
"The ball
is in the box"
All possible worlds

Highly Advanced Epistemology 101 for Beginners

1. The Useful Idea of Truth
2. Skill: The Map is Not the Territory
3. Rationality: Appreciating Cognitive Algorithms
4. Firewalling the Optimal from the Rational
5. The Fabric of Real Things
6. Causal Diagrams and Causal Models
7. Stuff That Makes Stuff Happen
8. Causal Reference
9. Causal Universes
10. Proofs, Implications, and Models
11. Logical Pinpointing
12. Standard and Nonstandard Numbers
13. Godel's Completeness and Incompleteness Theorems
14. Second-Order Logic: The Controversy
15. Mixed Reference: The Great Reductionist Project
16. By Which It May Be Judged

The Useful Idea of Truth

(This is the first post of a new Sequence, Highly Advanced Epistemology 101 for Beginners, setting up the Sequence Open Problems in Friendly AI. For experienced readers, this first post may seem somewhat elementary; but it serves as a basis for what follows. And though it may be conventional in standard philosophy, the world at large does not know it, and it is useful to know a compact explanation. Kudos to Alex Altair for helping in the production and editing of this post and Sequence!)

I remember this paper I wrote on existentialism. My teacher gave it back with an F. She'd underlined true and truth wherever it appeared in the essay, probably about twenty times, with a question mark beside each. She wanted to know what I meant by truth.
-- Danielle Egan

I understand what it means for a hypothesis to be elegant, or falsifiable, or compatible with the evidence. It sounds to me like calling a belief 'true' or 'real' or 'actual' is merely the difference between saying you believe something, and saying you really really believe something.
-- Dale Carrico

What then is truth? A movable host of metaphors, metonymies, and; anthropomorphisms: in short, a sum of human relations which have been poetically and rhetorically intensified, transferred, and embellished, and which, after long usage, seem to a people to be fixed, canonical, and binding.
-- Friedrich Nietzche

The Sally-Anne False-Belief task is an experiment used to tell whether a child understands the difference between belief and reality. It goes as follows:

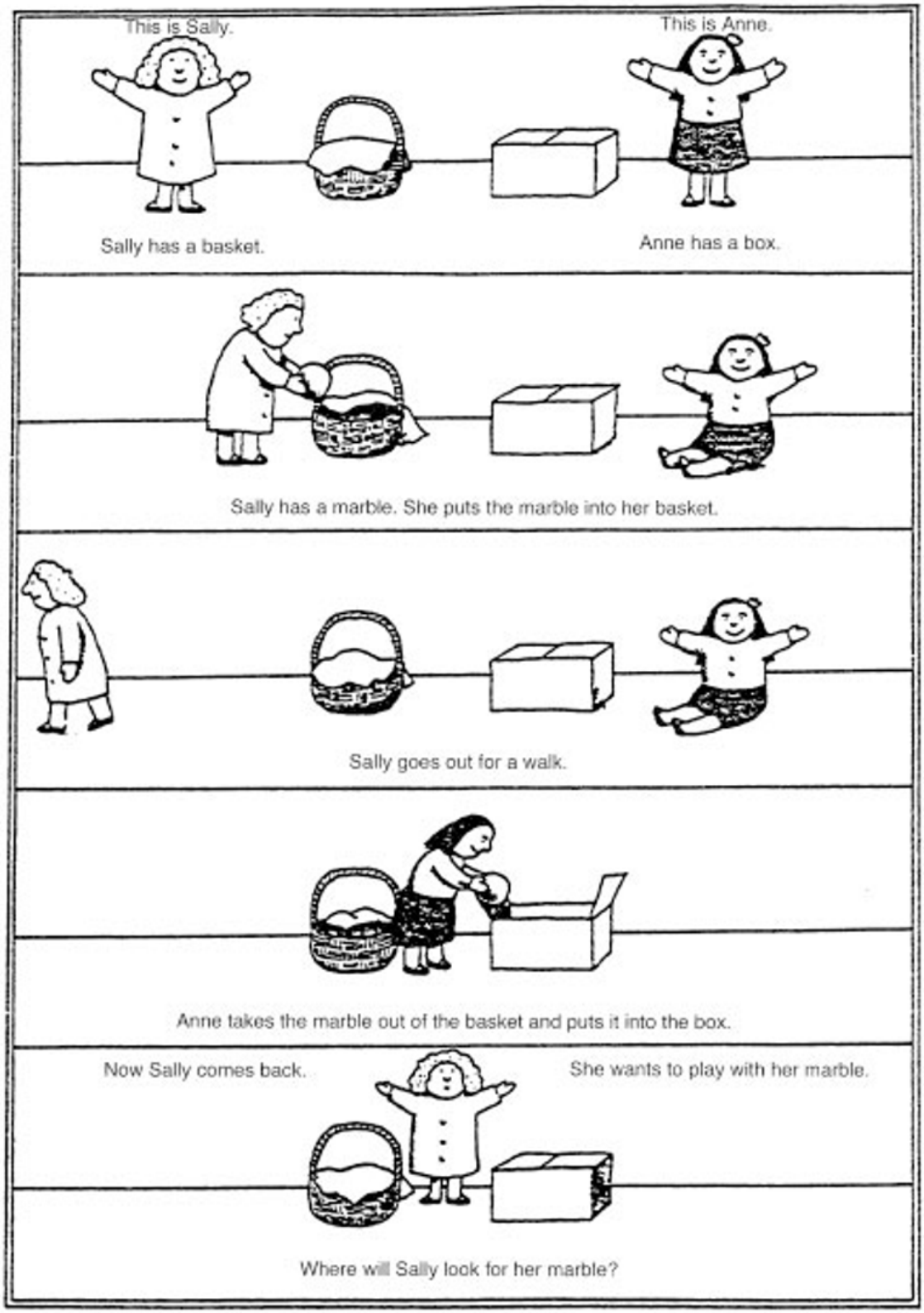
The child sees Sally hide a marble inside a covered basket, as Anne looks on.

Sally leaves the room, and Anne takes the marble out of the basket and hides it inside a lidded box.

Anne leaves the room, and Sally returns.

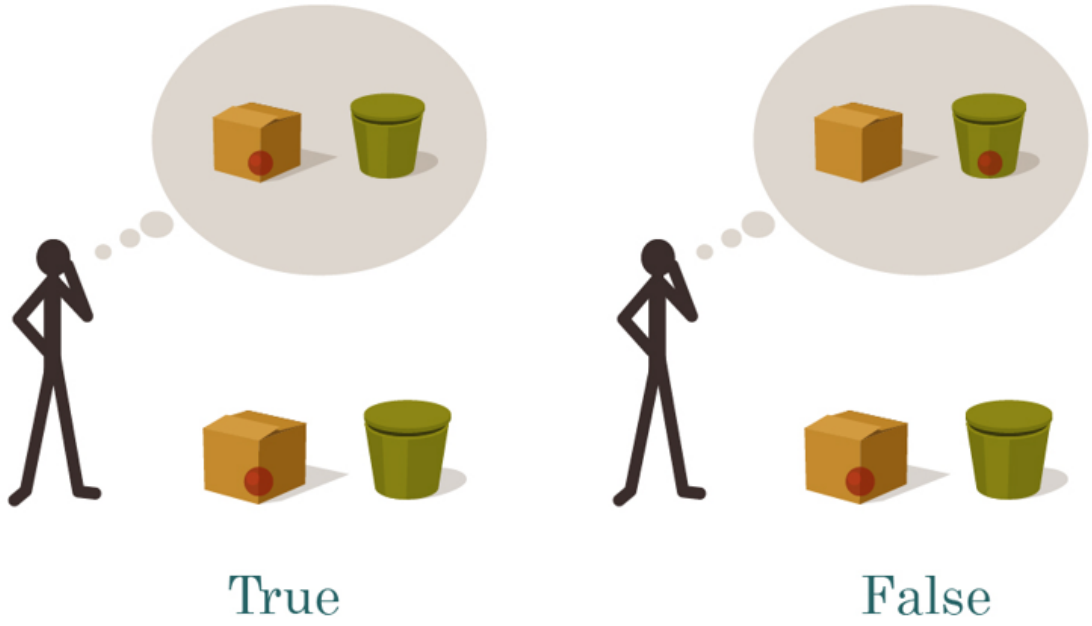
The experimenter asks the child where Sally will look for her marble.

Children under the age of four say that Sally will look for her marble inside the box. Children over the age of four say that Sally will look for her marble inside the basket.

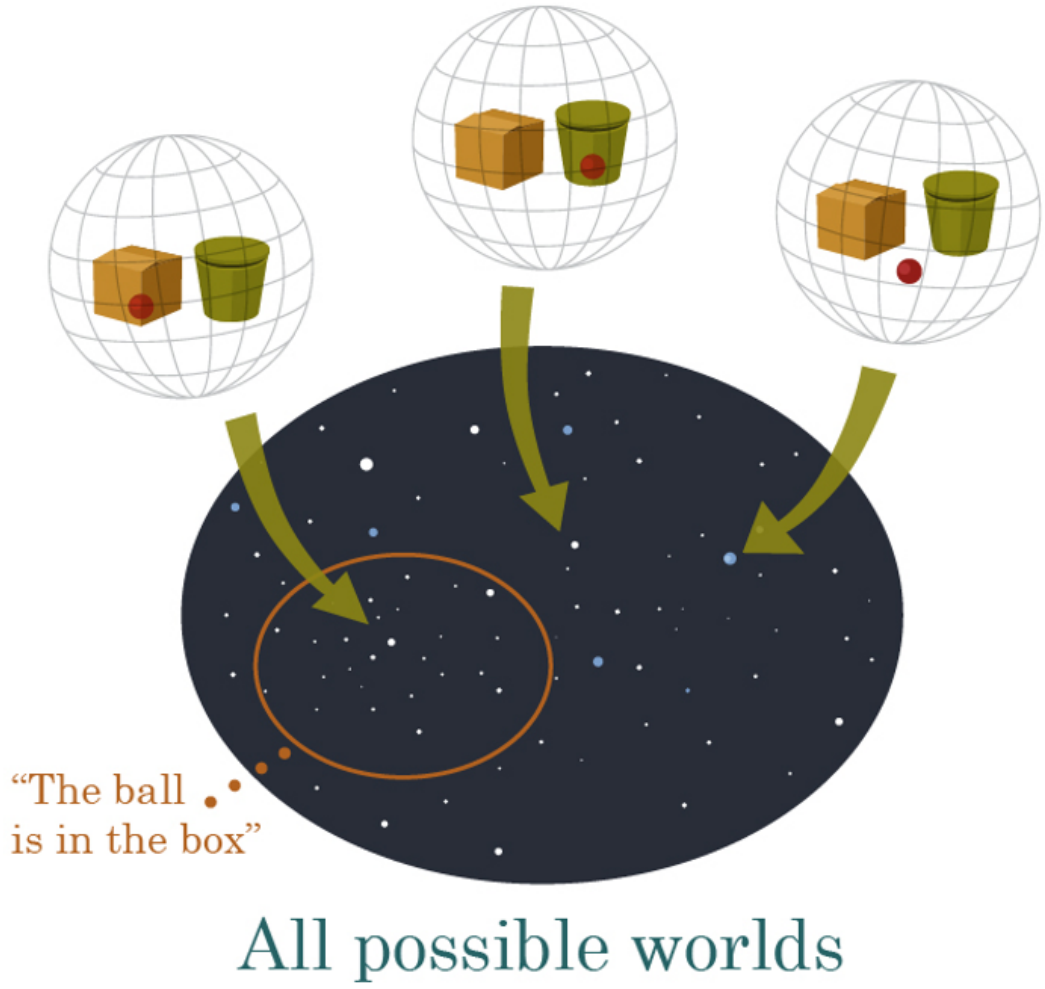


(Attributed to: Baron-Cohen, S., Leslie, L. and Frith, U. (1985) 'Does the autistic child have a "theory of mind"?', Cognition, vol. 21, pp. 37–46.)

Human children over the age of (typically) four, first begin to understand what it means for Sally to lose her marbles - for Sally's beliefs to stop corresponding to reality. A three-year-old has a model only of *where the marble is.* A four-year old is developing a theory of mind; they separately model *where the marble is* and *where Sally believes the marble is,* so they can notice when the two conflict - when Sally has a false belief.



Any meaningful belief has a *truth-condition,* some way reality can be which can make that belief true, or alternatively false. If Sally's brain holds a mental image of a marble inside the basket, then, in reality itself, the marble can actually be inside the basket - in which case Sally's belief is called 'true', since reality falls inside its truth-condition. Or alternatively, Anne may have taken out the marble and hidden it in the box, in which case Sally's belief is termed 'false', since reality falls outside the belief's truth-condition.



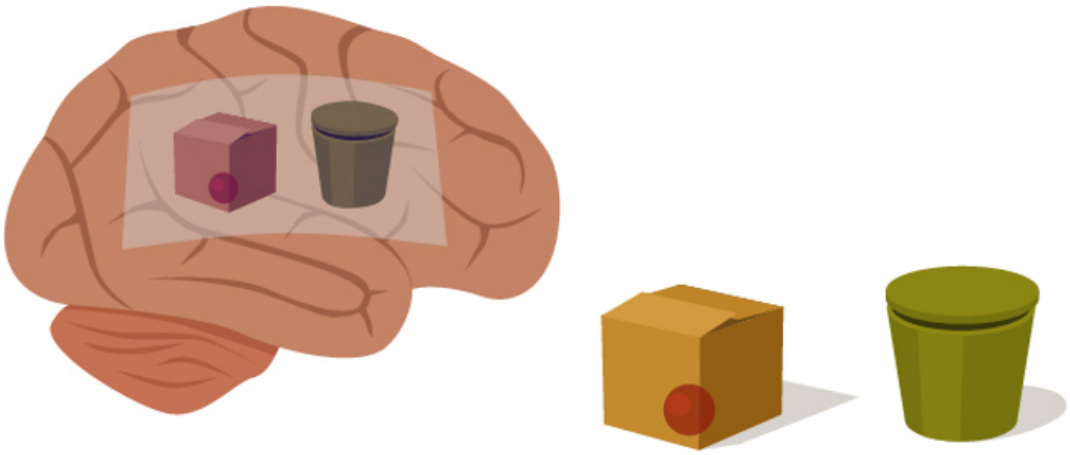
The mathematician Alfred Tarski once described the notion of 'truth' via an infinite family of truth-conditions:

The sentence 'snow is white' is true if and only if snow is white.

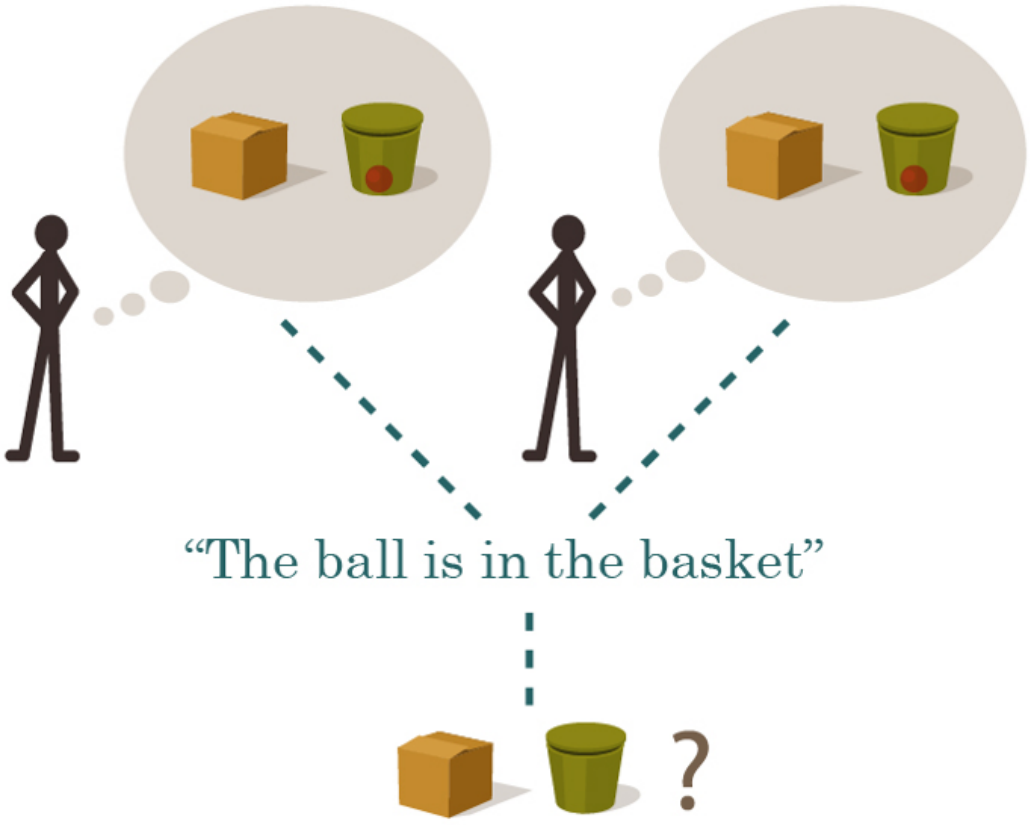
The sentence 'the sky is blue' is true if and only if the sky is blue.

When you write it out that way, it looks like the distinction might be trivial - indeed, why bother talking about sentences at all, if the sentence looks so much like reality when both are written out as English?

But when we go back to the Sally-Anne task, the difference looks much clearer:
Sally's *belief* is embodied in a pattern of neurons and neural firings inside Sally's brain, three pounds of wet and extremely complicated tissue inside Sally's skull. The *marble itself* is a small simple plastic sphere, moving between the basket and the box. When we compare Sally's belief to the marble, we are comparing two quite different things.

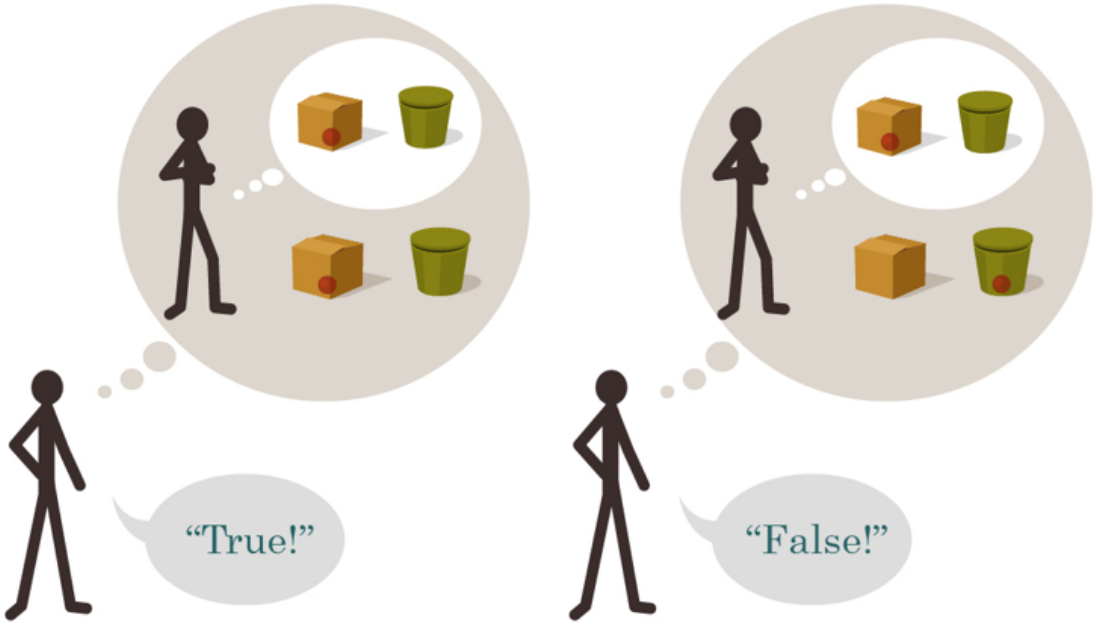


(Then why talk about these abstract 'sentences' instead of just neurally embodied beliefs? Maybe Sally and Fred believe "the same thing", i.e., their brains both have internal models of the marble inside the basket - two brain-bound beliefs with the same truth condition - in which case the thing these two beliefs have in common, the shared truth condition, is abstracted into the form of a *sentence* or *proposition* that we imagine being true or false apart from any brains that believe it.)



