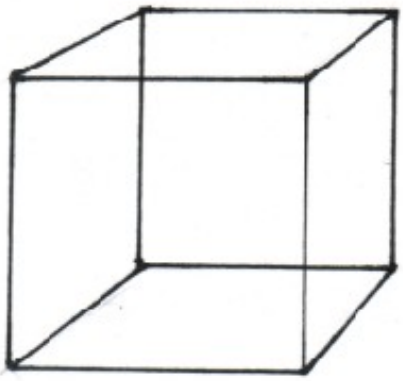
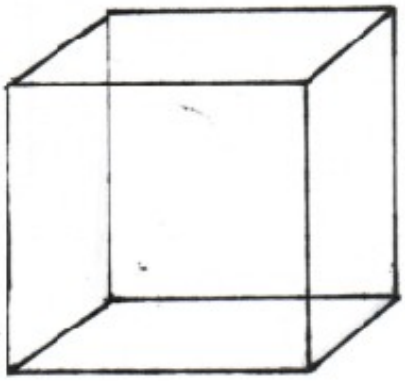


REAL CUBE



NECKER CUBE

Philosophy Corner

1. Philosophy of Numbers (part 1)
2. Philosophy of Numbers (part 2)
3. Dan Dennett on Stances
4. Empirical philosophy and inversions

Philosophy of Numbers (part 1)

This post is the first in a series of things that I think would be fun to discuss on LW. Part two is here.

It seems like there are (at least) two kinds of things we make statements about: physical things, like apples or cities, and logical things, like numbers or logical relations. And it's pretty interesting to question how accurate this seeming is. Are numbers really a "kind of thing," and what do we mean by that anyways? Can we unify these multiple kinds of things, or kinds of statements, into one kind, or not?

For a light review of standard answers, see this nice video. For more depth, you might see the SEP on abstract objects or philosophy of mathematics.

Compare the statements "There exists a city larger than Paris" versus "There exists a number greater than 17." It seems like we use much the same thought patterns to evaluate both these statements, and both seem to be true in the same ordinary sense. Yet the statement about cities seems true because of a correspondence to the external world, but there is no "17" object in a parsimonious predictive model of the world.

To this you might say, "What's the big deal? Even if I don't think numbers are physical objects, it's perfectly reasonable to make this tight analogy between cities and numbers in our reasoning. How is making a big issue out of this going to help us do anything practical?"

Well, in logical decision theory, a recent formulation of some ideas from TDT/UDT, the agent wants to make a causal model of the world that includes (in the model) "causal" effects of a fixed mathematical statement (speficially, the output of the agent's own algorithm). First of all, this is pretty novel and we don't really know how to formalize learning such a model. Second, it's pretty philosophically weird - how is a piece of math supposed to have something like a causal effect on trees and rocks? If we want to solve the practical problem, it might help to be less confused about numbers.

Plus, you know, it's interesting! Why do we think there's such a thing as "numbers," how come the same reasoning works for both numbers and cities, and what are the limits to this analogy, if any?

When one wants to outdo an entire branch of philosophy, it's nice to have some sort of advantage. And the sign of such an advantage is often a bunch of philosphers being loudly wrong about some related issue. But this case, I don't see the signs of an easy advantage. Modern philosophy of numbers doesn't seem to have a bunch of sharp divides or false confidence. Instead, most everyone seems pretty aware that they're confused, despite some fairly interesting ideas being available.

But, okay, I do have some ideas.

See, if you ask philosophers about something that might exist, their first instinct is to try to find a necessary-and-sufficient definition of this thing, more or less on its own terms. Over here, we're much more likely to think of how things are represented in

people's models of the world, and ask what chain of events led people to have that representation, which I think is some important philosophical technology.

This wouldn't be a proper post without a pile of links. So here are some options we might want to keep in mind: Taboo your words. Focus on origin or function, like in the example of "truth." Imagine what cognitive algorithm you're using. Keep your eye on the reductionist ball.

Since this post is labeled "part 1," you might expect that I'm going to end this without telling you exactly what I think about numbers. You'd be right!

