to real situations (including tests). Thus there needs to be a way of getting from a model to what the observer will see at each point in time. So I should introduce a new universe, the universe of observations, and THIS one consists of functions of the observer's proper time. The model itself can be whatever (and the extracting function, clearly, has to be the same for every model; it'd be trivial to find a different one that works for each model). Then the models don't actually have to be functions of proper time; they can be whatever. But there needs to be another extracting function which accepts the model and returns a function of the observer's proper time consisting of its predictions. Maybe the function can be probabilistic or something, since in reality (and possibly by logical necessity, assuming these axioms) any predictions of the future have some uncertainty. I don't see anything in particular to diagonalize regarding observations and predictions; saying the observations in the observer's head have to match the observations predicted by the model is just saying the whole ensemble has to match reality. Put another way, the real-world observations aren't mathematically defined; they're produced by an external world that we're trying to understand.

So I'll proceed trying to diagonalize the model itself, under the revised ansatz (for $\mathfrak{U}$ some collection of universes that are not necessarily functions of time, and some of which are physical):

Ansatz: There exists a function $S : \mathfrak{V} \to \mathfrak{V}$ (not including functions that are the identity on the physical universes) such that for every physical $u \in \mathfrak{U}$, $u = S(u)$.

5 12 February 2014

The extraction function thing is all wrong—the *model* needs to specify the observer, as if the observer's brain were saying "That's me!" Then we don't have to deal with the the-entire-model-is-the-observer-because-I'm-God problem (i.e. we don't need to worry about S being the identity—in fact, the requirement as phrased above essentially says it is the identity, although I intended to have more complex structure which simply worked out to be the identity). That's because set theory simply doesn't allow set to contain themselves (as far as I know?).

So instead of the above kind of universe, the new kind of universe considers ordered pairs $\langle u, w \rangle$, where u is a universe of the previous type and w is an observer that is somehow embedded in the universe (I still need to figure out how to guarantee that it's actually represented in the universe itself). Then all the extraction function S does is decode w.

Again, the challenge is now to make sure the observer obeys the same physics as everything else in the model.

6 14 February 2014

Of all the things to do on Valentine's evening.

It's time to define some things.

Theory: A (physical) theory is a sequence $\mathcal{T} = \langle \mathcal{G}, \mathcal{C}, w, Q \rangle$, where $\mathcal{G}$ is a *genre*, $\mathcal{C}$ is a *cast*, W (for "write") is an *author* from $\mathcal{C}$ to $\mathcal{G}$, and Q (for the Q in "critique") is a *critic* on $\mathcal{G}$, satisfying:

> For all $H \in \mathcal{G}$ (H for "history"), there exists $C \subseteq \mathcal{C}$ such that $H = W(C)$. That is, every possible story can be created by combining known character worldlines.

Intuitively, a theory is the assembly of all of the stories examinable by a particular set of physical laws (as opposed to those that are, so to speak, written in a different language—e.g. quantum mechanical wavefunction behavior cannot be analyzed with classical mechanics), together with a description of the particle worldlines in those models, rules for combining those particles into well-formed models, and rules for discriminating between stories that obey the physical laws in question and those that break them.

Genre: A (physical) genre is any nonempty set. Its elements are called (physical) *stories*. Intuitively, this is the set of all the tales that we can look at with our theory and decide between "Yes, this one obeys the physical laws of our theory," and "No, it doesn't."

Cast: A (physical) cast is also any nonempty set. Intuitively, this is the set of the *worldlines* of particles or objects (i.e. their presence throughout the entire story, not merely at a

particular time slice, since in general theories do not have time slices) from which we construct stories. The elements of a cast are called *characters*.

Author: Given a cast $\mathcal{C}$ and a genre $\mathcal{G}$, a (physical) author from $\mathcal{C}$ to $\mathcal{G}$ is a bijection $W : \mathcal{P}(\mathcal{C}) \to \mathcal{G}$. This function takes a list of worldlines and constructs the story corresponding to it. It's injective because the presence of every worldline makes a difference in the story, and it's surjective because every story is a combination of particular worldlines.

Critic: Given a genre $\mathcal{G}$, a (physical) critic Q on $\mathcal{G}$ is a unary relation on $\mathcal{G}$. It says whether a particular story is physical or not, according to some underlying set of physical laws. If the story is physical, it is called a (physical) *model*.

(Note: One thing the definition of theory is missing so far is a method of interpreting it, i.e. taking a model and producing predictions with it. I haven't decided yet whether it needs this level of detail (it would be useful if I decided to look into the narrator's role in interpretation) within the definition of the theory.)

Once we have the traditional kind of theory, we can define first-person theories, in which every story in the genre specifies a *narrator* (observer) from the cast.

First-person Theory: A first-person theory is a theory $\mathcal{T} = \langle \mathcal{G}, \mathcal{C}, W, Q \rangle$ satisfying, if g is some genre:

For all $H \in \mathcal{G}$, $H = \langle h, \mathcal{N} \rangle$ for some $\mathcal{N} \in \mathcal{C}$ and some $h \in g$. That is, each story specifies its narrator.

and where now

$W(C, \mathcal{N}) = \langle w(C), \mathcal{N} \rangle$ for some author w from $\mathcal{C}$ to g, and W is only defined if $\mathcal{N} \in C$ (so that it's bijective). That is, the narrator must be in the story in order to write it.

Now we just need to define the function that takes the narrator's worldline and extracts the story in their head, and then we can state the ansatz.