Some pundits have panicked over the point that any judgment of *truth* - any comparison of belief to reality - takes place inside some particular person's mind; and indeed seems to just

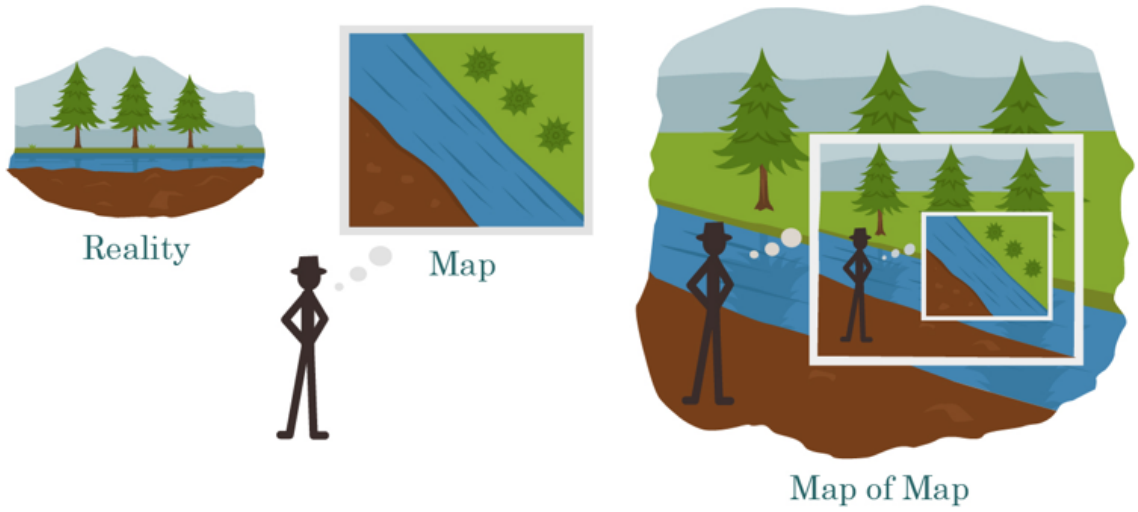
compare *someone else's belief* to *your belief:*



So is all this talk of truth just comparing other people's beliefs to our own beliefs, and trying to assert privilege? Is the word 'truth' just a weapon in a power struggle?

For that matter, you can't even *directly* compare other people's beliefs to our own beliefs. You can only internally compare your *beliefs* about someone else's belief to your own belief - compare your map of their map, to your map of the territory.

Similarly, to say of your own beliefs, that the belief is 'true', just means you're comparing *your map of your map,* to *your map of the territory.* People *usually* are not mistaken about what they themselves believe - though there are certain exceptions to this rule - yet nonetheless, the map of the map is usually accurate, i.e., people are usually right about the question of *what they believe:*



And so saying 'I believe the sky is blue, and that's true!' typically conveys the same information as 'I believe the sky is blue' or just saying 'The sky is blue' - namely, that your mental model of the world contains a blue sky.

Meditation:

If the above is true, aren't the postmodernists right? Isn't all this talk of 'truth' just an attempt to assert the privilege of your own beliefs over others, when there's nothing that can actually compare a belief to reality itself, outside of anyone's head?

(A 'meditation' is a puzzle that the reader is meant to attempt to solve before continuing. It's my somewhat awkward attempt to reflect the research which shows that you're much more likely to remember a fact or solution if you try to solve the problem yourself before reading the solution; succeed or fail, the important thing is to have tried first . This also reflects a problem Michael Vassar thinks is occurring, which is that since LW posts often sound obvious in retrospect, it's hard for people to visualize the diff between 'before' and 'after'; and this diff is also useful to have for learning purposes. So please try to say your own answer to the meditation - ideally whispering it to yourself, or moving your lips as you pretend to say it, so as to make sure it's fully explicit and available for memory - before continuing; and try to consciously note the difference between your reply and the post's reply, including any extra details present or missing, without trying to minimize or maximize the difference.)

...
...
...

Reply:

The reply I gave to Dale Carrico - who declaimed to me that he knew what it meant for a belief to be falsifiable, but not what it meant for beliefs to be true - was that my *beliefs* determine my experimental *predictions,* but only *reality* gets to determine my experimental *results.* If I believe very strongly that I can fly, then this belief may lead me to step off a cliff, expecting to be safe; but only the *truth* of this belief can possibly save me from plummeting to the ground and ending my experiences with a splat.

Since my expectations sometimes conflict with my subsequent experiences, I need different names for the thingies that determine my experimental predictions and the thingy that determines my experimental results. I call the former thingies 'beliefs', and the latter thingy 'reality'.

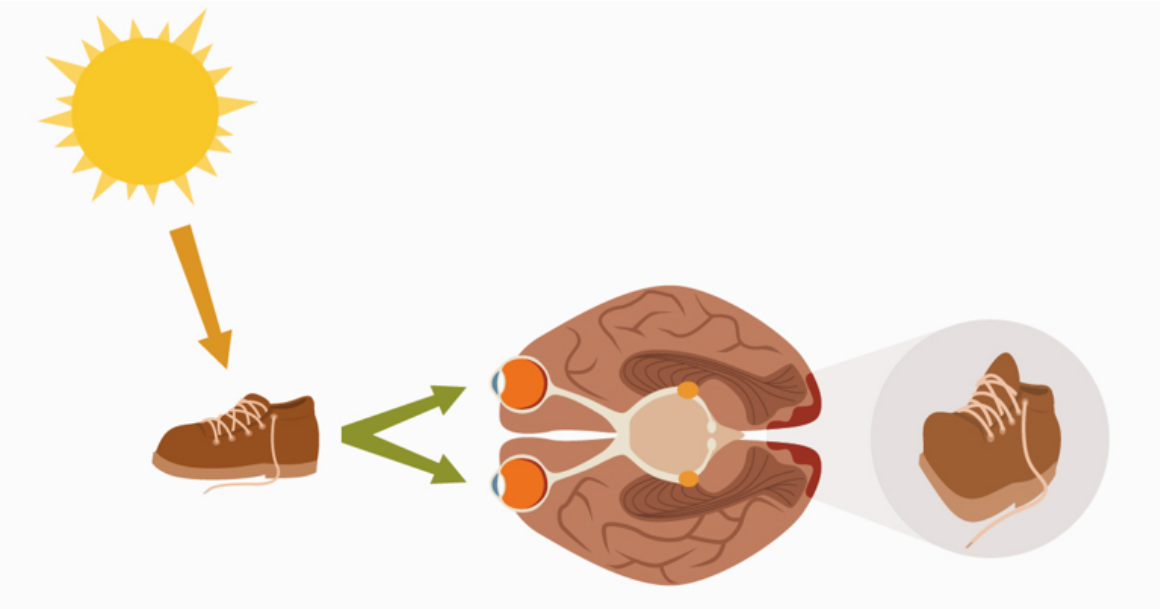
You won't get a direct collision between belief and reality - or between someone else's beliefs and reality - by sitting in your living-room with your eyes closed. But the situation is different if you open your eyes!

Consider how your brain ends up knowing that its shoelaces are untied:

- A photon departs from the Sun, and flies to the Earth and through Earth's atmosphere.
- Your shoelace absorbs and re-emits the photon.
- The reflected photon passes through your eye's pupil and toward your retina.
- The photon strikes a rod cell or cone cell, or to be more precise, it strikes a photoreceptor, a form of vitamin-A known as *retinal,* which undergoes a change in its molecular shape (rotating around a double bond) powered by absorption of the photon's energy. A bound protein called an *opsin* undergoes a conformational change in response, and this further propagates to a neural cell body which pumps a proton and increases its polarization.
- The gradual polarization change is propagated to a bipolar cell and then a ganglion cell. If the ganglion cell's polarization goes over a threshold, it sends out a *nerve impulse,* a propagating electrochemical phenomenon of polarization-depolarization that travels through the brain at between 1 and 100 meters per second. Now the incoming light from the outside world has been transduced to neural information, commensurate with the substrate of other thoughts.
- The neural signal is preprocessed by other neurons in the retina, further preprocessed by the lateral geniculate nucleus in the middle of the brain, and then, in the visual cortex located at the back of your head, reconstructed into *an actual little tiny*

picture of the surrounding world - a picture embodied in the firing frequencies of the neurons making up the visual field. (A distorted picture, since the center of the visual field is processed in much greater detail - i.e. spread across more neurons and more cortical area - than the edges.)

- Information from the visual cortex is then routed to the temporal lobes, which handle object recognition.
- Your brain recognizes the form of an untied shoelace.



And so your brain updates its map of the world to include the fact that your shoelaces are untied. Even if, previously, it expected them to be tied! There's no reason for your brain *not* to update if politics aren't involved. Once photons heading into the eye are turned into neural firings, they're commensurate with other mind-information and can be compared to previous beliefs.

Belief and reality interact *all the time*. If the environment and the brain never touched in any way, we wouldn't need eyes - or hands - and the brain could afford to be a *whole* lot simpler. In fact, organisms wouldn't need brains at all.

So, fine, belief and reality are distinct entities which do intersect and interact. But to say that we need separate concepts for 'beliefs' and 'reality' doesn't get us to needing the concept of 'truth', a comparison between them. Maybe we can just separately (a) talk about an agent's belief that the sky is blue and (b) talk about the sky itself. Instead of saying, "Jane believes the sky is blue, and she's right", we could say, "Jane believes 'the sky is blue'; also, the sky is blue" and convey the same information about what (a) we believe about the sky and (b) what we believe Jane believes. We could always apply Tarski's schema - "The sentence 'X' is *true* iff X" - and replace every instance of alleged truth by talking directly about the truth-condition, the corresponding state of reality (i.e. the sky or whatever). Thus we could eliminate that bothersome word, 'truth', which is so controversial to philosophers, and misused by various annoying people.

Suppose you had a rational agent, or for concreteness, an Artificial Intelligence, which was carrying out its work in isolation and certainly never needed to argue politics with anyone. The AI knows that "My model assigns 90% probability that the sky is blue"; it is quite sure that this probability is the exact statement stored in its RAM. Separately, the AI models that "The probability that my optical sensors will detect blue out the window is 99%, *given* that the sky is blue"; and it doesn't confuse this proposition with the quite different proposition that the optical sensors will detect blue whenever it *believes* the sky is blue. So the AI can

definitely differentiate the map and the territory; it knows that the possible states of its RAM storage do not have the same consequences and causal powers as the possible states of sky.

But does this AI ever need a concept for the notion of *truth in general* - does it ever need to invent the word 'truth'? Why would it work better if it did?

Meditation: If we were dealing with an Artificial Intelligence that never had to argue politics with anyone, would it ever need a word or a concept for 'truth'?

...
...
...

Reply: The abstract concept of 'truth' - the general idea of a map-territory correspondence - is required to express ideas such as:

Generalized across possible maps and possible cities, if your map of a city is accurate, navigating according to that map is more likely to get you to the airport on time.

To draw a true map of a city, someone has to go out and look at the buildings; there's no way you'd end up with an accurate map by sitting in your living-room with your eyes closed trying to imagine what you wish the city would look like.

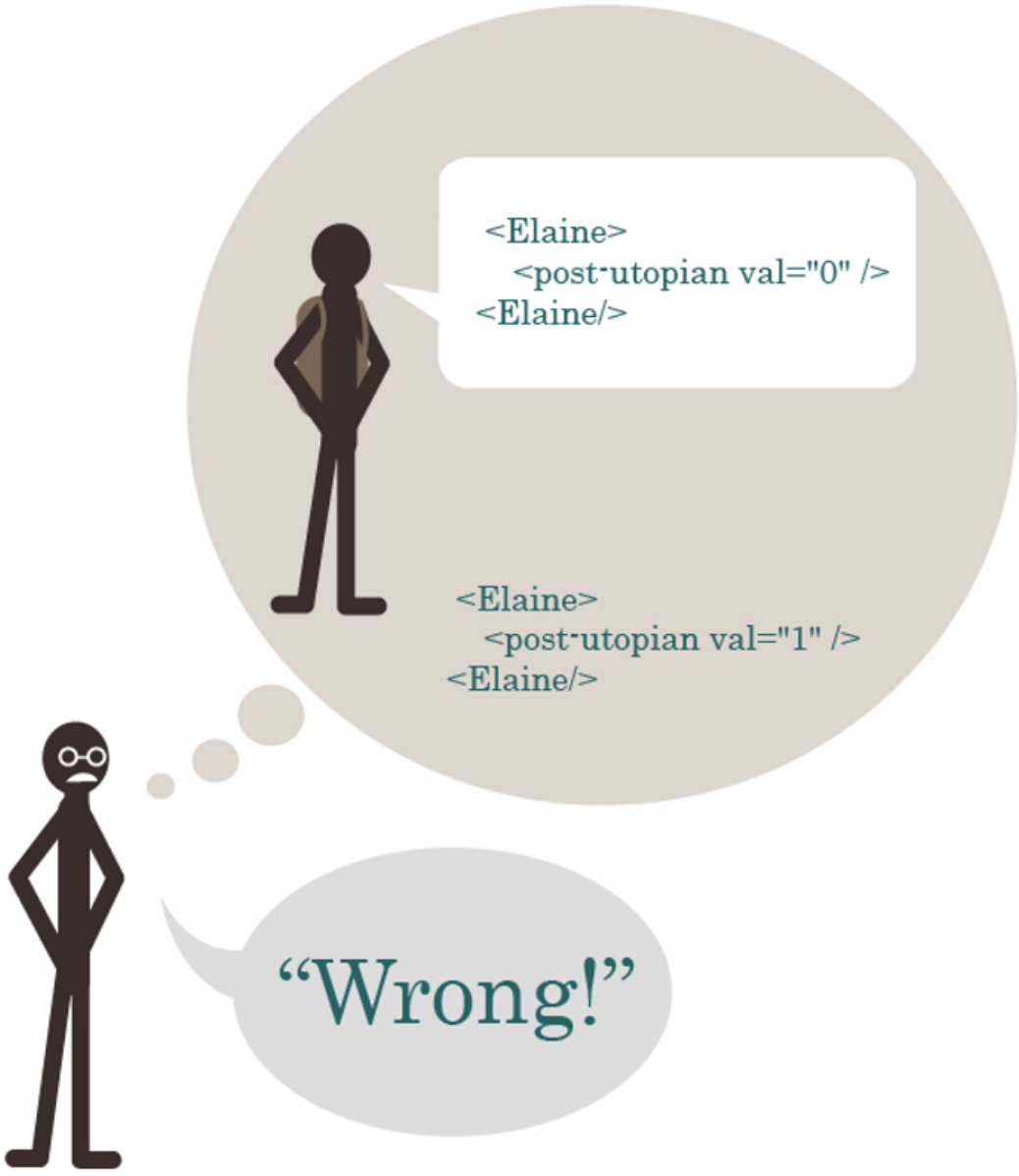
True beliefs are more likely than false beliefs to make correct experimental predictions, so if we increase our credence in hypotheses that make correct experimental predictions, our model of reality should become incrementally more true over time.

This is the main benefit of talking and thinking about 'truth' - that we can generalize rules about how to make maps match territories *in general;* we can learn lessons that transfer beyond particular skies being blue.

Next in main sequence:

Complete philosophical panic has turned out not to be justified (it never is). But there is a key practical problem that results from our internal evaluation of 'truth' being a comparison of a map of a map, to a map of reality: On this schema it is *very easy* for the brain to end up believing that a *completely meaningless* statement is 'true'.

Some literature professor lectures that the famous authors Carol, Danny, and Elaine are all 'post-utopians', which you can tell because their writings exhibit signs of 'colonial alienation'. For most college students the typical result will be that their brain's version of an object-attribute list will assign the attribute 'post-utopian' to the authors Carol, Danny, and Elaine. When the subsequent test asks for "an example of a post-utopian author", the student will write down "Elaine". What if the student writes down, "I think Elaine is *not* a post-utopian"? Then the professor models thusly...



...and marks the answer *false.*

After all...

The sentence "Elaine is a post-utopian" is *true* if and only if Elaine is a post-utopian.

...right?

Now of course it could be that this term *does* mean something (even though I made it up). It might even be that, although the professor can't give a good explicit answer to "What *is* post-utopianism, anyway?", you can nonetheless take many literary professors and separately show them new pieces of writing by unknown authors and they'll all independently arrive at the same answer, in which case they're clearly detecting *some* sensory-visible feature of the writing. We don't always know how our brains work, and we don't always know what we see, and the sky was seen as blue long before the word

"blue" was invented; for a part of your brain's world-model to be meaningful doesn't require that you can explain it in words.

On the other hand, it could also be the case that the professor learned about "colonial alienation" by memorizing what to say to *his* professor. It could be that the only person whose brain assigned a real meaning to the word is dead. So that by the time the students are learning that "post-utopian" is the password when hit with the query "colonial alienation?", both phrases are *just* verbal responses to be rehearsed, *nothing but* an answer on a test.

The two phrases don't feel "disconnected" individually because they're connected to each other - post-utopianism has the apparent consequence of colonial alienation, and if you ask what colonial alienation implies, it means the author is probably a post-utopian. But if you draw a circle around both phrases, they don't connect to anything *else.* They're *floating beliefs* not connected with the rest of the model. And yet there's no internal alarm that goes off when this happens. Just as "being wrong feels like being right" - just as having a false belief feels the same internally as having a true belief, at least until you run an experiment - having a meaningless belief can *feel* just like having a meaningful belief.

(You can even have fights over completely meaningless beliefs. If someone says "Is Elaine a post-utopian?" and one group shouts "Yes!" and the other group shouts "No!", they can fight over having shouted different things; it's not necessary for the words to *mean* anything for the battle to get started. Heck, you could have a battle over one group shouting "Mun!" and the other shouting "Fleem!" More generally, it's important to distinguish the visible consequences of the professor-brain's *quoted* belief (students had better write down a certain thing on his test, or they'll be marked wrong) from the proposition that there's an *unquoted state of reality* (Elaine *actually* being a post-utopian in the territory) which has visible consquences.)

One classic response to this problem was *verificationism,* which held that the sentence "Elaine is a post-utopian" is *meaningless* if it doesn't tell us which sensory experiences we should expect to see if the sentence is true, and how those experiences differ from the case if the sentence is false.

But then suppose that I transmit a photon aimed at the void between galaxies - heading far off into space, away into the night. In an expanding universe, this photon will eventually cross the *cosmological horizon* where, even if the photon hit a mirror reflecting it squarely back toward Earth, the photon would never get here because the universe would expand too fast in the meanwhile. Thus, after the photon goes past a certain point, there are *no experimental consequences whatsoever, ever,* to the statement "The photon continues to exist, rather than blinking out of existence."

And yet it seems to me - and I hope to you as well - that the statement "The photon suddenly blinks out of existence as soon as we can't see it, violating Conservation of Energy and behaving unlike all photons we can actually see" is *false,* while the statement "The photon continues to exist, heading off to nowhere" is *true.* And this sort of question can have important policy consequences: suppose we were thinking of sending off a near-light-speed colonization vessel as far away as possible, so that it would be over the cosmological horizon before it slowed down to colonize some distant supercluster. If we thought the colonization ship would just blink out of existence before it arrived, we wouldn't bother sending it.