But I do want to prompt you with some questions I think are more key than "what is math, really?", and corresponding things I think might be hints.

- **Why do we say that numbers "exist?"**

  Why do we need a property called "existence" in the first place, even just for trees and rocks? Those Eliezer-posts about truth may hint at one point of view.

- **Why would we want to say that certain abstract sentences are "true?"**

  Do statements about math have the same properties Eliezer outlined as making "true" a useful word? Why would we want talk about labels on mathematical sentences if math is just a bunch of tautologies?

- **Does it make sense to evaluate "There exists a city larger than Paris" and "There exists a number greater than 17" the same way?**

  What cognitive algorithms could we be using? What are their disadvantages?

- **Does this line of reasoning actually help us implement LDT?**

  I got nothing.

Philosophy of Numbers (part 2)

A post in a series of things I think would be fun to discuss on LW. Part one is here.

I

As it turns out, I asked my leading questions in precisely the reverse order I'd like to answer them in. I'll start with a simple picture of how we evaluate the truth of mathematical statements, then defend that this makes sense in terms of how we understand "truth," and only last mention existence.

Back to the comparison between "There exists a city larger than Paris" and "There exists a number greater than 17." When we evaluate the statement about Paris we check our map of the world, find that Paris doesn't seem extremely big, and maybe think of some larger cities.

We can use exactly the same thought process on the statement about 17: check our map, quickly recognize that 17 isn't very big, and maybe think of some bigger numbers or the stored principle that there is no largest integer. A large chunk of our issue now collapses into the question "Why does the map containing 17 seem so similar to the map containing Paris?"

<Digression>

> We use the metaphor of map and territory a lot, but let's take a moment to delve a little deeper. My "map" is really more like a huge collection of names, images, memories, scents, impressions, etcetera, all associated with each other in a big web. When I see the word "Paris" I can very quickly figure out how strongly that thing is associated with "city size," and by thinking about "city size" I can tell you some city names that seem more closely-associated with that than "Paris."
>
> "17" is a little trickier, because to explain how I can have associations with "17" in my big web of association, I also need to explain why I don't need a planet-sized brain to hold my impressions of all possible numbers you could have shown me.
>
> The answer is that there's not really a separate token in my head for "17," and not for "Paris" either. My brain doesn't keep a discrete label for everything, instead it stores and manipulates mental representations that are the collective pattern of lots of neurons, and therefore inhabit some high-dimensional space. For example, 17 and 18 might have mental representations that are close together in representation-space. And I can easily represent 87438 despite never having thought about that number before, because I can map the symbols to the right point in representation-space.