Analysis: Given a first-person theory $\mathcal{T} = \langle \mathcal{G}, \mathcal{C}, W, Q \rangle$, a (physical) analysis on $\mathcal{T}$ is a function $M : \mathcal{C} \to \mathcal{G}$ (where M is for "meta"). Things this function spits out are called *substories* or *submodels*.

And the ansatz:

Ansatz: A theory $\mathcal{T} = \langle \mathcal{G}, \mathcal{C}, W, Q \rangle$ is only physical if it is first-person and there exists an analysis M such that for all $H \in Q$, $(H)_0 = M((H)_1)$.

The one problem I can see with this so far is that it assumes the narrator will have perfect knowledge of the model. Aha, this is where the prediction part of things comes into play: instead of requiring that the two be the same, we can require that they make the same predictions in the region in which the submodel makes predictions. No, actually that

doesn't seem strong enough—in my model of this universe, the narrator (sub-me) has a a model which precisely matches the model itself.

Also, I wonder if holography is at all related to this. I'm noticing that you kinda have to embed the whole universe (or maybe just part) in the observer, which sounds like holography.

I'm trying to figure out a good place to start with these. It's tempting to start with narrator-only theories based on e.g. classical mechanics, but it seems like that's always possible because just put the entire universe into the narrator part of the sequence. Maybe I'll start with two-character first-person theories based on classical mechanics.

7 15 February 2014

Revisions of the above: the narrator's submodel should change with their proper time—it's not like the narrator believes the world works the same way their entire life. Then the diagonalization we want is that if we associate the model with a t_{now}, the submodel *at that time* should equal the model. This means we'll have two functions of the narrator's worldline: the function which returns observations as a function of time, and the function which returns submodels as a function of time.

We should also require that in two different models, if the narrators' observations up to a certain point are the same, the narrators should have the same submodel. This is where the choice of interpretation (that word might be misleading) becomes important—the same function used to make real-world predictions based off of the model governs the behavior of submodels, and thus via the ansatz the model itself.

Now to formalize this.

8 19 February 2014

Actually the author should not be injective. The cast should include different ensembles of the same fundamental characters so that large-scale things like humans can be observers. When you combine ensembles, it's possible you'll get the same set of worldlines in multiple ways, so we need the author not to be injective.

An alternative is to allow an observer to be a set of characters and keep the author injective. I think this is a bad idea because physical theories may not have fundamental characters. In quantum mechanics, any wavefunction can be written as the linear combination of other wavefunctions; there's no unique fundamental set. If we want quantum mechanics to be a unique theory, we need a unique cast, so this doesn't work.

Also, a thought on the "observations are the same up to a certain point $\Rightarrow$ narrators have the same submodel" ansatz. This has to do with that if I have a particular set of memories of what happened, I should come to a unique conclusion about the initial conditions for my model. But my memories are a property of me *now*, so it should be possible to look at the state of the observer only at t_{now} and figure out those things.

While talking to Han today I figured out how to formalize the requirement that a model's predictions stem only from what the narrator observes. Along similar lines to the above, things the narrator observes should be reflected in the state of the narrator, so instead of

looking at the entire model to get the functions of proper time that return the model's submodel and predictions, we require that we look only at the narrator's worldline to get those functions.

9 30 March 2014

These are some things I've written down on various sticky notes that I wanted to get in here, even though they're not the most updated thoughts I've had.

This one is from sometime last spring: "A particle that thinks about itself: what if a *thing* encoded information (e.g. position) about itself as itself? Then it would make total sense for that information to be an eigenstate of the operator that takes the thing and returns the information. Note connection with Fixed Point Theorem (i.e. Diagonal Lemma). Or... maybe it's a computer that has to calculate its own future."

The rest are all from this quarter.

"Predictions for a given observer are just some FOL model whose universe is what the observer can observe."

"To diagonalize, i.e. disprove that for any event either the model predicts it or the model anti-predicts it, examine statement 'the model anti-predicts this statement' but using the ' "yields falsehood when preceded by its quotation" yields falsewhod when preceded by its quotation' format"

"Prediction fn: $P : Q \to \mathrm{Fn}_{O,t}$ where O is the set of possible observations and $\mathrm{Fn}_{O,t}$ is the set of functions $f : \mathbb{R} \to O$. I.e. prediction function returns a function of time—at each time, that function predicts what observations will be seen. In the case of first-person theories, we need O to be such that $\forall o \in O$, $o = \langle \cdot, h \rangle$ for some $h \in G$. I.e. person notices what submodel they have each time."

"Can have multiple of same character???"

"Use predictions to integrate with language of formal logic. Remove fixed-point requirement?"

"When you take the prediction template 'When you take the prediction template ____ and embed it into itself, you get an incorrect prediction.' and embed it into itself, you get an incorrect prediction."

"Maybe it's not so much about finding a diagonal prediction as about making the quantum mechanical disturb-something-in-order-to-observe-it problem a diagonalization of the existing system. This can most likely be accomplished by introducing submodel-posessing specified characters into the model and making it a two-step strange loop. The reason one step is difficult is that the narrator does not need to resort to rules about the universe as a whole (just the narrator) in order to predict their own submodel (since they're aware of it)."

The last two quotes reflect most accurately my thoughts now: I can find a way of making that diagonal prediction (thought: by making the computation of the answer take some time, so that I don't have it already when I ask the question), but it seems a bit unrelated to the quantum measurement problem. Maybe I should follow along in David H. Wolpert's steps using my new model of physics and see what turns up.