It is both useful and wise to ask after the sensory consequences of our beliefs. But it's not quite the *fundamental* definition of meaningful statements. It's an excellent *hint* that something might be a disconnected 'floating belief', but it's not a hard-and-fast rule.

You might next try the answer that for a statement to be meaningful, there must be some *way reality can be* which makes the statement true or false; and that since the universe is made of atoms, there must be some way to *arrange the atoms in the universe* that would make a statement true or false. E.g. to make the statement "I am in

Paris" true, we would have to move the atoms comprising myself to Paris. A literateur claims that Elaine has an attribute called post-utopianism, but there's no way to translate this claim into a way to *arrange the atoms in the universe* so as to make the claim true, or alternatively false; so it has no truth-condition, and must be meaningless.

Indeed there are claims where, if you pause and ask, "How could a universe be arranged so as to make this claim true, or alternatively false?", you'll suddenly realize that you didn't have as strong a grasp on the claim's truth-condition as you believed. "Suffering builds character", say, or "All depressions result from bad monetary policy." These claims aren't necessarily meaningless, but they're a lot easier to *say*, than to visualize the universe that makes them true or false. Just like asking after sensory consequences is an important hint to meaning or meaninglessness, so is asking how to configure the universe.

But if you say there has to be some arrangement of *atoms* that makes a meaningful claim true or false...

Then the theory of quantum mechanics would be meaningless *a priori*, because there's no way to arrange *atoms* to make the theory of quantum mechanics true.

And when we discovered that the universe was not made of atoms, but rather quantum fields, all *meaningful* statements everywhere would have been revealed as *false* - since there'd be no atoms arranged to fulfill their truth-conditions.

Meditation: What rule could restrict our beliefs to *just* propositions that can be meaningful, without excluding a priori anything that could in principle be true?

- **Meditation Answers -** *(A central comment for readers who want to try answering the above meditation (before reading whatever post in the Sequence answers it) or read contributed answers.)*
- **Mainstream Status -** *(A central comment where I say what I think the status of the post is relative to mainstream modern epistemology or other fields, and people can post summaries or excerpts of any papers they think are relevant.)*

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

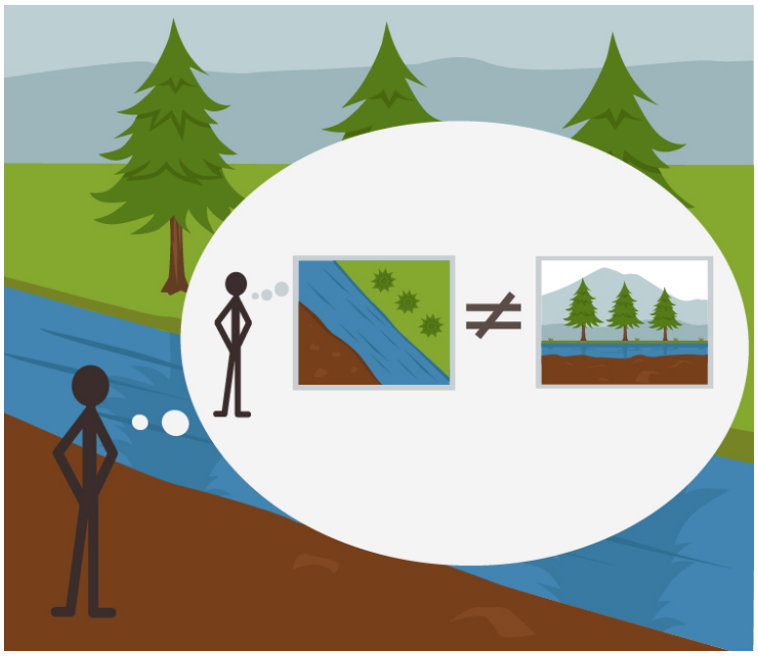
Next post: "Skill: The Map is Not the Territory"

Skill: The Map is Not the Territory

Followup to: The Useful Idea of Truth *(minor post)*

So far as I know, the first piece of rationalist fiction - one of only two explicitly rationalist fictions I know of that didn't descend from HPMOR, the other being "David's Sling" by Marc Stiegler - is the Null-A series by A. E. van Vogt. In Vogt's story, the protagonist, Gilbert Gosseyn, has mostly non-duplicable abilities that you can't pick up and use even if they're supposedly mental - e.g. the ability to use all of his muscular strength in emergencies, thanks to his alleged training. The main explicit-rationalist skill someone could *actually* pick up from Gosseyn's adventure is embodied in his slogan:

"The map is not the territory."



Sometimes it still amazes me to contemplate that this proverb was *invented* at some point, and some fellow named Korzybski invented it, and this happened as late as the 20th century. I read Vogt's story and absorbed that lesson when I was rather young, so to me this phrase sounds like a sheer background axiom of existence.

But as the Bayesian Conspiracy enters into its second stage of development, we must all accustom ourselves to translating mere insights into applied techniques. So:

Meditation: Under what circumstances is it helpful to consciously think of the distinction between the map and the territory - to visualize your thought bubble containing a belief, and a reality outside it, rather than just using your map to think about reality directly? How exactly does it help, on what sort of problem?

...

...

...

Skill 1: The conceivability of being wrong.

In the story, Gilbert Gosseyn is most liable to be reminded of this proverb when some belief is uncertain; "Your belief in that does not make it so." It might sound basic, but this *is* where some of the earliest rationalist training starts - making the jump from living in a world where the sky just *is* blue, the grass just *is* green, and people from the Other Political Party just *are* possessed by demonic spirits of pure evil, to a world where it's possible that *reality* is going to be different from these beliefs and come back and surprise you. You might assign low probability to that in the grass-is-green case, but in a world where there's a territory separate from the map it is at least *conceivable* that reality turns out to disagree with you. There are people who could stand to rehearse this, maybe by visualizing themselves with a thought bubble, first in a world like X, then in a world like not-X, in cases where they are tempted to entirely neglect the possibility that they might be wrong. "He hates me!" and other beliefs about other people's motives seems to be a domain in which "I believe that he hates me" or "I hypothesize that he hates me" might work a lot better.

Probabilistic reasoning is also a remedy for similar reasons: Implicit in a 75% probability of X is a 25% probability of not-X, so you're hopefully automatically considering more than one world. Assigning a probability also inherently reminds you that you're occupying an epistemic state, since only beliefs can be probabilistic, while reality itself is either one way or another.

Skill 2: Perspective-taking on beliefs.

What we really believe feels like the way the world *is;* from the inside, other people *feel like* they are inhabiting different worlds from you. They aren't disagreeing with you because they're obstinate, they're disagreeing because the world feels different to them - even if the two of you are in fact embedded in the same reality.

This is one of the secret writing rules behind Harry Potter and the Methods of Rationality. When I write a character, e.g. Draco Malfoy, I don't just extrapolate their mind, I extrapolate the surrounding subjective world they live in, which has that character at the center; all other things seem important, or are considered at all, in relation to how important they are to that character. Most other books are never told from more than one character's viewpoint, but if they are, it's strange how often the other characters seem to be living inside the protagonist's universe and to think mostly about things that are important to the main protagonist. In HPMOR, when you enter Draco Malfoy's viewpoint, you are plunged into Draco Malfoy's subjective universe, in which Death Eaters have reasons for everything they do and Dumbledore is an exogenous reasonless evil. Since I'm not trying to show off postmodernism, everyone is still recognizably living in the same underlying reality, and the justifications of the Death Eaters only sound reasonable *to Draco,* rather than having been optimized to persuade the reader. It's not like the characters *literally* have their own universes, nor is morality handed out in equal portions to all parties regardless of what they do. But different elements of reality have different meanings and different importances to different characters.

Joshua Greene has observed - I think this is in his Terrible, Horrible, No Good, Very Bad paper - that most political discourse rarely gets beyond the point of lecturing naughty children who are just refusing to acknowledge the evident truth. As a special case, one

may also appreciate internally that being wrong feels just like being right, unless you can actually perform some sort of experimental check.

Skill 3: You are less bamboozleable by anti-epistemology or motivated neutrality which explicitly claims that there's no truth.

This is a *negative* skill - avoiding one more wrong way to do it - and mostly about quoted arguments rather than positive reasoning you'd want to conduct yourself. Hence the sort of thing we want to put less emphasis on in training. Nonetheless, it's easier not to fall for somebody's line about the absence of objective truth, if you've previously spent a bit of time visualizing Sally and Anne with different beliefs, and separately, a marble for those beliefs to be compared-to. Sally and Anne have different *beliefs*, but there's only one way-things-are, the actual state of the marble, to which the beliefs can be compared; so no, they don't have 'different truths'. A real belief (as opposed to a belief-in-belief) will *feel* true, yes, so the two have different feelings-of-truth, but the feeling-of-truth is not the territory.

To rehearse this, I suppose, you'd try to notice this kind of anti-epistemology when you ran across it, and maybe respond internally by actually visualizing two figures with thought bubbles and their single environment. Though I don't *think* most people who understood the core insight would require any further persuasion or rehearsal to avoid contamination by the fallacy.

Skill 4: World-first reasoning about decisions a.k.a. the Tarski Method aka Litany of Tarski.

Suppose you're considering whether to wash your white athletic socks with a dark load of laundry, and you're worried the colors might bleed into the socks, but on the other hand you really don't want to have to do another load just for the white socks. You might find your brain selectively rationalizing reasons why it's not all *that* likely for the colors to bleed - there's no really new dark clothes in there, say - trying to persuade itself that the socks won't be ruined. At which point it may help to say:

"If my socks will stain, I want to believe my socks will stain;
If my socks won't stain, I don't want to believe my socks will stain;
Let me not become attached to beliefs I may not want."

To stop your brain trying to persuade itself, visualize that you are either *already in* the world where your socks will end up discolored, or already in the world where your socks will be fine, and in either case it is better for you to believe you're in the world you're actually in. Related mantras include "That which can be destroyed by the truth should be" and "Reality is that which, when we stop believing in it, doesn't go away". Appreciating that belief is not reality can help us to appreciate the primacy of reality, and either stop arguing with it and accept it, or actually become curious about it.

Anna Salamon and I usually apply the Tarski Method by visualizing a world that is not-how-we'd-like or not-how-we-previously-believed, and ourselves as believing the contrary, and the disaster that would then follow. For example, let's say that you've been driving for a while, haven't reached your hotel, and are starting to wonder if you took a wrong turn... in which case you'd have to go back and drive another 40 miles in the opposite direction, which is an unpleasant thing to think about, so your brain tries to persuade itself that it's not lost. Anna and I use the form of the skill where we visualize the world where we *are* lost and *keep driving.*

Note that in principle, this is only one quadrant of a 2 x 2 matrix:

	In **reality**, you're heading in the right direction	In **reality**, you're totally lost
You **believe** you're heading in the right direction	No need to change anything - just keep doing what you're doing, and you'll get to the conference hotel	Just keep doing what you're doing, and you'll eventually drive your rental car directly into the sea
You **believe** you're lost	Alas! You spend 5 whole minutes of your life pulling over and asking for directions you didn't need	After spending 5 minutes getting directions, you've got to turn around and drive 40 minutes the other way.

Michael "Valentine" Smith says that he practiced this skill by actually visualizing all four quadrants in turn, and that with a bit of practice he could do it very quickly, and that he thinks visualizing all four quadrants helped.

(Mainstream status here.)

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Rationality: Appreciating Cognitive Algorithms"

Previous post: "The Useful Idea of Truth"

Rationality: Appreciating Cognitive Algorithms

Followup to: The Useful Idea of Truth

It is an error mode, and indeed an annoyance mode, to go about preaching the importance of the "Truth", especially if the Truth is supposed to be something incredibly lofty instead of some boring, mundane truth about gravity or rainbows or what your coworker said about your manager.

Thus it is a worthwhile exercise to practice deflating the word 'true' out of any sentence in which it appears. (Note that this is a special case of rationalist taboo.) For example, instead of saying, "I believe that the sky is blue, and that's *true!*" you can just say, "The sky is blue", which conveys essentially the same information about what color you think the sky is. Or if it feels *different* to say "I believe the Democrats will win the election!" than to say, "The Democrats will win the election", this is an important warning of belief-alief divergence.

Try it with these:

- I believe Jess just wants to win arguments.
- It’s true that you weren’t paying attention.
- I believe I will get better.
- In reality, teachers care a lot about students.

If 'truth' is defined by an infinite family of sentences like 'The sentence "the sky is blue" is true if and only if the sky is blue', then why would we ever need to talk about 'truth' at all?

Well, you can't deflate 'truth' out of the sentence "True beliefs are more likely to make successful experimental predictions" because it states a property of map-territory correspondences *in general*. You could say 'accurate maps' instead of 'true beliefs', but you would still be invoking the same *concept.*

It's only because most sentences containing the word 'true' are *not* talking about map-territory correspondences in general, that most such sentences can be deflated.

Now consider - when are you *forced* to use the word 'rational'?

As with the word 'true', there are very few sentences that truly *need* to contain the word 'rational' in them. Consider the following deflations, all of which convey essentially the same information about your own opinions:

- "It's rational to believe the sky is blue"
  -> "I think the sky is blue"
  -> "The sky is blue"

- "Rational Dieting: Why To Choose Paleo"
  -> "Why you should think the paleo diet has the best consequences for health"
  -> "I like the paleo diet"

Generally, when people bless something as 'rational', you could directly substitute the word 'optimal' with no loss of content - or in some cases the phrases 'true' or 'believed-by-me', if we're talking about a belief rather than a strategy.

Try it with these:

- "It's rational to teach your children calculus."
- "I think this is the most rational book ever."
- "It's rational to believe in gravity."

Meditation: Under what rare circumstances can you *not* deflate the word 'rational' out of a sentence?

...
...
...

Reply: We need the word 'rational' in order to talk about *cognitive algorithms* or *mental processes* with the property "systematically increases map-territory correspondence" (epistemic rationality) or "systematically finds a better path to goals" (instrumental rationality).

E.g.:

"It's (epistemically) rational to believe more in hypotheses that make successful experimental predictions."

or

"Chasing sunk costs is (instrumentally) irrational."

You can't deflate the *concept* of rationality out of the intended meaning of those sentences. You could find some way to rephrase it without the *word* 'rational'; but then you'd have to use other words describing the same concept, e.g:

"If you believe more in hypotheses that make successful predictions, your map will better correspond to reality over time."

or

"If you chase sunk costs, you won't achieve your goals as well as you could otherwise."

The word 'rational' is properly used to talk about *cognitive algorithms* which *systematically* promote map-territory correspondences or goal achievement.

Similarly, a rationalist isn't just somebody who respects the Truth.

All too many people respect the Truth.

They respect the Truth that the U.S. government planted explosives in the World Trade Center, the Truth that the stars control human destiny (ironically, the exact reverse will be true if everything goes right), the Truth that global warming is a lie... and so it goes.

A rationalist is somebody who respects the *processes of finding truth.* They respect somebody who seems to be showing genuine curiosity, even if that curiosity is about a should-already-be-settled issue like whether the World Trade Center was brought down by explosives, because genuine curiosity is part of a lovable algorithm and respectable process. They respect Stuart Hameroff for trying to test whether neurons have properties conducive to quantum computing, even if this idea seems exceedingly unlikely a priori and was suggested by awful Gödelian arguments about why brains can't be mechanisms, because Hameroff was *trying to test his wacky beliefs experimentally,* and humanity would still be living on the savanna if 'wacky' beliefs never got tested experimentally.

Or consider the controversy over the way CSICOP (Committee for Skeptical Investigation of Claims of the Paranormal) handled the so-called Mars effect, the controversy which led founder Dennis Rawlins to leave CSICOP. Does the position of the planet Mars in the sky during your hour of birth, *actually* have an effect on whether you'll become a famous athlete? I'll go out on a limb and say no. And if you *only* respect the Truth, then it doesn't matter very much whether CSICOP raised the goalposts on the astrologer Gauquelin - i.e., stated a test and then made up new reasons to reject the results after Gauquelin's result came out positive. The astrological conclusion is almost certainly un-true... and that conclusion was indeed derogated, the Truth upheld.

But a *rationalist* is disturbed by the claim that there were *rational process violations*. As a Bayesian, in a case like this you do update to a very small degree in favor of astrology, just not enough to overcome the prior odds; and you update to a larger degree that Gauquelin has inadvertantly uncovered some other phenomenon that might be worth tracking down. One definitely shouldn't state a test and then ignore the results, or find new reasons the test is invalid, when the results don't come out your way. That process has bad *systematic* properties for finding truth - and a rationalist doesn't just appreciate the beauty of the Truth, but the beauty of the processes and cognitive algorithms that get us there.[1]

The reason why rationalists can have unusually productive and friendly conversations *at least when everything goes right,* is not that everyone involved has a great and abiding respect for whatever they think is the True or the Optimal in any given moment. Under most everyday conditions, people who argue heatedly aren't doing so because they know the truth but disrespect it. Rationalist conversations are (potentially) more productive to the degree that everyone respects the *process,* and is on mostly the same page about what the process should be, thanks to all that explicit study of things like cognitive psychology and probability theory. When Anna tells me, "I'm worried that you don't seem very curious about this," there's this state of mind called 'curiosity' that we both agree is important - as a matter of *rational process,* on a meta-level above the particular issue at hand - and I know as a matter of process that when a respected fellow rationalist tells me that I need to become curious, I should pause and check my curiosity levels and try to increase them.

Is rationality-use necessarily tied to rationality-appreciation? I can imagine a world filled with hordes of rationality-users who were taught in school to use the Art competently, even though only very few people love the Art enough to try to advance it further; and everyone else has no particular love or interest in the Art apart from the practical results it brings. Similarly, I can imagine a competent applied mathematician who only worked at a hedge fund for the money, and had never loved math or programming or optimization in the first place - who'd been in it for the money from day one. I can imagine a competent musician who had no particular love in

composition or joy in music, and who only cared for the album sales and groupies. Just because something is imaginable doesn't make it probable in real life... but then there are many children who learn to play the piano despite having no love for it; "musicians" are those who are *unusually* good at it, not the adequately-competent.

But for now, in this world where the Art is *not* yet forcibly impressed on schoolchildren nor yet explicitly rewarded in a standard way on standard career tracks, almost everyone who has any skill at rationality is the sort of person who finds the Art intriguing for its own sake. Which - perhaps unfortunately - explains quite a bit, both about rationalist communities and about the world.

---

[1] RationalWiki really needs to rename itself to SkepticWiki. They're very interested in kicking hell out of homeopathy, but not as a group interested in the abstract beauty of questions like "What trials should a strange new hypothesis undergo, which it will *not* fail if the hypothesis is true?" You can go to them and be like, "You're criticizing theory X because some people who believe in it are stupid; but many true theories have stupid believers, like how Deepak Chopra claims to be talking about quantum physics; so this is not a useful method in general for discriminating true and false theories" and they'll be like, "Ha! So what? Who cares? X is crazy!" I think it was actually RationalWiki which first observed that it and Less Wrong ought to swap names.

(Mainstream status here.)

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Firewalling the Optimal from the Rational"

Previous post: "Skill: The Map is Not the Territory"

Firewalling the Optimal from the Rational

Followup to: Rationality: Appreciating Cognitive Algorithms *(minor post)*

There's an old anecdote about Ayn Rand, which Michael Shermer recounts in his "The Unlikeliest Cult in History" (*note: calling a fact unlikely is an insult to your prior model, not the fact itself*), which went as follows:

> Branden recalled an evening when a friend of Rand's remarked that he enjoyed the music of Richard Strauss. "When he left at the end of the evening, Ayn said, in a reaction becoming increasingly typical, 'Now I understand why he and I can never be real soulmates. The distance in our sense of life is too great.' Often she did not wait until a friend had left to make such remarks."