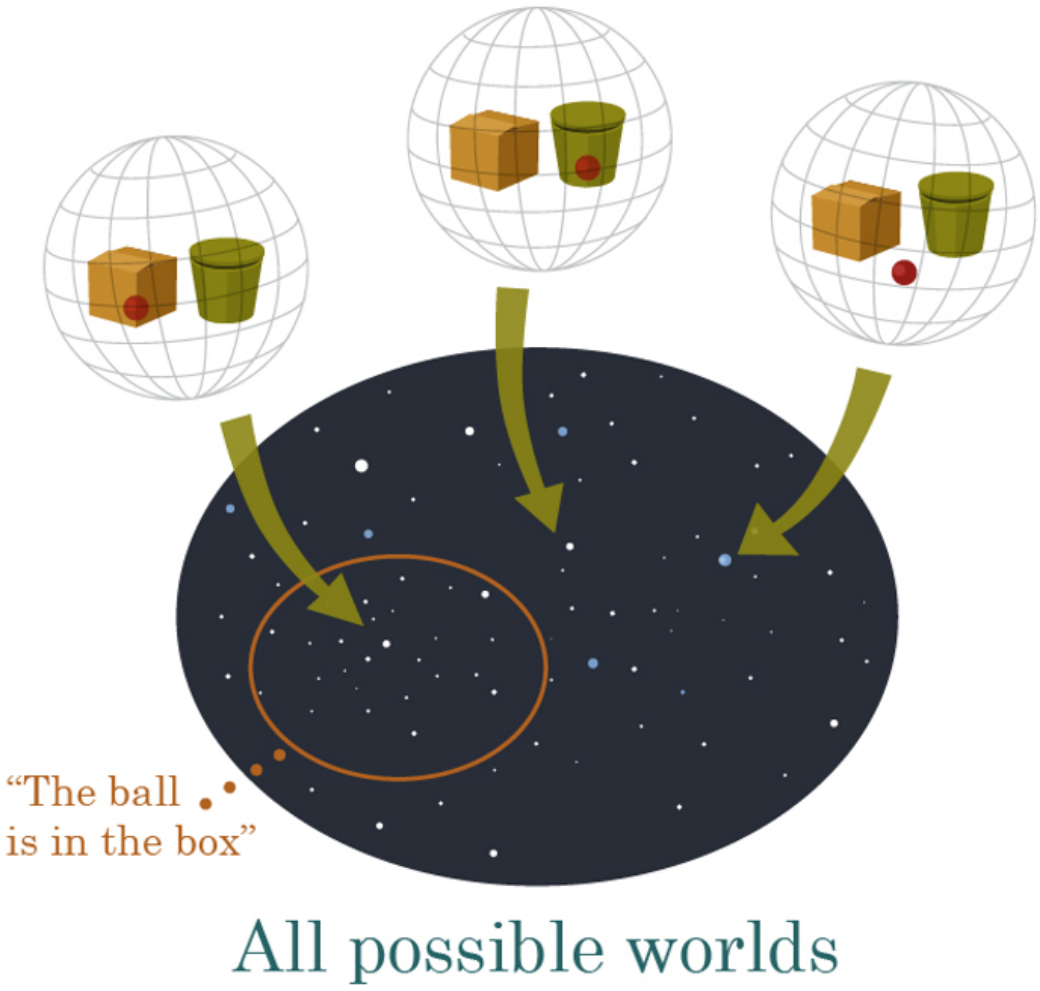
</Digression>

If we really do evaluate mathematical statements the same way we evaluate statements about our map of the external world, then that would explain why both evaluations seem to return the same type of "true" or "false." It's also convenient for evaluating the truth of mixed mathematical and empirical statements like "The number of pens on my table is less than 3 factorial." But we still need to fit this apparent-truth of mathematical statements with our conception of truth as a correspondence between map and territory.

II

An important fact about our models of the world is that they're capable of modeling things that aren't real. Suppose our world contains a red ball. We might hypothesize many different world-models and variations on models, each with a different past and future trajectory for

the red ball. Psychologically, this feels like we are imagining different possible worlds, at most one of which can be real.



To make a statement like "The ball is in the box" is to imply that we are in one specific fraction of the possible worlds. This statement is false in some possible worlds and true in others, but we should only endorse that the ball is in the box if, in our one true world, the ball is actually in the box.

Each statement about the red ball that we can evaluate as true or false can be thought of as defining a set of the possible worlds where that statement is true. "The volume of the ball contains a neutrino" is true in almost every world, while "The ball is in a volcano" is true in almost none. Knowing true statements gives us helps us narrow down which possible world we're actually in.

<Digression> More technically, knowing true statements helps us pick models that predict the world well. All this talk of possible worlds is a convenient metaphor. </Digression>

Moving closer to the point: "The ball has bounced a prime number of times" also defines a perfectly valid set of possible worlds. So. Does "3 is a prime number" define a set of possible worlds?

If we were really committed to answering "no" to this, we would have to undergo strange contortions, like being able to evaluate "The ball has bounced three times and the ball has bounced a prime number of times," but not "The ball has bounced three times and three is a prime number." Being able to compare the empirical with the abstract suggests the ability to compare the abstract with the abstract.

If we answer "yes," the set of possible worlds where 3 is a prime number seems like "all of them." (Or perhaps only almost all of them.) Math is then a bunch of tautologies.