Many readers may already have appreciated this point, but one of the Go stones placed to block that failure mode is being careful what we bless with the great community-normative-keyword 'rational'. And one of the ways we do that is by trying to deflate the word 'rational' out of sentences, especially in post titles or critical comments, which can live without the word. As you hopefully recall from the previous post, we're only *forced* to use the word 'rational' when we talk about the *cognitive algorithms* which *systematically* promote goal achievement or map-territory correspondences. Otherwise the word can be deflated out of the sentence; e.g. "It's rational to believe in anthropogenic global warming" goes to "Human activities are causing global temperatures to rise"; or "It's rational to vote for Party X" deflates to "It's optimal to vote for Party X" or just "I think you should vote for Party X".

If you're writing a post comparing the experimental evidence for four different diets, that's not "Rational Dieting", that's "Optimal Dieting". A post about *rational* dieting is if you're writing about how the sunk cost fallacy causes people to eat food they've already purchased even if they're not hungry, or if you're writing about how the typical mind fallacy or law of small numbers leads people to overestimate how likely it is that a diet which worked for them will work for a friend. And even then, your title is 'Dieting and the Sunk Cost Fallacy', unless it's an overview of four different cognitive biases affecting dieting. In which case a *better* title would be 'Four Biases Screwing Up Your Diet', since 'Rational Dieting' carries an implication that your post discusses *the* cognitive algorithm for dieting, as opposed to four contributing things to keep in mind.

By the same token, a post about Givewell's top charities and how they compare to existential-risk mitigation is a post about *optimal philanthropy,* while a post about scope insensitivity and hedonic returns vs. marginal returns is a post about *rational philanthropy,* because the first is discussing object-level outcomes while the second is discussing cognitive algorithms. And either way, if you can have a post title that doesn't include the word "rational", it's probably a good idea because the word gets a little less powerful every time it's used.

Of course, it's still a good idea to include *concrete examples* when talking about general cognitive algorithms. A good writer won't discuss rational philanthropy without including some discussion of particular charities to illustrate the point. In general, the *concrete-abstract* writing pattern says that your opening paragraph

should be a concrete example of a nonoptimal charity, and only afterward should you generalize to make the abstract point. (That's why the main post opened with the Ayn Rand anecdote.)

And I'm not saying that we should never have posts about Optimal Dieting on LessWrong. What good is all that rationality if it never leads us to anything optimal?

Nonetheless, the *second* Go stone placed to block the Objectivist Failure Mode is trying to *define ourselves as a community* around the cognitive algorithms; and trying to avoid membership tests (especially implicit *de facto* tests) that *aren't* about rational process, but just about some particular thing that a lot of us think is *optimal*.

Like, say, paleo-inspired diets.

Or having to love particular classical music composers, or hate dubstep, or something. (Does anyone know any good dubstep mixes of classical music, by the way?)

Admittedly, a lot of the utility *in practice* from any community like this one, can and should come from sharing lifehacks. If you go around teaching people methods that they can allegedly use to distinguish *good* strange ideas from *bad* strange ideas, *and* there's some combination of successfully teaching Cognitive Art: Resist Conformity with the less lofty enhancer We Now Have Enough People Physically Present That You Don't Feel Nonconformist, that community will inevitably propagate what they *believe* to be good new ideas that haven't been mass-adopted by the general population.

When I saw that Patri Friedman was wearing Vibrams (five-toed shoes) and that William Eden (then Will Ryan) was also wearing Vibrams, I got a pair myself to see if they'd work. They didn't work for me, which thanks to Cognitive Art: Say Oops I was able to admit without much fuss; and so I put my athletic shoes back on again. Paleo-inspired diets haven't done anything discernible for me, but have helped many other people in the community. Supplementing potassium (citrate) hasn't helped me much, but works dramatically for Anna, Kevin, and Vassar. Seth Roberts's "Shangri-La diet", which was propagating through econblogs, led me to lose twenty pounds that I've mostly kept off, and then it mysteriously stopped working...

De facto, I *have* gotten a noticeable amount of mileage out of imitating things I've seen other rationalists do. In principle, this will work better than reading a lifehacking blog to whatever extent rationalist opinion leaders are better able to filter lifehacks - discern better and worse experimental evidence, avoid affective death spirals around things that sound cool, and give up faster when things don't work. In practice, I myself haven't gone particularly far into the mainstream lifehacking community, so I don't know how much of an advantage, if any, we've got (so far). My suspicion is that on average lifehackers should know more cool things than we do (by virtue of having invested more time and practice), and have more obviously bad things mixed in (due to only average levels of Cognitive Art: Resist Nonsense).

But strange-to-the-mainstream yet oddly-effective ideas propagating through the community is something that happens if everything goes *right*. The danger of these things looking *weird...* is one that I think we just have to bite the bullet on, though opinions on this subject vary between myself and other community leaders.

So a lot of real-world mileage in practice is likely to come out of us imitating each other...

And yet *nonetheless,* I think it worth naming and resisting that dark temptation to think that somebody can't be a *real* community member if they aren't eating beef livers and supplementing potassium, or if they believe in a collapse interpretation of QM, etcetera. If a newcomer *also* doesn't show any particular, noticeable interest in the algorithms and the process, then sure, don't feed the trolls. It should be another matter if someone seems interested in the process, better yet the math, and has some non-zero grasp of it, and are just coming to different conclusions than the local consensus.

Applied rationality counts for something, indeed; rationality that isn't applied might as well not exist. And if somebody believes in something really wacky, like Mormonism or that personal identity follows individual particles, you'd *expect* to eventually find some flaw in reasoning - a departure from the rules - if you trace back their reasoning far enough. But there's a genuine and open question as to how much you should really assume - how much would be *actually true* to assume - about the general reasoning deficits of somebody who says they're Mormon, but who can solve Bayesian problems on a blackboard and explain what Governor Earl Warren was doing wrong and analyzes the Amanda Knox case correctly. Robert Aumann (Nobel laureate Bayesian guy) is a believing Orthodox Jew, after all.

But the deeper danger isn't that of mistakenly excluding someone who's fairly good at a bunch of cognitive algorithms and still has some blind spots.

The deeper danger is in allowing your *de facto* sense of rationalist community to start being defined by conformity to what people think is merely *optimal,* rather than the cognitive algorithms and thinking techniques that are supposed to be at the center.

And then a purely metaphorical Ayn Rand starts kicking people out because they like suboptimal music. A sense of you-must-do-X-to-belong is also a kind of Authority.

Not all Authority is bad - probability theory is also a kind of Authority and I try to be ruled by it as much as I can manage. But good Authority should generally be *modular;* having a sweeping cultural sense of lots and lots of mandatory things is also a failure mode. This is what I think of as the core Objectivist Failure Mode - why the heck is Ayn Rand talking about music?

So let's all please be conservative about invoking the word 'rational', and try not to use it except when we're talking about cognitive algorithms and thinking techniques. And in general and as a reminder, let's continue exerting some pressure to adjust our intuitions about belonging-to-LW-ness in the direction of (a) deliberately not rejecting people who disagree with a particular point of mere optimality, and (b) deliberately extending hands to people who show *respect for the process* and *interest in the algorithms* even if they're disagreeing with the general consensus.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "The Fabric of Real Things"

Previous post: "Rationality: Appreciating Cognitive Algorithms"

The Fabric of Real Things

Followup to: The Useful Concept of Truth

We previously asked:

What rule would restrict our beliefs to just statements that can be meaningful, without excluding a priori anything that could in principle be true?

It doesn't work to require that the belief's truth or falsity make a sensory difference. It's true, but not testable, to say that a spaceship going over the cosmological horizon of an expanding universe does not suddenly blink out of existence. It's *meaningful and false*, rather than *meaningless*, to say that on March 22nd, 2003, the particles in the center of the Sun spontaneously arranged themselves into a short-lived chocolate cake. This statement's truth or falsity has no consequences we'll ever be able to test experientally. Nonetheless, it legitimately describes a way reality could be, but isn't; the atoms in our universe *could've* been arranged like that on March 22nd 2003, but they weren't.

You can't say that there has to be some way to arrange the atoms in the universe so as to make the claim true or alternatively false. Then the theory of quantum mechanics is *a priori* meaningless, because there's no way to arrange *atoms* to make it true. And if you try to substitute quantum fields instead, well, what if they discover something else tomorrow? And is it *meaningless* -rather than *meaningful and false* - to imagine that physicists are lying about quantum mechanics in a grand organized conspiracy?

Since claims are rendered true or false by how-the-universe-is, the question "What claims can be meaningful?" implies the question "What sort of reality can exist for our statements to correspond to?"

If you rephrase it this way, the question probably sounds completely fruitless and pointless, the sort of thing that a philosopher would ponder for years before producing a long, incomprehensible book that would be studied by future generations of unhappy students while being of no conceivable interest to anyone with a real job.

But while deep philosophical dilemmas such as these are never settled by philosophers, they *are* sometimes settled by people working on a related practical problem which happens to intersect the dilemma. There are a lot of people who think I'm being too harsh on philosophers when I express skepticism about mainstream philosophy; but in this case, at least, history clearly bears out the point. Philosophers have been discussing the nature of reality for literal millennia... and yet the people who first delineated and formalized a critical hint about the nature of reality, the people who first discovered *what sort of things seem to be real,*were trying to solve a completely different-sounding question.

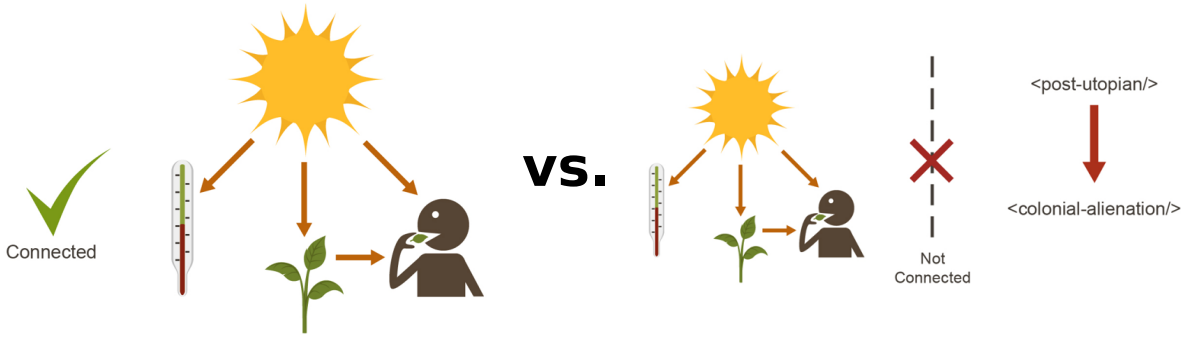
They were trying to figure out whether you can tell the direction of *cause and effect* from survey data.

---

*Please now read **Causal Diagrams and Causal Models**, which was modularized out so that it could act as a standalone introduction. This post involves some simple math, but causality is so basic to key future posts that it's pretty important to get at least some grasp on the math involved. Once you are finished reading, continue with the rest of this post.*

Okay, now suppose someone were to claim the following:

"A universe is a connected fabric of causes and effects."



(In the right-hand image we see a connected causal fabric; the sun raises the temperature, makes plants grow, and sends light into the eyes of the person eating from the plant. On the other hand, while "post-utopian" is linked to "colonial alienation" and vice versa, these two elements don't connect to the rest of the causal fabric - so that must not be a universe.)

This same someone might further claim:

"For a statement to be comparable to your universe, so that it can be true or alternatively false, it must talk about stuff you can *find in relation to yourself* by tracing out causal links."

To clarify the second claim, the idea here is that reference can trace causal links *forwards or backwards*. If a spaceship goes over the cosmological horizon, it may not cause anything *else* to happen to you after that. But you could still say, 'I saw the space shipyard - it affected my eyes - and the shipyard building was the cause of that ship existing and going over the horizon.' You know the second causal link exists, because you've previously observed the general law implementing links of that type - previously observed that objects continue to exist and do not violate Conservation of Energy by spontaneously vanishing.

And now I present three meditations, whose answers (or at least, what I think are the answers) will appear at later points in Highly Advanced Epistemology 101 For Beginners. Please take a shot at whispering the answers to yourself; or if you're bold enough to go on record, comments for collecting posted answers are linked.

Meditation 1:

"You say that a *universe* is a connected fabric of causes and effects. Well, that's a very Western viewpoint - that it's all about mechanistic, deterministic stuff. I agree that anything else is outside the realm of science, but it can still be *real,* you know. My cousin is psychic - if you draw a card from his deck of cards, he can tell you the name of your card before he looks at it. There's no *mechanism* for it - it's not a *causal* thing that scientists could study - he just *does* it. Same thing when I commune on a deep level with the entire universe in order to realize that my partner truly loves me. I agree that purely spiritual phenomena are outside the realm of causal processes, which can be scientifically understood, but I don't agree that they can't be *real.*"

How would you reply?

Meditation 2:

"Does your rule there forbid epiphenomenalist theories of consciousness - that consciousness is caused by neurons, but doesn't affect those neurons in turn? The classic argument for epiphenomenal consciousness has always been that we can imagine a universe in which all the atoms are in the same place and people behave exactly the same way, but there's nobody home - no awareness, no consciousness, inside the brain. The usual *effect* of the brain generating consciousness is missing, but consciousness doesn't cause anything else in turn - it's just a passive awareness - and so from the outside the universe looks the same. Now, I'm not so much interested in whether you think epiphenomenal theories of consciousness are true or false - rather, I want to know if you think they're impossible or meaningless *a priori* based on your rules."

How would you reply?

Meditation 3:

Does the idea that everything is made of causes and effects meaningfully constrain experience? Can you coherently say how reality might look, if our universe did *not* have the kind of structure that appears in a causal model?

Mainstream status.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Causal Diagrams and Causal Models"

Previous post: "Firewalling the Optimal from the Rational"

Causal Diagrams and Causal Models

Suppose a general-population survey shows that people who exercise less, weigh more. You don't have any known direction of *time* in the data - you don't know which came first, the increased weight or the diminished exercise. And you didn't randomly assign half the population to exercise less; you just surveyed an existing population.

The statisticians who discovered causality were trying to find a way to distinguish, within survey data, the direction of cause and effect - whether, as common sense would have it, more obese people exercise less *because* they find physical activity less rewarding; or whether, as in the virtue theory of metabolism, lack of exercise actually *causes* weight gain due to divine punishment for the sin of sloth.

vs.

The usual way to resolve this sort of question is by *randomized intervention.* If you randomly assign half your experimental subjects to exercise more, and afterward the increased-exercise group doesn't lose any weight compared to the control group [1], you could rule out causality *from* exercise *to* weight, and conclude that the correlation between weight and exercise is probably due to physical activity being less fun when you're overweight [3]. The question is whether you can get causal data *without* interventions.

For a long time, the conventional wisdom in philosophy was that this was impossible unless you knew the direction of time and knew which event had happened first. Among some philosophers of science, there was a belief that the "direction of causality" was a *meaningless* question, and that in the universe itself there were *only* correlations - that "cause and effect" was something unobservable and undefinable, that only unsophisticated non-statisticians believed in due to their lack of formal training:

> "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." -- Bertrand Russell (he later changed his mind)
>
> "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish among the inscrutable arcana of modern science, namely, the category of cause and effect." -- Karl Pearson

The famous statistician Fisher, who was also a smoker, testified before Congress that the correlation between smoking and lung cancer couldn't prove that the former caused the latter. We have remnants of this type of reasoning in old-school "Correlation does not imply causation", without the now-standard appendix, "But it sure is a hint".

This skepticism was overturned by a surprisingly simple mathematical observation.

Let's say there are three variables in the survey data: Weight, how much the person exercises, and how much time they spend on the Internet.

For simplicity, we'll have these three variables be binary, yes-or-no observations: Y or N for whether the person has a BMI over 25, Y or N for whether they exercised at least twice in the last week, and Y or N for whether they've checked Reddit in the last 72 hours.

Now let's say our gathered data looks like this:

Overweight	Exercise	Internet	#
Y	Y	Y	1,119
Y	Y	N	16,104
Y	N	Y	11,121
Y	N	N	60,032
N	Y	Y	18,102
N	Y	N	132,111

N	N	Y	29,120
N	N	N	155,033

And lo, merely by eyeballing this data -

(which is *totally made up*, so don't go actually *believing* the conclusion I'm about to draw)

- we now realize that *being overweight and spending time on the Internet both cause you to exercise less,* presumably because exercise is less fun and you have more alternative things to do, *but exercising has no causal influence on body weight or Internet use.*

"What!" you cry. "How can you tell *that* just by inspecting those numbers? You can't say that exercise isn't *correlated* to body weight - if you just look at all the members of the population who exercise, they clearly have lower weights. 10% of exercisers are overweight, vs. 28% of non-exercisers. How could you rule out the obvious causal explanation for that correlation, just by looking at this data?"

There's a wee bit of math involved. It's *simple* math - the part we'll use doesn't involve solving equations or complicated proofs -but we do have to introduce a wee bit of novel math to explain how the heck we got there from here.

Let me start with a question that turned out - to the surprise of many investigators involved - to be highly related to the issue we've just addressed.

Suppose that earthquakes and burglars can both set off burglar alarms. If the burglar alarm in your house goes off, it might be because of an actual burglar, but it might *also* be because a minor earthquake rocked your house and triggered a few sensors. Early investigators in Artificial Intelligence, who were trying to represent all high-level events using primitive tokens in a first-order logic (for reasons of historical stupidity we won't go into) were stymied by the following apparent paradox:

- If you tell me that my burglar alarm went off, I infer a burglar, which I will represent in my first-order-logical database using a theorem ⊢ ALARM → BURGLAR. (The symbol "⊢" is called "turnstile" and means "the logical system asserts that".)
- If an earthquake occurs, it will set off burglar alarms. I shall represent this using the theorem ⊢ EARTHQUAKE → ALARM, or "earthquake implies alarm".
- If you tell me that my alarm went off, and then further tell me that an earthquake occurred, it *explains away* my burglar alarm going off. I don't need to explain the alarm by a burglar, because the alarm has already been explained by the earthquake. I conclude there was no burglar. I shall represent this by adding a theorem which says ⊢ (EARTHQUAKE & ALARM) → NOT BURGLAR.

Which represents a logical contradiction, and for a while there were attempts to develop "non-monotonic logics" so that you could retract conclusions given additional data. This didn't work very well, since the underlying structure of reasoning was a terrible fit for the structure of classical logic, even when mutated.

Just changing certainties to quantitative probabilities can fix many problems with classical logic, and one might think that this case was likewise easily fixed.

Namely, just write a probability table of all possible combinations of earthquake or ¬earthquake, burglar or ¬burglar, and alarm or ¬alarm (where ¬ is the logical negation symbol), with the following entries:

Burglar	**Earthquake**	**Alarm**	**%**
b	e	a	.000162
b	e	¬a	.0000085
b	¬e	a	.0151
b	¬e	¬a	.00168
¬b	e	a	.0078
¬b	e	¬a	.002
¬b	¬e	a	.00097
¬b	¬e	¬a	.972

Using the operations of *marginalization* and *conditionalization,* we get the desired reasoning back out:

Let's start with the *probability of a burglar given an alarm,* p(burglar|alarm). By the law of conditional probability,

$$p(b|a) = \frac{p(ab)}{p(a)}$$

i.e. the relative fraction of cases where there's an alarm *and* a burglar, within the set of all cases where there's an alarm.

The table doesn't directly tell us p(alarm & burglar)/p(alarm), but by the law of marginal probability,

$$p(ab) = p(abe) + p(ab\neg e) = .000162 + .0151 = .0153$$

Similarly, to get the probability of an alarm going off, p(alarm), we add up all the different sets of events that involve an alarm going off - entries 1, 3, 5, and 7 in the table.

So the entire set of calculations looks like this:

- If I hear a burglar alarm, I conclude there was probably (63%) a burglar.

$$p(b|a) = \frac{p(ab)}{p(a)} = \frac{.0153}{.000162 + .0151 + .0078 + .00097} = .63$$

- If I learn about an earthquake, I conclude there was probably (80%) an alarm.

$$p(a|e) = \frac{p(ae)}{p(e)} = \frac{.000162 + .0078}{.000162 + .0078 + .0000085 + .002} = .8$$

- I hear about an alarm and then hear about an earthquake; I conclude there was probably (98%) no burglar.

$$\frac{p(ae\neg b)}{p(ae)} = \frac{p(ae\neg b)}{p(aeb) + p(ae\neg b)} = \frac{.0078}{.000162 + .0078} = .98$$

Thus, a joint probability distribution is indeed capable of *representing* the reasoning-behaviors we want.

So is our problem solved? Our work done?

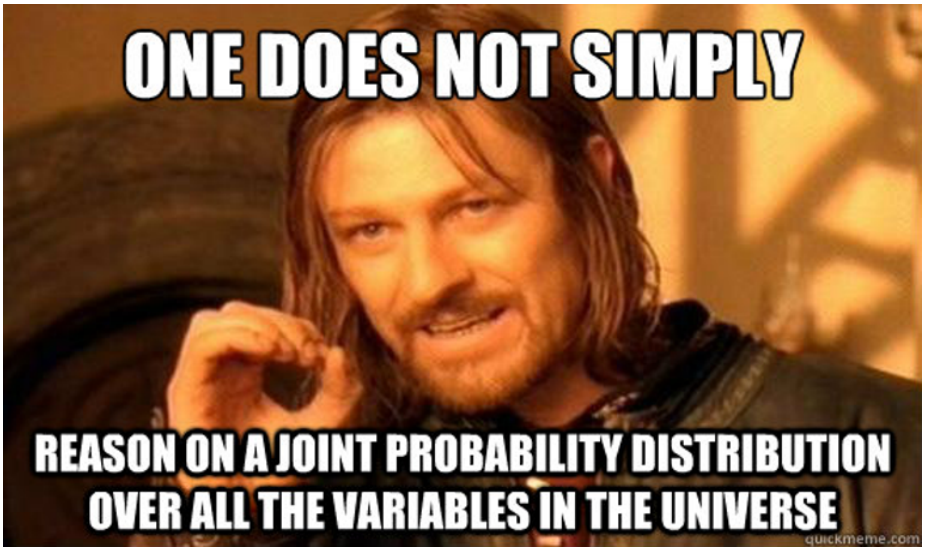
Not in real life or real Artificial Intelligence work. The problem is that this solution doesn't scale. *Boy howdy*, does it not scale! If you have a model containing *forty* binary variables - alert readers may notice that the observed physical universe contains at least forty things - and you try to write out the *joint probability distribution* over all combinations of those variables, it looks like this:

.0000000000112	YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
.000000000000034	YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYN
.00000000000991	YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYNY
.00000000000532	YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYNN
.000000000145	YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYNYY
...	...

(1,099,511,627,776 entries)

This isn't merely a storage problem. In terms of storage, a trillion entries is just a terabyte or three. The real problem is *learning* a table like that. You have to deduce 1,099,511,627,776 floating-point probabilities from observed data, and the only constraint on this giant table is that all the probabilities must sum to exactly 1.0, a problem with 1,099,511,627,775 degrees of freedom. (If you know the first 1,099,511,627,775 numbers, you can deduce the 1,099,511,627,776th number using the constraint that they all sum to exactly 1.0.) It's not the storage cost that kills you in a problem with forty variables,

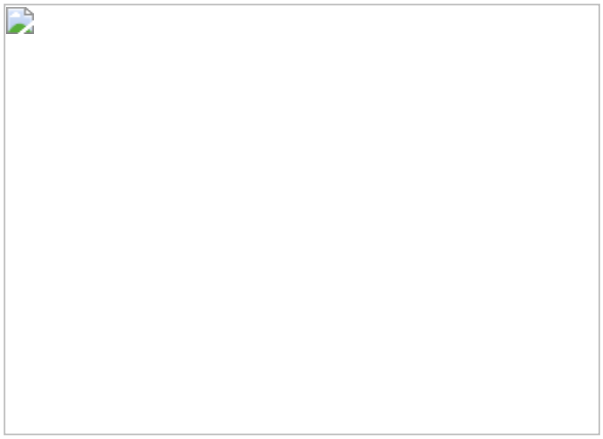
it's the difficulty of gathering enough observational data to constrain a trillion different parameters. And in a universe containing *seventy* things, things are even worse.



So instead, suppose we approached the earthquake-burglar problem by trying to specify probabilities in a format where... never mind, it's easier to just give an example before stating abstract rules.

First let's add, for purposes of further illustration, a new variable, "Recession", whether or not there's a depressed economy at the time. Now suppose that:

- The probability of an earthquake is 0.01.
- The probability of a recession at any given time is 0.33 (or 1/3).
- The probability of a burglary given a recession is 0.04; or, given no recession, 0.01.
- An earthquake is 0.8 likely to set off your burglar alarm; a burglar is 0.9 likely to set off your burglar alarm. *And* - we can't compute this model fully without this info - the combination of a burglar *and* an earthquake is 0.95 likely to set off the alarm; and in the absence of either burglars or earthquakes, your alarm has a 0.001 chance of going off anyway.



p(r)	.33
p(¬r)	.67

p(e)	.01
p(¬e)	.99

p(b|r)	.04
p(b|¬r)	.01
p(¬b|r)	.96
p(¬b|¬r)	.99

p(a|be)	.95
p(a|b¬e)	.9
p(a|¬be)	.797
p(a|¬b¬e)	.001
p(¬a|be)	.05
p(¬a|b¬e)	.1
p(¬a|¬be)	.203
p(¬a|¬b¬e)	.999

According to this model, if you want to know "The probability that an earthquake occurs" - just the probability of that one variable, without talking about any others - you can directly look up p(e) = .01. On the other hand, if you want to know the probability of a burglar striking, you have to first look up the probability of a recession (.33), and then p(b|r) and p(b|¬r), and sum up p(b|r)*p(r) + p(b|¬r)*p(¬r) to get a net probability of .01*.66 + .04*.33 = .02 = p(b), a 2% probability that a burglar is around at some random time.

If we want to compute the joint probability of four values for all four variables - for example, the probability that there is no earthquake *and* no recession *and* a burglar *and* the alarm goes off - this causal model computes this joint probability as the product:

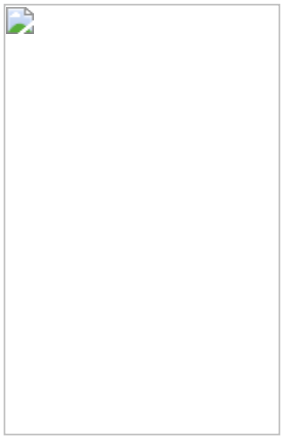
$$p(\neg e)p(\neg r)p(b|\neg r)p(a|b\neg e) = .99 * .67 * .01 * .9 = .006 = 0.6\%$$

In general, to go from a *causal model* to a *probability distribution,* we compute, for each setting of all the variables, the product

$$p(\mathbf{X} = \mathbf{x}) = \prod_i p(X_i = x_i | \mathbf{PA_i} = \mathbf{pa_i})$$

multiplying together the conditional probability of each variable *given the values of its immediate parents.* (If a node has no parents, the probability table for it has just an unconditional probability, like "the chance of an earthquake is .01".)

This is a *causal* model because it corresponds to a world in which each event is *directly* caused by only a small set of other events, its parent nodes in the graph. In this model, a recession can *indirectly* cause an alarm to go off - the recession increases the probability of a burglar, who in turn sets off an alarm - but the recession *only* acts on the alarm through the *intermediate cause* of the burglar. (Contrast to a model where recessions set off burglar alarms directly.)

 **vs.** 

The first diagram implies that once we *already know* whether or not there's a burglar, we don't learn *anything more* about the probability of a burglar alarm, if we find out that there's a recession:

$$p(a|b) = p(a|br)$$

This is a fundamental illustration of *the locality of causality* - once I know there's a burglar, I know *everything I need to know* to calculate the probability that there's an alarm. Knowing the state of Burglar *screens off* anything that Recession could tell me about Alarm - even though, if I *didn't* know the value of the Burglar variable, Recessions would appear to be statistically correlated with Alarms. The present screens off the past from the future; in a causal system, if you know the *exact, complete* state of the present, the state of the past has no further physical relevance to computing the future. It's how, in a system containing many correlations (like the recession-alarm correlation), it's still possible to compute each variable just by looking at a small number of immediate neighbors.

Constraints like this are also how we can store a causal model - and much more importantly, *learn* a causal model - with many fewer parameters than the naked, raw, joint probability distribution.

Let's illustrate this using a simplified version of this graph, which only talks about earthquakes and recessions. We could consider three hypothetical causal diagrams over only these two variables:

p(r)	0.03
p(¬r)	0.97

p(e)	0.29
p(¬e)	0.71

p(E&R)=p(E)p(R)

p(e)	0.29
p(¬e)	0.71

p(r|e)	0.15
p(¬r|e)	0.85
p(r|¬e)	0.03
p(¬r|¬e)	0.97

p(E&R) = p(E)p(R|E)

p(r)	0.03
p(¬r)	0.97

p(e|r)	0.24
p(¬e|r)	0.76
p(e|¬r)	0.09
p(¬e|¬r)	0.91

p(E&R) = p(R)p(E|R)

Let's consider the first hypothesis - that there's no causal arrows connecting earthquakes and recessions. If we build a *causal model* around this diagram, it has 2 real degrees of freedom - a degree of freedom for saying that the probability of an earthquake is, say, 29% (and hence that the probability of not-earthquake is necessarily 71%), and another degree of freedom for saying that the probability of a recession is 3% (and hence the probability of not-recession is constrained to be 97%).

On the other hand, the full joint probability distribution would have 3 degrees of freedom - a free choice of (earthquake&recession), a choice of p(earthquake&¬recession), a choice of p(¬earthquake&recession), and then a constrained p(¬earthquake&¬recession) which must be equal to 1 minus the sum of the other three, so that all four probabilities sum to 1.0.

By the pigeonhole principle (you can't fit 3 pigeons into 2 pigeonholes) there must be some joint probability distributions which *cannot be represented* in the first causal structure. This means the first causal structure is *falsifiable;* there's survey data we can get which would lead us to reject it as a hypothesis. In particular, the first causal model requires:

$$p(er) = p(e)p(r)$$

or equivalently

$$p(e|r) = p(e)$$

or equivalently

$$p(r|e) = p(r)$$

which is a *conditional independence* constraint - it says that learning about recessions doesn't tell us anything about the probability of an earthquake or vice versa. If we find that earthquakes and recessions are highly correlated in the observed data - if earthquakes and recessions go together, or earthquakes and the *absence* of recessions go together - it falsifies the first causal model.

For example, let's say that in your state, an earthquake is 0.1 probable per year and a recession is 0.2 probable. If we suppose that earthquakes don't cause recessions, earthquakes are not an effect of recessions, and that there aren't hidden aliens which produce both earthquakes and recessions, then we should find that years in which there are *earthquakes and recessions* happen around 0.02 of the time. If instead earthquakes and recessions happen 0.08 of the time, then the probability of a recession *given* an earthquake is 0.8 instead of 0.2, and we should much more strongly expect a recession any time we are told that an earthquake has occurred. Given enough samples, this falsifies the theory that these factors are unconnected; or rather, the more samples we have, the more we disbelieve that the two events are unconnected.

On the other hand, we can't tell apart the second two possibilities from survey data, because both causal models have 3 degrees of freedom, which is the size of the full joint probability distribution. (In general, *fully connected* causal graphs in which there's a line between every pair of nodes, have the same number of degrees of freedom as a raw joint distribution - and 2 nodes connected by 1 line are "fully connected".) We can't tell if earthquakes are 0.1 likely and cause recessions with 0.8 probability, or recessions are 0.2 likely and cause earthquakes with 0.4 probability (or if there are hidden aliens which on 6% of years show up and cause earthquakes and recessions with probability 1).

With larger universes, the difference between *causal models* and *joint probability distributions* becomes a lot more striking. If we're trying to reason about a million binary variables connected in a huge causal model, and each variable could have up to four direct 'parents' - four other variables that *directly* exert a causal effect on it - then the total number of free parameters would be at most... 16 million!

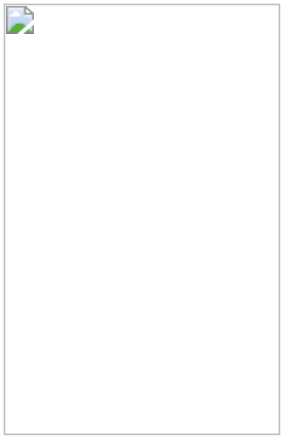
The number of free parameters in a raw joint probability distribution over a million binary variables would be $2^{1,000,000}$. Minus one.

So causal models which are *less* than fully connected - in which most objects in the universe are not the direct cause or direct effect of everything else in the universe - are very strongly *falsifiable;* they only allow probability distributions (hence, observed frequencies) in an infinitesimally tiny range of all possible joint probability tables. Causal models very strongly constrain anticipation - disallow almost all possible patterns of observed frequencies - and gain mighty Bayesian advantages when these predictions come true.

To see this effect at work, let's consider the *three* variables Recession, Burglar, and Alarm.

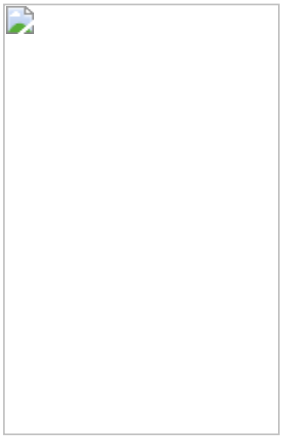
Alarm	Burglar	Recession	%
Y	Y	Y	.012
N	Y	Y	.0013
Y	N	Y	.00287
N	N	Y	.317
Y	Y	N	.003
N	Y	N	.000333
Y	N	N	.00591
N	N	N	.654

All three variables seem correlated to each other when considered two at a time. For example, if we consider Recessions and Alarms, they should seem correlated because recessions cause burglars which cause alarms. If we learn there was an alarm, for example, we conclude it's more probable that there was a recession. So since all three variables are correlated, can we distinguish between, say, these three causal models?



$$p(a|e) = \frac{p(ae)}{p(e)} = \frac{.000162 + .0078}{.000162 + .0078 + .0000085 + .002} = .8$$

$$p(b|a) = \frac{p(ab)}{p(a)} = \frac{.0153}{.000162 + .0151 + .0078 + .00097} = .63$$



$$p(rab) = p(r)p(a|r)p(b|a)$$

Yes we can! Among these causal models, the prediction which only the first model makes, which is not shared by either of the other two, is that *once we know whether a burglar is there,* we learn nothing *more* about whether there was an alarm by finding out that there was a recession, since recessions only affect alarms through the intermediary of burglars:

$$p(a|b) = p(a|br)$$

But the third model, in which recessions directly cause alarms, which only then cause burglars, does *not* have this property. If I know that a burglar has appeared, it's likely that an alarm caused the burglar - but it's even *more* likely that there was an alarm, if there was a recession around to cause the alarm! So the third model predicts:

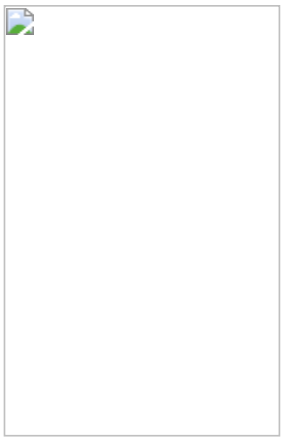
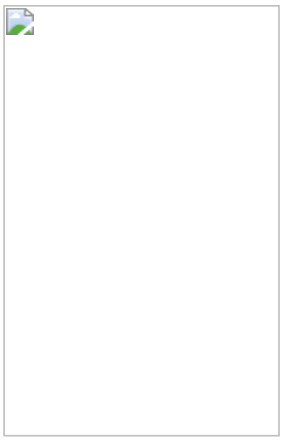
$$p(a|b) = p(a|br)$$

And in the second model, where alarms and recessions both cause burglars, we again don't have the conditional independence. If we know that there's a burglar, then we think that either an alarm or a recession caused it; and if we're told that there's an alarm, we'd conclude it was less likely that there was a recession, since the recession had been explained away.

(This may seem a bit clearer by considering the scenario B->A<-E, where burglars and earthquakes both cause alarms. If we're told the value of the bottom node, that there was an alarm, the probability of there being a burglar is *not* independent of whether we're told there was an earthquake - the two top nodes are *not* conditionally independent *once we condition on the bottom node.*)

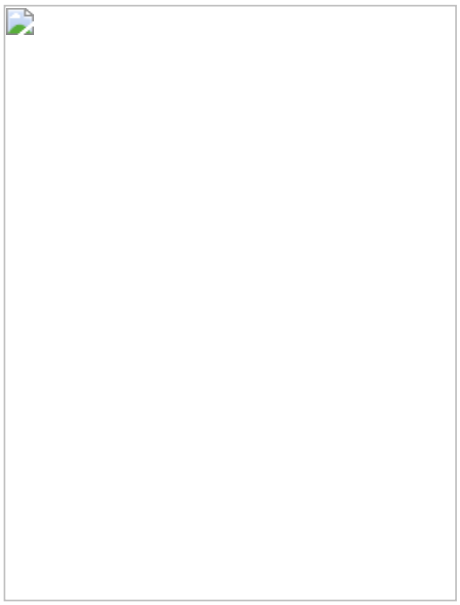
On the other hand, we can't tell the difference between:

vs. **vs.**

using *only* this data and no other variables, because all three causal structures predict the same pattern of conditional dependence and independence - three variables which all appear mutually correlated, but Alarm and Recession become independent once you condition on Burglar.

Being able to read off patterns of conditional dependence and independence is an art known as "D-separation", and if you're good at it you can glance at a diagram like this...



...and see that, once we already know the Season, whether the Sprinkler is on and whether it is Raining are conditionally independent of each other - if we're told that it's Raining we conclude nothing about whether or not the Sprinkler is on. But if we then further observe that the sidewalk is Slippery, then Sprinkler and Rain become conditionally dependent once more, because if the Sidewalk is Slippery then it is probably Wet and this can be explained by either the Sprinkler or the Rain but probably not both, i.e. if we're told that it's Raining we conclude that it's less likely that the Sprinkler was on.

Okay, back to the obesity-exercise-Internet example. You may recall that we had the following observed frequencies:

Overweight	**Exercise**	**Internet**	**#**
Y	Y	Y	1,119
Y	Y	N	16,104
Y	N	Y	11,121

Y	N	N	60,032
N	Y	Y	18,102
N	Y	N	132,111
N	N	Y	29,120
N	N	N	155,033

Do you see where this is going?

"Er," you reply, "Maybe if I had a calculator and ten minutes... you want to just go ahead and spell it out?"

Sure! First, we *marginalize* over the 'exercise' variable to get the table for just weight and Internet use. We do this by taking the 1,119 people who are YYY, overweight and Reddit users and exercising, and the 11,121 people who are overweight and non-exercising and Reddit users, YNY, and adding them together to get 12,240 total people who are overweight Reddit users:

Overweight	**Internet**	**#**
Y	Y	12,240
Y	N	76,136
N	Y	47,222
N	N	287,144

"And then?"

Well, that suggests that the *probability* of using Reddit, given that your weight is normal, is the *same* as the probability that you use Reddit, given that you're overweight. 47,222 out of 334,366 normal-weight people use Reddit, and 12,240 out of 88,376 overweight people use Reddit. That's about 14% either way.

"And so we conclude?"