But this raises an important problem: if mathematical truths are tautologous, then that would seem to render having a mental map of mathematics unnecessary - you can just evaluate statements purely on whether they obey their axioms. Conversely, if mathematical statements are always true or always false, then they're not useful, because learning them doesn't refine our predictions of the world. To resolve this apparent problem, we'll need a very powerful force: human ignorance.

Even though mathematical statements are theoretically evaluable from a small set of axioms, in practice that is much, much too hard for humans to do at runtime. Instead, we have to build up our knowledge of math slowly, associate important results with each other and with their real-world applications, and be able to place new knowledge in context of the old.

So it is precisely human badness at math that makes us keep a mental map of mathematics that's structured like our map of the world. The fact that our map doesn't start completely filled in also means that we can learn new things about math. It also leads directly into my last leading question from part one: why might we think numbers exist?

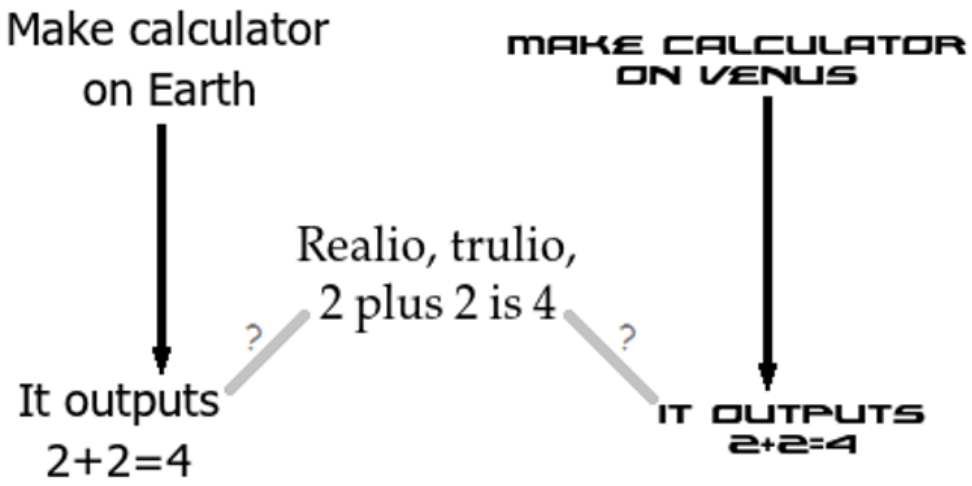
III

The reasons to feel like numbers exist are pretty similar to the reasons to feel like the physical world exists. For starters, our observations don't always turn out how we'd predict. The stuff that generates the predictions, we call belief, and the stuff that generates the observations, we call reality.

Sometimes, you have beliefs about mathematical statements even if you can't prove them. You might think, say, P!=NP, not by reasoning from the axioms, but by reasoning from the shape of your map. And when this heuristic reasoning fails, as it occasionally does, it feels like you're encountering an external reality, even if there's no causal thing that could be providing the feedback.

We also feel more like things exist when we model them as objective, rather than subjective. When we use our model of the world to imagine changing peoples' opinions about an objective thing, our model says that the objective thing doesn't change. Mathematical truths fulfil this property nicely - details left to the reader.

Lastly, things that we think exist have relationships with other elements in our map of the world. Things are associated with properties, like color and size - numbers definitely have properties. And although numbers are not connected to rocks in a causal model of the world, it seems like we say "2+2=4" because 2+2=4. But the "because" back there is not a causal relationship - rather it's an association our brain makes that's something like logical implication.



So maybe I do understand those mysterious links in LDT (artist's representation above) better than I did before. They're a toy-model representation of a connection that seems very natural in our brains, between different things that we have in the same map of the world.

Epilogue

I played a bit coy in this post - I talk a big game about understanding numbers, but here we are at the end and rather than telling you whether numbers really exist or not, I've just harped on what makes people *feel* like things exist.

To give away the game completely: I avoided the question because whether numbers "really exist" can end up getting stuck in the center node of the classic blegg/rube classifier. When faced with a red egg, the solution is usually not to figure out if it's "really a blegg or a rube." The solution is to be able to think about it as a red egg. And the even better solution is to understand the function of sorting these objects so that we can use categorizations in contexts where it's useful.