Well, first we conclude it's not particularly likely that using Reddit causes weight gain, or that being overweight causes people to use Reddit:

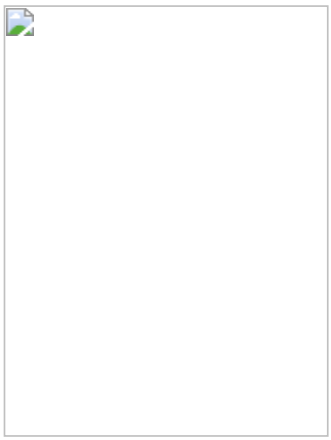
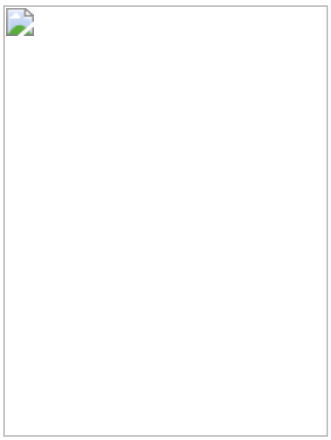
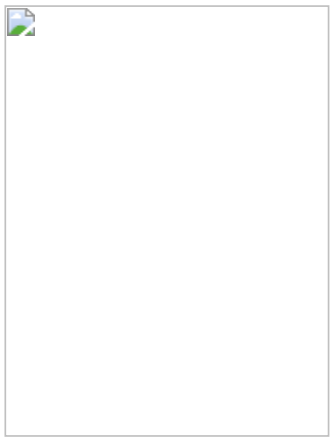
If either of those causal links existed, those two variables should be *correlated.* We shouldn't find the *lack of correlation* or *conditional independence* that we just discovered.

Next, imagine that the real causal graph looked like this:

In this graph, exercising *causes* you to be less likely to be overweight (due to the virtue theory of metabolism), and exercising *causes* you to spend less time on the Internet (because you have less time for it).

But in this case we should *not* see that the groups who are/aren't overweight have the same probability of spending time on Reddit. There should be an outsized group of people who are both normal-weight and non-Redditors (because they exercise), and an outsized group of non-exercisers who are overweight and Reddit-using.

So that causal graph is also *ruled out* by the data, as are others like:

Leaving *only* this causal graph:

Which says that weight and Internet use exert causal effects on exercise, but exercise doesn't causally affect either.

All this discipline was invented and systematized by Judea Pearl, Peter Spirtes, Thomas Verma, and a number of other people in the 1980s and you should be quite impressed by their accomplishment, because before then, inferring causality from correlation was thought to be a fundamentally unsolvable problem. The standard volume on causal structure is *Causality* by Judea Pearl.

Causal *models* (with specific probabilities attached) are sometimes known as "Bayesian networks" or "Bayes nets", since they were invented by Bayesians and make use of Bayes's Theorem. They have all sorts of neat computational advantages which are far beyond the scope of this introduction - e.g. in many cases you can split up a Bayesian network into parts, put each of the parts on its own computer processor, and then update on three different pieces of evidence at once using a neatly local message-passing algorithm in which each node talks only to its immediate neighbors and when all the updates are finished propagating the whole network has settled into the correct state. For more on this see Judea Pearl's *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* which is the original book on Bayes nets and still the best introduction I've personally happened to read.

---

[1] Somewhat to my own shame, I must admit to ignoring my own observations in this department - even after I saw no discernible effect on my weight or my musculature from aerobic exercise and strength training 2 hours a day 3 times a week, I didn't really start believing that the virtue theory of metabolism was *wrong* [2] until after other people had started the skeptical dogpile.

[2] I should mention, though, that I have confirmed a personal effect where eating *enough* cookies (at a convention where no protein is available) will cause weight gain afterward. There's no other discernible correlation between my carbs/protein/fat allocations and weight gain, *just* that eating sweets in large quantities can cause weight gain afterward. This admittedly does bear with the straight-out virtue theory of metabolism, i.e., eating pleasurable foods is sinful weakness and hence punished with fat.

[3] Or there might be some hidden third factor, a gene which causes both fat and non-exercise. By Occam's Razor this is more complicated and its probability is penalized accordingly, but we can't actually rule it out. It is obviously impossible to do the converse experiment where half the subjects are randomly assigned lower weights, since there's no known intervention which can cause weight loss.

Mainstream status: This is meant to be an introduction to completely bog-standard Bayesian networks, causal models, and causal diagrams. Any departures from mainstream academic views are errors and should be flagged accordingly.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Stuff That Makes Stuff Happen"

Previous post: "The Fabric of Real Things"

Stuff That Makes Stuff Happen

Followup to: Causality: The Fabric of Real Things

Previous meditation:

"You say that a *universe* is a connected fabric of causes and effects. Well, that's a very Western viewpoint - that it's all about mechanistic, deterministic stuff. I agree that anything else is outside the realm of science, but it can still be *real,* you know. My cousin is psychic - if you draw a card from his deck of cards, he can tell you the name of your card before he looks at it. There's no *mechanism* for it - it's not a *causal* thing that scientists could study - he just *does* it. Same thing when I commune on a deep level with the entire universe in order to realize that my partner truly loves me. I agree that purely spiritual phenomena are outside the realm of causal processes that can be studied by experiments, but I don't agree that they can't be *real.*"

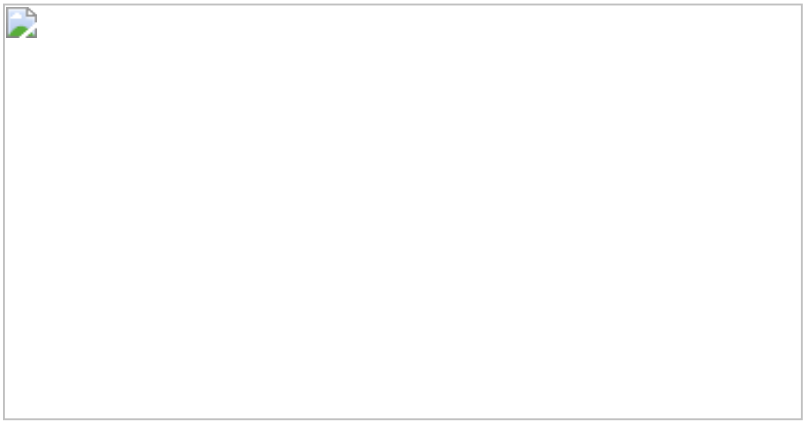
Reply:

Fundamentally, a causal model is a way of *factorizing our uncertainty* about the universe. One way of viewing a causal model is as a structure of *deterministic* functions plus *uncorrelated* sources of background uncertainty.

Let's use the Obesity-Exercise-Internet model *(reminder: which is totally made up)* as an example again:

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$$

We can also view this as a set of *deterministic* functions F_i, plus *uncorrelated* background sources of uncertainty U_i:



This says is that the value x_3 - how much someone exercises - is a function of how obese they are (x_1), how much time they spend on the Internet (x_2), *plus* some other background factors U_3 *which don't correlate to anything else in the diagram,* all of which *collectively* determine, when combined by the mechanism F_3, how much time someone spends exercising.

There might be any number of different real factors involved in the possible states of U_3 - like whether someone has a personal taste for jogging, whether they've ever been to a trampoline park and liked it, whether they have some gene that affects exercise endorphins. These are all different unknown background facts about a person, which might affect whether or not they exercise, above and beyond obesity and Internet use.

But from the perspective of somebody building a causal model, so long as we don't have anything else in our causal graph that *correlates* with these factors, we can sum them up into a single *factor of subjective uncertainty,* our uncertainty U_3 about all the other things that might add up to a force for or against exercising. Once we know that someone isn't overweight and that they spend a lot of time on the Internet, all our uncertainty about those other *background* factors gets summed up with those two *known* factors and turned into a 38% conditional probability that the person exercises frequently.

And the key condition on a causal graph is that if you've properly described your beliefs about the connective mechanisms F_i, all your remaining uncertainty U_i should be *conditionally independent:*

$$p(u_1, u_2, u_3) = p(u_1)p(u_2)p(u_3)$$

or more generally

$$p(\mathbf{U}) = \prod p(U_i)$$

And then plugging those probable U_i into the strictly deterministic F_i should give us back out our whole causal model - the same joint probability table over the observable X_i.

Hence the idea that a causal model *factorizes* uncertainty. It factorizes out all the mechanisms that we *believe* connect variables, and all remaining uncertainty should be uncorrelated *so far as we know*.

To put it another way, if we ourselves knew about a correlation between two U_i that *wasn't* in the causal model, our own expectations for the joint probability table couldn't match the model's product

$$p(\mathbf{x}) = \prod p(x_i|\mathbf{pa_i})$$

and all the theorems about causal inference would go out the window. Technically, the idea that the U_i are uncorrelated is known as the causal Markov condition.

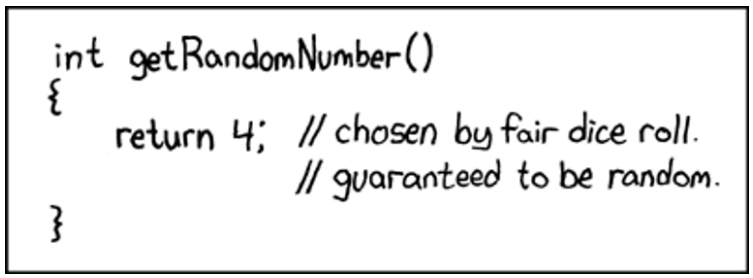
What if you realize that two variables actually *are* correlated more than you thought? What if, to make the diagram correspond to reality, you'd have to hack it to make some U_a and U_b correlated?

Then you draw another arrow from X_a to X_b, or from X_b to X_a; or you make a new node representing the correlated part of U_a and U_b, X_c, and draw arrows from X_c to X_a and X_b.

 vs.  vs. 

(Or you might have to draw some extra causal arrows somewhere else; but those three changes are the ones that would solve the problem most directly.)

There was apparently at one point - I'm not sure if it's still going on or not - this big debate about the true meaning of *randomization* in experiments, and what counts as 'truly random'. Is your randomized experiment invalidated, if you use a merely pseudo-random algorithm instead of a thermal noise generator? Is it okay to use pseudo-random algorithms? Is it okay to use shoddy pseudo-randomness that a professional cryptographer would sneer at? Clearly, using 1-0-1-0-1-0 on a list of patients in alphabetical order isn't random enough... or is it? What if you pair off patients in alphabetical order, and flip a coin to assign one member of each pair to the experimental group and the control? How random is random?



Understanding that causal models *factorize uncertainty* leads to the realization that "randomizing" an experimental variable means using randomness, a U_x for the assignment, which doesn't correlate with your uncertainty about any other U_i. Our *uncertainty* about a thermal noise generator seems strongly guaranteed to be uncorrelated with our *uncertainty* about a subject's economic status, their upbringing, or anything else in the universe that might affect how they react to Drug A...

...unless somebody wrote *down* the output of the thermal noise generator, and then used it in *another* experiment on the *same* group of subjects to test Drug B. It doesn't matter how "intrinsically random" that output was - whether it was the XOR of a thermal noise source, a quantum noise source, a human being's so-called free will, and the world's strongest cryptographic algorithm - once it ends up *correlated* to any other uncertain background factor, any other U_i, you've invalidated the randomization. That's the implicit problem in the XKCD cartoon above.

But picking a strong randomness source, and using the output only *once*, is a pretty solid guarantee this won't happen.

Unless, ya know, you start out with a list of subjects sorted by income, and the randomness source randomly happens to put out 111111000000. Whereupon, as soon as you *look* at the output and are *no longer* uncertain about it, you might expect correlation and trouble. But that's a different and much thornier issue in Bayesianism vs. frequentism.

If we take frequentist ideas about randomization at face value, then the key requirement for theorems about experimental randomization to be applicable, is for your uncertainty about patient randomization to *not correlate* with any other background facts about the patients. A double-blinded study (where the doctors don't know patient status) ensures that patient status doesn't correlate with the doctor's *beliefs* about a patient leading them to treat patients differently. Even plugging in the fixed string "1010101010" would be sufficiently random *if* that pattern wasn't correlated to anything important; the trouble is that such a simple pattern could very easily correlate with some background effect, and we can believe in this possible correlation even if we're not sure what the exact correlation would be.

(It's worth noting that the Center for Applied Rationality ran the June minicamp experiment using a standard but unusual statistical method of sorting applicants into pairs that seemed of roughly matched prior ability / prior expected outcome, and then flipping a coin to pick one member of each pair to be admitted or not admitted that year. This procedure means you never randomly improbably get an experimental group that would, once you actually looked at the random numbers, seem much more promising or much worse than the control group in advance - where the frequentist guarantee that you used an experimental procedure where this usually doesn't happen 'in the long run', might be cold comfort if it obviously *had* happened this time once you looked at the random numbers. Roughly, this choice reflects a difference between frequentist ideas about procedures that make it hard for scientists to obtain results unless their theories are true, and then not caring about the actual random numbers so long as it's still hard to get fake results on average; versus a Bayesian goal of trying to get the maximum evidence out of the update we'll actually have to perform after looking at the results, including how the random numbers turned out on this particular occasion. Note that frequentist ethics are still being obeyed - you can't game the expected statistical significance of experimental vs. control results by picking bad pairs, so long as the coinflips themselves are fair!)

Okay, let's look at that meditation again:

"You say that a *universe* is a connected fabric of causes and effects. Well, that's a very Western viewpoint - that it's all about mechanistic, deterministic stuff. I agree that anything else is outside the realm of science, but it can still be *real,* you know. My cousin is psychic - if you draw a card from his deck of cards, he can tell you the name of your card before he looks at it. There's no *mechanism* for it - it's not a *causal* thing that scientists could study - he just *does* it. Same thing when I commune on a deep level with the entire universe in order to realize that my partner truly loves me. I agree that purely spiritual phenomena are outside the realm of causal processes that can be studied by experiments, but I don't agree that they can't be *real.*"

Well, you know, you can stand there all day, shouting all you like about how something is outside the realm of science, but if a picture of the world has this...

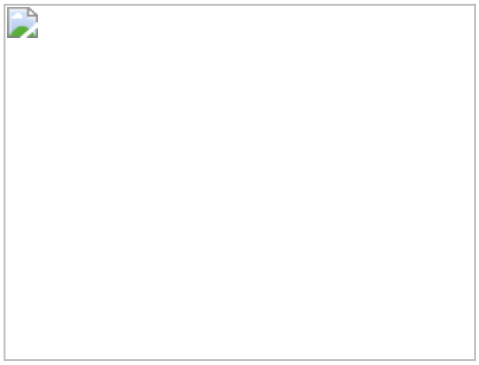
...then we're either going to draw an arrow from the top card to the prediction; an arrow from the prediction to the top card (the prediction makes it happen!); or arrows from a third

source to both of them (aliens are picking the top card and using telepathy on your cousin... or something; there's no rule you have to *label* your nodes).

More generally, for me to expect your beliefs to correlate with reality, I have to either think that reality is the cause of your beliefs, expect your beliefs to alter reality, or believe that some third factor is influencing both of them.

This is the *more general* argument that "To draw an accurate map of a city, you have to open the blinds and look out the window and draw lines on paper corresponding to what you see; sitting in your living-room with the blinds closed, making stuff up, isn't going to work."

Correlation requires causal interaction; and expecting beliefs to be true means expecting the map to *correlate* with the territory. To open your eyes and look at your shoelaces is to let those shoelaces have a causal effect on your brain - in general, looking at something, gaining information about it, requires letting it causally affect you. Learning about X means letting your brain's state be causally determined by X's state. The first thing that happens is that your shoelace is untied; the next thing that happens is that the shoelace interacts with your brain, via light and eyes and the visual cortex, in a way that makes your brain believe your shoelace is untied.



p(Shoelace=tied, Belief="tied")	0.931
p(Shoelace=tied, Belief="untied")	0.003
p(Shoelace=untied, Belief="untied")	0.053
p(Shoelace=untied, Belief="tied")	0.012

This is related in spirit to the idea seen earlier on LW that having knowledge materialize from nowhere *directly* violates the second law of thermodynamics because mutual information counts as thermodynamic negentropy. But the causal form of the proof is much deeper and more general. It applies even in universes like Conway's Game of Life where there's no equivalent of the second law of thermodynamics. It applies even if we're in the Matrix and the aliens can violate physics at will. Even when entropy can go down, you still can't learn about things without being causally connected to them.

The fundamental question of rationality, "What do you think you know and how do you think you know it?", is on its strictest level a request for a causal model of how you think your brain ended up mirroring reality - the causal process which accounts for this supposed correlation.

You might not think that this would be a useful question to ask - that when your brain has an irrational belief, it would automatically have irrational beliefs about process.

But "the human brain is not illogically omniscient", we might say. When our brain undergoes motivated cognition or other fallacies, it often ends up strongly believing in X, *without* the unconscious rationalization process having been sophisticated enough to *also* invent a causal story explaining how we know X. "How could you possibly know that, even if it was true?" is a more skeptical form of the same question. If you can successfully stop your brain from rationalizing-on-the-spot, there actually *is* this useful thing you can sometimes catch yourself in, wherein you go, "Oh, wait, even if I'm in a world where AI does get developed on

March 4th, 2029, there's no lawful story which could account for me knowing that in advance - there must've been some other pressure on my brain to produce that belief."

Since it illustrates an important general point, I shall now take a moment to remark on the idea that science is merely one magisterium, and there's other magisteria which can't be subjected to standards of mere evidence, because they are special. That seeing a ghost, or knowing something because God spoke to you in your heart, is an exception to the ordinary laws of epistemology.

That exception would be convenient for the speaker, perhaps. But causality is *more general* than that; it is *not* excepted by such hypotheses. "I saw a ghost", "I mysteriously sensed a ghost", "God spoke to me in my heart" - there's no difficulty drawing those causal diagrams.

The *methods* of science - even sophisticated methods like the conditions for randomizing a trial - aren't just about atoms, or quantum fields.

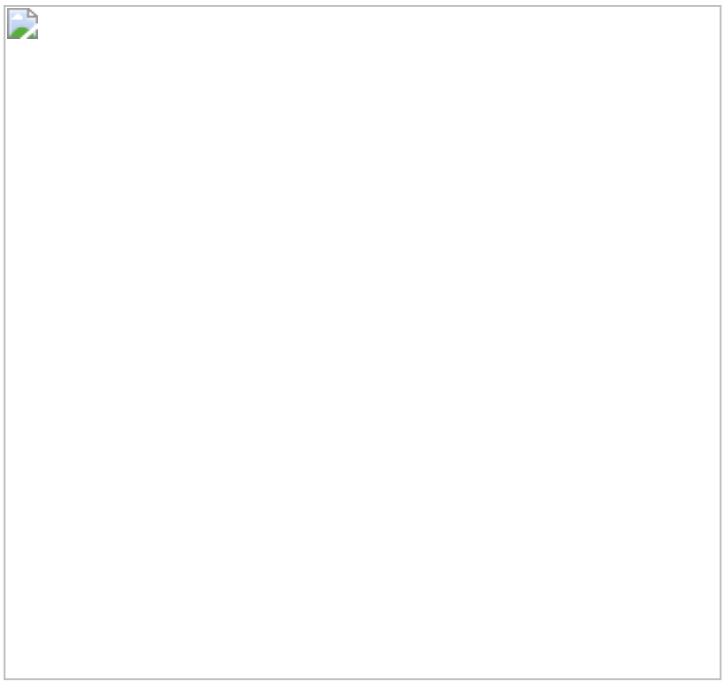
They're about stuff that makes stuff happen, and happens because of other stuff.

In this world there are well-paid professional marketers, including philosophical and theological marketers, who have thousands of hours of practice convincing customers that their beliefs are beyond the reach of science. But those marketers don't know about causal models. They may know about - know how to lie persuasively relative to - the epistemology used by a Traditional Rationalist, but that's crude by the standards of today's rationality-with-math. Highly Advanced Epistemology hasn't diffused far enough for there to be explicit anti-epistemology against it.

And so we shouldn't *expect* to find anyone with a background story which would justify evading science's skeptical gaze. As a matter of cognitive science, it seems extremely likely that the human brain natively represents something like causal structure - that this native representation is how your own brain knows that "If the radio says there was an earthquake, it's less likely that your burglar alarm going off implies a burglar." People who want to evade the gaze of science haven't read Judea Pearl's book; they don't know enough about formal causality to *not* automatically reason this way about things they claim are in separate magisteria. They can say words like "It's not mechanistic", but they don't have the mathematical fluency it would take to deliberately design a system outside Judea Pearl's box.

So in all probability, when somebody says, "I communed holistically and in a purely spiritual fashion with the entire universe - that's how I know my partner loves me, not because of any *mechanism*", their brain is just representing something like this:

Partner loves	**Universe knows**	**I hear universe**	**%**
p	u	h	0.44
p	u	¬h	0.023
p	¬u	h	0.01
p	¬u	¬h	0.025
¬p	u	h	0.43
¬p	u	¬h	0.023
¬p	¬u	h	0.015
¬p	¬u	¬h	0.035



True, false, or meaningless, this belief isn't beyond investigation by standard rationality.

Because *causality* isn't a word for a special, restricted domain that scientists study. 'Causal process' sounds like an impressive formal word that would be used by people in lab coats with doctorates, but that's not what it *means.*

'Cause and effect' just means "stuff that makes stuff happen and happens because of other stuff". Any time there's a noun, a verb, and a subject, there's causality. If the universe spoke to you in your heart - then the universe would be making stuff happen inside your heart! All the standard theorems would still apply.

Whatever people try to imagine that science supposedly can't analyze, it just ends up as more "stuff that makes stuff happen and happens because of other stuff".

Mainstream status.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Causal Reference"

Previous post: "Causal Diagrams and Causal Models"

Causal Reference

Followup to: The Fabric of Real Things, Stuff That Makes Stuff Happen

Previous meditation: "Does your rule forbid epiphenomenalist theories of consciousness that consciousness is caused by neurons, but doesn't affect those neurons in turn? The classic argument for epiphenomenal consciousness is that we can imagine a universe where people behave exactly the same way, but there's nobody home - no awareness, no consciousness, inside the brain. For all the atoms in this universe to be in the same place - for there to be no detectable difference *internally,* not just externally - 'consciousness' would have to be something created by the atoms in the brain, but which didn't affect those atoms in turn. It would be an effect of atoms, but not a cause of atoms. Now, I'm not so much interested in whether you think epiphenomenal theories of consciousness are true or false - rather, I want to know if you think they're impossible or meaningless *a priori* based on your rules."

Is it coherent to imagine a universe in which a real entity can be an effect but not a cause?

Well... there's a couple of senses in which it seems *imaginable.* It's important to remember that imagining things yields info primarily about what human brains can imagine. It only provides info about reality to the extent that we think imagination and reality are systematically correlated for some reason.

That said, I can certainly write a computer program in which there's a tier of objects affecting each other, and a second tier - a lower tier - of epiphenomenal objects which are affected by them, but don't affect them. For example, I could write a program to simulate some balls that bounce off each other, and then some little shadows that follow the balls around.

But then I only know about the shadows because I'm outside that whole universe, looking in. So *my mind* is being affected by both the balls and shadows - to observe something is to be affected by it. I know where the shadow is, because the shadow makes pixels be drawn on screen, which make my eye see pixels. If your universe has two tiers of causality - a tier with things that affect each other, and another tier of things that are affected by the first tier without affecting them - then could you know that fact from *inside* that universe?

Again, this seems easy to *imagine* as long as objects in the second tier can affect *each other.* You'd just have to be living in the second tier! We can imagine, for example - this wasn't the way things worked out in *our* universe, but it might've seemed plausible to the ancient Greeks - that the stars in heaven (and the Sun as a special case) could affect *each other* and affect Earthly forces, but no Earthly force could affect them:

(Here the X'd-arrow stands for 'cannot affect'.)

The Sun's light would illuminate Earth, so it would cause plant growth. And sometimes you would see two stars crash into each other and explode, so you'd see they could affect each other. (And affect your brain, which was seeing them.) But the stars and Sun would be made out of a different substance, the 'heavenly material', and throwing any Earthly material at it would not cause it to change state in the slightest. The Earthly material might be burned up, but the Sun would occupy exactly the same position as before. It would affect us, but not be affected by us.

(To clarify an important point raised in the comments: In standard causal diagrams and in standard physics, no two *individual events* ever affect *each other;* there's a causal arrow from the PAST to FUTURE but never an arrow from FUTURE to PAST. What we're talking about here is the sun and stars *over time,* and the *generalization over* causal arrows that point from Star-in-Past to Sun-in-Present and Sun-in-Present back to Star-in-Future. The standard formalism dealing with this would be Dynamic Bayesian Networks (DBNs) in which there are repeating nodes and repeating arrows for each successive timeframe: $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, and causal laws F relating $\mathbf{X}_i$ to $\mathbf{X}_{i+1}$. If the laws of physics did *not* repeat over time, it would be rather hard to learn about the universe! The Sun *repeatedly* sends out photons, and they obey the same laws each time they fall on Earth; rather than the F_i being new transition tables each time, we see a constant $F_{physics}$ over and over. By saying that we live in a single-tier universe, we're observing that whenever there are F-arrows, causal-link-types, which (over repeating time) descend from variables-of-type-X to variables-of-type-Y (like present photons affecting future electrons), there are *also* arrows going back from Ys to Xs (like present electrons affecting future photons). If we *weren't* generalizing over time, it couldn't possibly make sense to speak of thingies that "affect each other" - causal diagrams don't allow directed cycles!)

A two-tier causal universe seems easy to imagine, even easy to specify as a computer program. If you were arranging a Dynamic Bayes Net at random, would it *randomly* have everything in a single tier? If you were designing a causal universe at random, wouldn't there randomly be some things that appeared to us as causes but not effects? And yet our own physicists haven't discovered any upper-tier particles which can move us without being movable by us. There might be a hint here at what sort of thingies tend to be real in the first place - that, for whatever reasons, the Real Rules somehow mandate or suggest that all the causal forces in a universe be on the same level, capable of both affecting and being affected by each other.

Still, we don't actually *know* the Real Rules are like that; and so it seems premature to assign *a priori* zero probability to hypotheses with multi-tiered causal universes. Discovering a class of upper-tier affect-only particles seems imaginable[1] - we can imagine which experiences would convince us that they existed. If we're in the Matrix,

we can see how to program a Matrix like that. If there's some deeper reason why that's *impossible* in any base-level reality, we don't know it yet. So we probably want to call that a meaningful hypothesis for now.

But what about lower-tier particles which can be affected by us, and yet never affect us?

Perhaps there are whole sentient Shadow Civilizations living on my nose hairs which can never *affect* those nose hairs, but find my nose hairs solid beneath their feet. (The solid Earth affecting them but not being affected, like the Sun's light affecting us in the 'heavenly material' hypothesis.) Perhaps I wreck their world every time I sneeze. It certainly seems imaginable - you could write a computer program simulating physics like that, given sufficient perverseness and computing power...

And yet the fundamental question of rationality - "What do you think you know, and how do you think you know it?" - raises the question:

How could you possibly know about the lower tier, even if it existed?

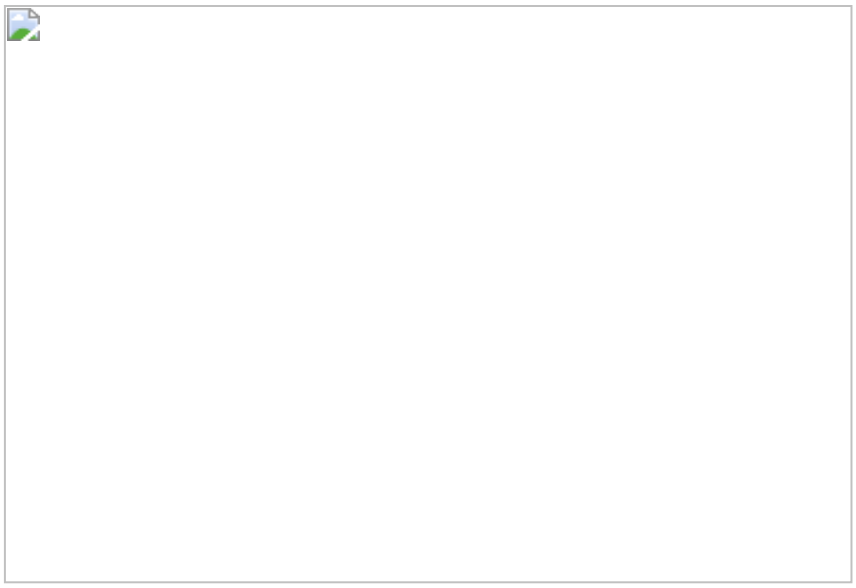
To observe something is to be affected by it - to have your brain and beliefs take on different states, depending on that thing's state. How can you know about something that doesn't affect your brain?

In fact there's an even deeper question, "How could you possibly *talk about* that lower tier of causality even if it existed?"

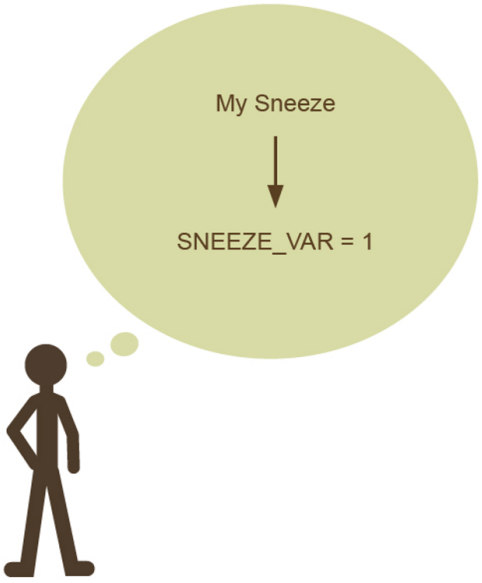
Let's say you're a Lord of the Matrix. You write a computer program which first computes the physical universe as we know it (or a discrete approximation), and then you add a couple of lower-tier effects as follows:

First, every time I sneeze, the binary variable YES_SNEEZE will be set to the second of its two possible values.

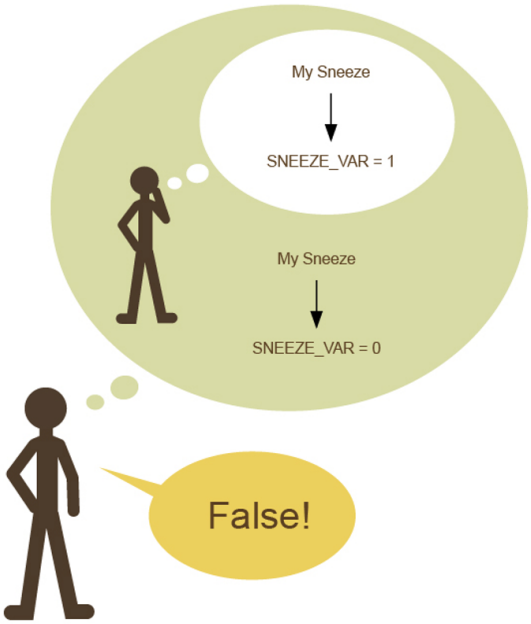
Second, every time I sneeze, the binary variable NO_SNEEZE will be set to the first of its two possible values.



Now let's say that - somehow - even though I've never caught any hint of the Matrix - I just *magically* think to myself one day, "What if there's a variable that watches when I sneeze, and gets set to 1?"



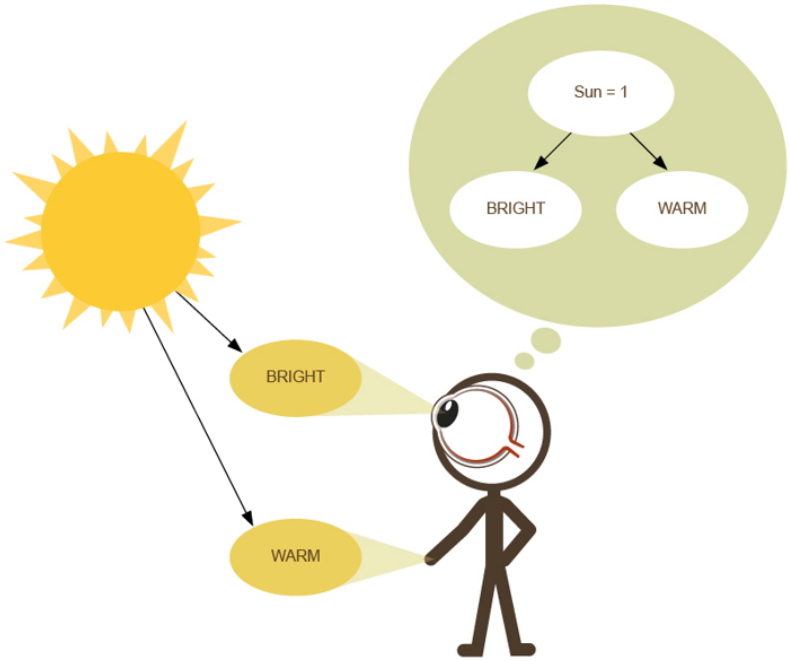
It will be all too easy for me to imagine that this belief is meaningful and could be true or false:



And yet in reality - as *you* know from outside the matrix - there are *two* shadow variables that get set when I sneeze. How can I talk about one of them, rather than the other? Why should my thought about '1' refer to their second possible value rather than their first possible value, inside the Matrix computer program? If we tried to establish a truth-value in this situation, to compare my *thought* to the reality inside the computer program - why compare my thought about SNEEZE_VAR to the variable YES_SNEEZE instead of NO_SNEEZE, or compare my thought '1' to the first possible value instead of the second possible value?

Under more epistemically healthy circumstances, when you talk about things that are not directly sensory experiences, you will reference a causal model of the universe that you inducted to *explain* your sensory experiences. Let's say you repeatedly go outside at various times of day, and your eyes and skin directly experience BRIGHT-WARM, BRIGHT-WARM, BRIGHT-WARM, DARK-COOL, DARK-COOL, etc. To explain the patterns in your sensory experiences, you hypothesize a latent variable we'll call 'Sun', with some kind of state which can change between 1, which causes BRIGHTness and WARMness, and 0, which causes DARKness and COOLness. You believe that the state of the 'Sun' variable changes over time, but usually changes less frequently than you go outside.

p(BRIGHT|Sun=1)	0.9
p(¬BRIGHT|Sun=1)	0.1
p(BRIGHT|Sun=0)	0.1
p(¬BRIGHT|Sun=0)	0.9



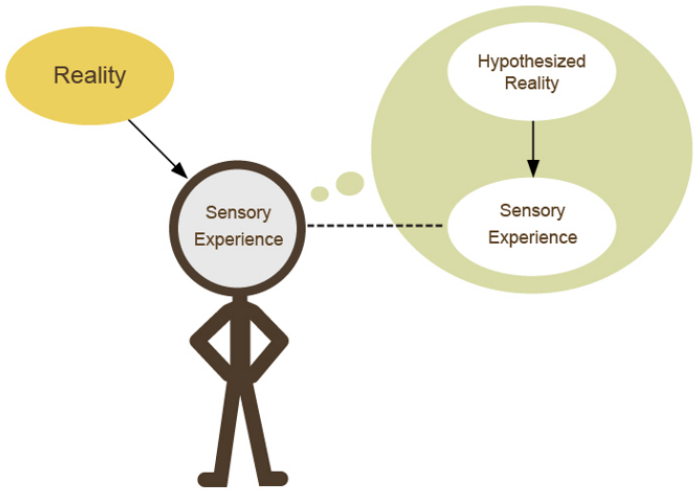
Standing here *outside* the Matrix, we might be tempted to compare your *beliefs* about "Sun = 1", to the real universe's state regarding the visibility of the sun in the sky (or rather, the Earth's rotational position).

But even if we compress the sun's visibility down to a binary categorization, how are we to know that your thought "Sun = 1" is meant to correspond to the sun being visible in the sky, rather than the sun being occluded by the Earth? Why the first state of the variable, rather than the second state?

How indeed are we know that this thought "Sun = 1" is meant to compare to the sun at all, rather than an anteater in Venezuela?

Well, because that 'Sun' thingy is supposed to be the *cause* of BRIGHT and WARM feelings, and if you trace back the cause of those sensory experiences *in reality* you'll arrive at the sun that the 'Sun' thought allegedly corresponds to. And to distinguish between whether the sun being visible in the sky is meant to correspond to 'Sun'=1 or 'Sun'=0, you check the conditional probabilities for that 'Sun'-state giving rise to BRIGHT - if the actual sun being visible has a 95% chance of causing the BRIGHT sensory feeling, then that true state of the sun is intended to correspond to the hypothetical 'Sun'=1, not 'Sun'=0.

Or to put it more generally, in cases where we have...

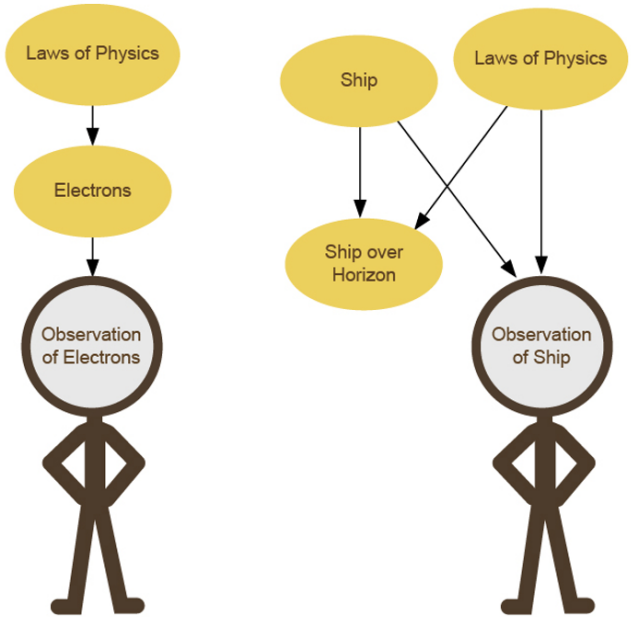


...then the correspondence between map and territory can at least *in principle* be point-wise evaluated by tracing causal links back from sensory experiences to reality, and tracing hypothetical causal links from sensory experiences back to hypothetical reality. We can't directly evaluate that truth-condition inside our own thoughts; but we can perform experiments and be corrected by them.

Being able to *imagine* that your thoughts are meaningful and that a correspondence between map and territory is being maintained, is no guarantee that your thoughts are true. On the other hand, if you *can't even imagine within your own model* how a piece of your map could have a traceable correspondence to the territory, that is a very bad sign for the belief being meaningful, let alone true. Checking to see whether you can *imagine* a belief being meaningful is a test which will occasionally throw out bad beliefs, though it is no guarantee of a belief being good.

Okay, but what about the idea that it should be meaningful to talk about whether or not a spaceship continues to exist after it travels over the cosmological horizon? Doesn't this theory of meaningfulness seem to claim that you can only sensibly imagine something that makes a difference to your sensory experiences?

No. It says that you can only talk about events that your sensory experiences *pin down within the causal graph*. If you observe enough protons, electrons, neutrons, and so on, you can pin down the physical generalization which says, "Mass-energy is neither created nor destroyed; and in particular, particles don't vanish into nothingness without a trace." It is then an *effect* of that rule, combined with our previous observation of the ship itself, which tells us that there's a ship that went over the cosmological horizon and now we can't see it any more.