Understanding why we feel the way we do about numbers is really an exercise in looking at the surrounding nodes. The core claim of this article is that two things that normally agree - "should be a basic object in a parsimonous causal model of the world" and "can usefully be thought about using certain expectations and habits developed for physical objects" - diverge here, and so we should strive to replace tension about whether numbers "really exist" with understanding of how we think about numbers.

My aim was for a standard LW-ian view of numbers. I feel like I learned a lot writing this, and hopefully some of that feeling rubs off on the reader. (Thank you for reading, by the way.) I'll be back with something completely different next week.

Dan Dennett on Stances

This is a linkpost for
https://ase.tufts.edu/cogstud/dennett/papers/intentionalsystems.pdf

Part of a series. Index is here.

I'm writing up a post on consciousness, and I realized that this is worth linking separately. The gist of it is that we can think of a person as having feelings and intentions, and we can think of a person as a colleciton of atoms obeying the laws of physics, and both are valid ways of thinking. The person is the same thing out in reality either way - it's just the model we're using that changes, and, specifically, what kind of predictions our model is good at making.

You can really see where Dennett was influenced by Less Wrong. Especially in his work from the `70s.

Anyhow, I think this review is a real Dennett classic, up there with Eliminate The Middletoad. If you have a favorite in a similar vein, I'd love to hear about it.

Empirical philosophy and inversions

A regular installment. Index is here.

This post is in large part a linkpost for an excellent talk on experimental philosophy, given by Ned Block (don't be put off by the title): https://www.youtube.com/watch?v=6lHHxcxurhQ . Apologies to those who dislike videos, but (especially at 1.5x or 2x speed) it's faster and more fun than reading a bunch of his papers, I swear.

Here's an example: one experiment Block talks about sticks electrodes to peoples' heads, and then subtly shows them a geometrical shape while they're doing another task. In the after-experiment report, some participants report they noticed the shape, and their electrode data can then be reviewed to see what brain activity was neccessary for noticing the shape even if they didn't know they had to notice it. It turns out, your brain doesn't need to be very active for you to be able to recall seeing the shape.

Block uses results like this to defend his thesis about the richness of conscious perception, and how early in the brain's perceptual systems activity can be experienced consciously. But this forms an interesting contrast with a deflationary view of consciousness.

Our agent of deflation is Marvin Minsky. Here's a video of him being deflationist. He has a favorite point, which is that people associate consciousness with lots of tasks, like being able to remember smells, or being able to imagine applying verbs to nouns, et cetera, but that this grouping is a human-made category, and thinking about these things as a group can get in the way of understanding them. The stuff we call conscious activity can, he says, be broken up into lots of sub-processes like smelling and abstract-verb-imagining that have a strong internal coherence but not much overlap with each other.

Which brings us back to Ned Block and consciousness of perception. It's possible to look at the several experiments Block talks about, not as different probes of a unified consciousness, but as probes of several functions of the brain that fall under the umbrella of consciousness.

Another of Block's examples is presenting different images to each eye, and using eye-tracking to determine which image the subject is experiencing. It's natural and effortless for us to think that this sort of consciousness is the same thing as the consciousness of remembering the shape, form the first example. It takes a weird, effortful inversion of perspective for me to think about what if would be like if the brain functions determining the two experiments had very little overlap.

This is the reason I linked to Dennett's article on the intentional stance earlier - Minsky's view can be thought of as deliniating a "conscious stance" as separate from a "process stance." In this view, consciousness is just a convenient way to predict mental things without looking too close. And so from this view, the kinds of brain activity the empirical philosophers are trying to pin down - where do you store representation of things you see, how fast do you put those representations into long-term memory, what where do you figure out which eye's signals are dominant in

determining your representation of the visual field, et cetera - are actually at a level of description below consciousness.

You may already be grumbling about the distinction between hard and easy problems of consciousness. These grumblings are fair, and we will get to that later. I just thought this was too much fun to not share.