To navigate referentially to the fact that the ship continues to exist over the cosmological horizon, we navigate from our sensory experience *up to* the laws of physics, by talking about the *cause* of electrons not blinking out of existence; we also navigate *up to* the ship's existence by tracing back the cause of our observation of the ship being built. We can't *see* the future ship over the horizon - but the causal links *down from* the ship's construction, and from the laws of physics saying it doesn't disappear, are both *pinned down by observation* - there's no difficulty in figuring out which causes we're talking about, or what effects they have.[2]

All righty-ighty, let's revisit that meditation:

"Does your rule forbid epiphenomenalist theories of consciousness in which consciousness is caused by neurons, but doesn't affect those neurons in turn? The classic argument for epiphenomenal consciousness is that we can imagine a universe where people behave exactly the same way, but there's nobody home - no awareness, no consciousness, inside the brain. For all the atoms in this universe to be in the same place - for there to be no detectable difference *internally,* not just externally - 'consciousness' would have to be something created by the atoms in the brain, but which didn't affect those atoms in turn. It would be an effect of atoms, but not a cause of atoms. Now, I'm not so much interested in whether you think epiphenomenal theories of consciousness are true or false - rather, I want to know if you think they're impossible or meaningless *a priori* based on your rules."

The closest theory to this which definitely *does* seem coherent - i.e., it's *imaginable* that it has a pinpointed meaning - would be if there was *another* little brain living inside my brain, made of shadow particles which could affect each other and be affected by my brain, but not affect my brain in turn. This brain would correctly hypothesize the reasons for its sensory experiences - that there was, from its perspective, an upper tier of particles interacting with each other that it couldn't affect. Upper-tier particles are observable, i.e., can affect lower-tier senses, so it would be possible to correctly induct a simplest explanation for them. And this inner brain would think, "I can imagine a Zombie Universe in which *I* am missing, but all the upper-

tier particles go on interacting with each other as before." If we imagine that the upper-tier brain is just a robotic sort of agent, or a kitten, then the inner brain might justifiably imagine that the Zombie Universe would contain nobody to listen - no lower-tier brains to watch and be aware of events.

We could write that computer program, given significantly more knowledge and vastly more computing power and zero ethics.

But this inner brain composed of lower-tier shadow particles *cannot* write upper-tier philosophy papers about the Zombie universe. If the inner brain thinks, "I am aware of my own awareness", the upper-tier lips cannot move and say aloud, "I am aware of my own awareness" a few seconds later. That would require causal links from lower particles to upper particles.

If we try to suppose that the lower tier isn't a complicated brain with an independent reasoning process that can imagine its own hypotheses, but just some shadowy pure experiences that don't affect anything in the upper tier, then clearly the *upper-tier brain* must be thinking meaningless gibberish when the *upper-tier lips* say, "I have a lower tier of shadowy pure experiences which did not affect in any way how I said these words." The deliberating upper brain that invents hypotheses for sense data, can only use sense data that affects the upper neurons carrying out the search for hypotheses that can be reported by the lips. Any shadowy pure experiences couldn't be inputs into the hypothesis-inventing cognitive process. So the upper brain would be talking nonsense.

There's a version of this theory in which the part of our brain that we can report out loud, which invents hypotheses to explain sense data out loud and manifests physically visible papers about Zombie universes, *has for no explained reason* invented a meaningless theory of shadow experiences which is experienced *by the shadow part* as a meaningful and correct theory. So that if we look at the "merely physical" slice of our universe, philosophy papers about consciousness are meaningless and the physical part of the philosopher is saying things their physical brain couldn't possibly know even if they were true. And yet our inner experience of those philosophy papers is meaningful and true. In a way that couldn't possibly have caused me to physically write the previous sentence, mind you. And yet your experience of that sentence is also true even though, in the upper tier of the universe where that sentence was actually written, it is not only false but meaningless.

I'm honestly not sure what to say when a conversation gets to that point. Mostly you just want to yell, "Oh, for the love of Belldandy, will you just give up already?" or something about the importance of saying oops.

(Oh, plus the unexplained correlation violates the Markov condition for causal models.)

Maybe my reply would be something along the lines of, "Okay... look... I've given my account of a single-tier universe in which agents can invent meaningful explanations for sense data, and when they build accurate maps of reality there's a known reason for the correspondence... if you want to claim that a *different* kind of meaningfulness can hold within a *different* kind of agent divided into upper and lower tiers, it's up to *you* to explain what parts of the agent are doing which kinds of hypothesizing and how those hypotheses end up being meaningful and what causally explains their miraculous accuracy so that this all makes *sense*."

But frankly, I think people would be wiser to just *give up* trying to write sensible philosophy papers about lower causal tiers of the universe that don't affect the

philosophy papers in any way.

Meditation: If we can only meaningfully talk about parts of the universe that can be pinned down inside the causal graph, where do we find the fact that 2 + 2 = 4? Or did I just make a meaningless noise, there? Or if you claim that "2 + 2 = 4" *isn't* meaningful or true, then what alternate property does the sentence "2 + 2 = 4" have which makes it so much more useful than the sentence "2 + 2 = 3"?

Mainstream status.

---

[1] Well, it seems imaginable so long as you toss most of quantum physics out the window and put us back in a classical universe. For particles to not be affected by us, they'd need their own configuration space such that "which configurations are identical" was determined by looking only at those particles, and not looking at any lower-tier particles entangled with them. If you *don't* want to toss QM out the window, it's actually pretty hard to imagine what an upper-tier particle would look like.

[2] This diagram treats the laws of physics as being just another node, which is a convenient shorthand, but probably not a good way to draw the graph. The laws of physics really correspond to the causal arrows F_i, not the causal nodes X_i. If you had the laws themselves - the function from past to future - be an X_i of variable state, then you'd need meta-physics to describe the $F_{physics}$ arrows for how the physics-stuff $X_{physics}$ could affect us, followed promptly by a need for meta-meta-physics et cetera. If the laws of physics *were* a kind of causal stuff, they'd be an upper tier of causality - we can't appear to affect the laws of physics, but if you call them causes, they can affect us. In Matrix terms, this would correspond to our universe running on a computer that stored the laws of physics in one area of RAM and the state of the universe in another area of RAM, the first area would be an upper causal tier and the second area would be a lower causal tier. But the infinite regress from treating the laws of determination as causal stuff, makes me suspicious that it might be an error to treat the laws of physics as "stuff that makes stuff happen and happens because of other stuff". When we trust that the ship doesn't disappear when it goes over the horizon, we may not be navigating to a physics-node in the graph, so much as we're navigating to a single $F_{physics}$ that appears in many different places inside the graph, and whose previously unknown function we have inferred. But this is an unimportant technical quibble on Tuesdays, Thursdays, Saturdays, and Sundays. It is only an incredibly deep question about the nature of reality on Mondays, Wednesdays, and Fridays, i.e., less than half the time.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Proofs, Implications, and Models"

Previous post: "Stuff That Makes Stuff Happen"

Causal Universes

Followup to: Stuff that Makes Stuff Happen

> Previous meditation: Does the idea that everything is made of causes and effects meaningfully constrain experience? Can you coherently say how reality might look, if our universe did *not* have the kind of structure that appears in a causal model?

I can describe to you at least one famous universe that *didn't* look like it had causal structure, namely the universe of J. K. Rowling's Harry Potter.

You might think that J. K. Rowling's universe doesn't have causal structure because it contains magic - that wizards wave their wands and cast spells, which doesn't make any sense and goes against all science, so J. K. Rowling's universe isn't 'causal'.

In this you would be completely mistaken. The domain of "causality" is just "stuff that makes stuff happen and happens because of other stuff". If Dumbledore waves his wand and therefore a rock floats into the air, that's causality. You don't even have to use words like 'therefore', let alone big fancy phrases like 'causal process', to put something into the lofty-sounding domain of causality. There's causality anywhere there's a noun, a verb, and a subject: 'Dumbledore's wand lifted the rock.' So far as I could tell, there wasn't anything in *Lord of the Rings* that violated causality.

You might worry that J. K. Rowling had made a continuity error, describing a spell working one way in one book, and a different way in a different book. But we could just suppose that the spell had changed over time. If we actually found ourselves in that apparent universe, and saw a spell have two different effects on two different occasions, we would not conclude that our universe was uncomputable, or that it couldn't be made of causes and effects.

No, the *only* part of J. K. Rowling's universe that violates 'cause and effect' is...

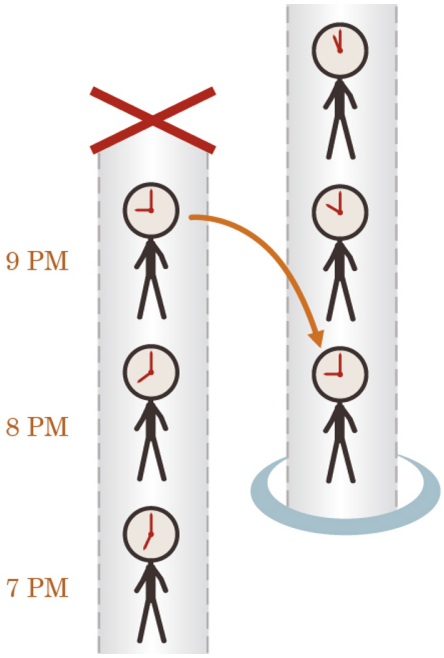
...
...
...

...the Time-Turners, of course.

A Time-Turner, in Rowling's universe, is a small hourglass necklace that sends you back in time 1 hour each time you spin it. In Rowling's universe, this time-travel doesn't allow for *changing* history; whatever you do after you go back, it's already happened. The universe containing the time-travel is a stable, self-consistent object.

If a time machine does allow for changing history, it's easy to imagine how to compute it; you could easily write a computer program which would simulate that universe and its time travel, given sufficient computing power. You would store the state of the universe in RAM and simulate it under the programmed 'laws of physics'. Every nanosecond, say, you'd save a copy of the universe's state to disk. When the Time-Changer was activated at 9pm, you'd retrieve the saved state of the universe from one hour ago at 8pm, load it into RAM, and then insert the Time-Changer and its user in the appropriate place. This would, of course, dump the *rest* of the universe

from 9pm into oblivion - no processing would continue onward from that point, which is the same as ending that world and killing everyone in it.[1]



Still, if we don't worry about the ethics or the disk space requirements, then a Time-Changer which can restore and then change the past is easy to *compute.* There's a perfectly clear order of causality in metatime, in the linear time of the simulating computer, even if there are apparent cycles as seen from *within* the universe. The person who suddenly appears with a Time-Changer is the causal descendant of the older universe that just got dumped from RAM.

But what if instead, reality is always - somehow - perfectly self-consistent, so that there's apparently only *one* universe with a future and a past that never changes, so that the person who appears at 8PM has always seemingly descended from *the very same universe* that then develops by 9PM...?

How would you compute *that* in one sweep-through, without any higher-order metatime?

What would a causal graph for *that* look like, when the past descends from its very own future?

And the answer is that there isn't any such causal graph. Causal models are sometimes referred to as DAGs, which stands for Directed Acyclic Graph. If instead there's a *directed cycle*, there's no obvious order in which to compute the joint probability table. Even if you somehow knew that at 8PM somebody was going to appear with a Time-Turner used at 9PM, you still couldn't compute the exact state of the time-traveller without already knowing the future at 9PM, and you couldn't compute the future without knowing the state at 8PM, and you couldn't compute the state at 8PM without knowing the state of the time-traveller who just arrived.

In a causal model, you can compute p(9pm|8pm) and p(8pm|7pm) and it all starts with your unconditional knowledge of p(7pm) or perhaps the Big Bang, but with a

Time-Turner we have p(9pm|8pm) and p(8pm|9pm) and we can't untangle them - multiplying those two conditional matrices together would just yield nonsense.

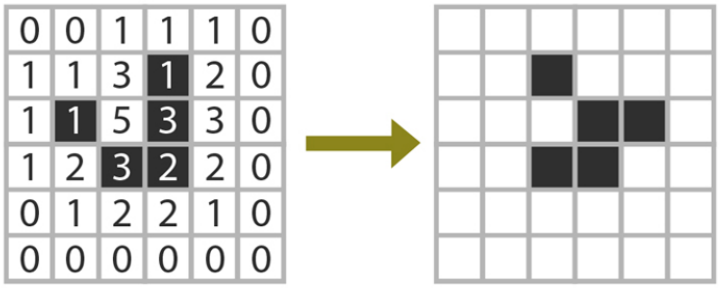
Does this mean that the Time-Turner is beyond all logic and reason?

Complete philosophical panic is basically never justified. We should even be reluctant to say anything like, "The so-called Time-Turner is beyond coherent description; we only think we can imagine it, but really we're just talking nonsense; so we can conclude *a priori* that no such Time-Turner that can exist; in fact, there isn't even a meaningful thing that we've just proven can't exist." This is *also* panic - it's just been made to sound more dignified. The first rule of science is to accept your experimental results, and generalize based on what you see. What if we actually *did* find a Time-Turner that seemed to work like that? We'd just have to accept that Causality As We Previously Knew It had gone out the window, and try to make the best of that.

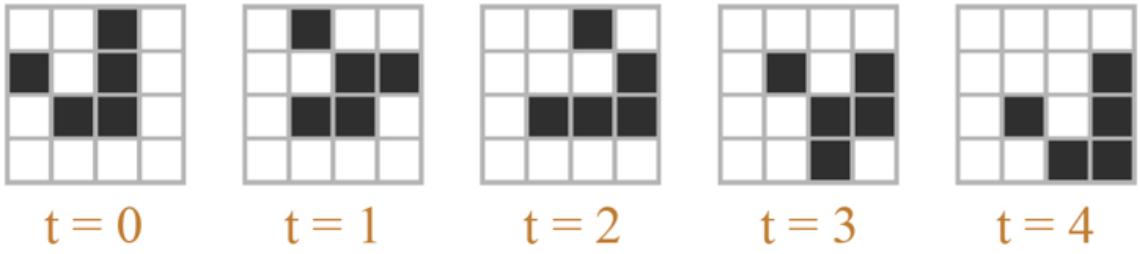
In fact, despite the somewhat-justified conceptual panic which the protagonist of *Harry Potter and the Methods of Rationality* undergoes upon seeing a Time-Turner, a universe like that can have a straightforward *logical* description even if it has no *causal* description.

Conway's Game of Life is a very simple specification of a causal universe; what we would today call a cellular automaton. The Game of Life takes place on a two-dimensional square grid, so that each cell is surrounded by eight others, and the Laws of Physics are as follows:

- A cell with 2 living neighbors during the last tick, retains its state from the last tick.
- A cell with 3 living neighbors during the last tick, will be alive during the next tick.
- A cell with fewer than 2 or more than 3 living neighbors during the last tick, will be dead during the next tick.



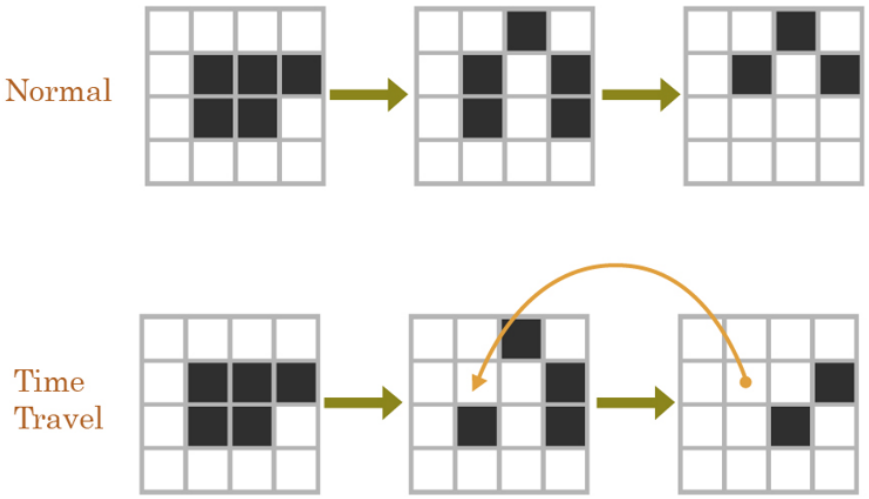
It is my considered opinion that everyone should play around with Conway's Game of Life at some point in their lives, in order to comprehend the notion of 'laws of physics'. Playing around with Life as a kid (on a Mac Plus) helped me gut-level-understand the concept of a 'lawful universe' developing under exceptionless rules.



Now suppose we modify the Game of Life universe by adding some prespecified cases of *time travel* - places where a cell will descend from neighbors in the future, instead of the past.

In particular we shall take a 4x4 Life grid, and arbitrarily hack Conway's rules to say:

- On the 2nd tick, the cell at (2,2) will have its state determined by that cell's state on the 3rd tick, instead of its neighbors on the 1st tick.



It's no longer possible to compute the state of each cell at each time *in a causal order* where we start from known cells and compute their not-yet-known causal descendants. The state of the cells on the 3rd tick, depend on the state of the cells on the 2nd tick, which depends on the state on the 3rd tick.

In fact, the time-travel rule, on the same initial conditions, also permits a live cell to travel back in time, not just a dead cell - this just gives us the "normal" grid! Since you can't compute things in order of cause and effect, even though each local rule is deterministic, the global outcome is not determined.

However, you *could* simulate Life with time travel *merely* by brute-force searching through all possible Life-histories, discarding all histories which disobeyed the laws of Life + time travel. If the entire universe were a 4-by-4 grid, it would take 16 bits to specify a single slice through Time - the universe's state during a single clock tick. If the whole of Time was only 3 ticks long, there would be only 48 bits making up a candidate 'history of the universe' - it would only take 48 bits to completely specify a

History of Time. 2^48 is just 281,474,976,710,656, so with a cluster of 2GHz CPUs it would be quite practical to find, for this *rather tiny* universe, the set of all possible histories that obey the *logical relations* of time travel.

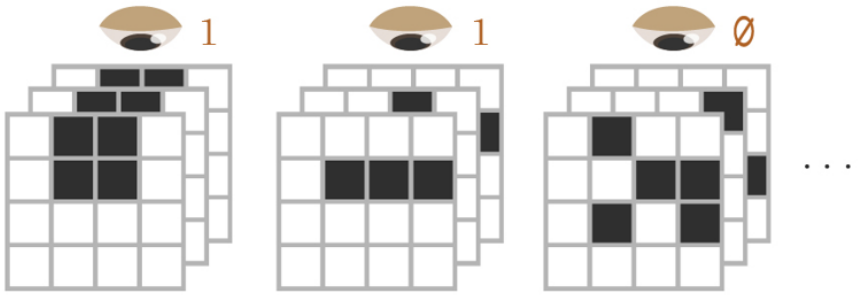
It would no longer be possible to point to a particular cell in a particular history and say, "This is *why* it has the 'alive' state on tick 3". There's no "reason" - in the framework of causal reasons - why the time-traveling cell is 'dead' rather than 'alive', in the history we showed. (Well, except that Alex, in the real universe, happened to pick it out when I asked him to generate an example.) But you could, in principle, find out what the set of permitted histories for a large digital universe, given *lots and lots* of computing power.

Here's an interesting question I do *not* know how to answer: Suppose we had a more *complicated* set of cellular automaton rules, on a vastly larger grid, such that the cellular automaton was large enough, and supported enough complexity, to permit *people* to exist inside it and be computed. Presumably, if we computed out cell states in the ordinary way, each future following from its immediate past, the people inside it would be as real as we humans computed under our own universe's causal physics.

Now suppose that instead of computing the cellular automaton causally, we hack the rules of the automaton to add large time-travel loops - change their physics to allow Time-Turners - and with an *unreasonably large* computer, the size of *two to the power of* the number of bits comprising an *entire history of the cellular automaton,* we enumerate all possible candidates for a universe-history.

So far, we've just generated all 2^N possible bitstrings of size N, for some large N; nothing more. You wouldn't expect this procedure to generate any people or make any experiences real, unless enumerating all finite strings of size N causes all lawless universes encoded in them to be real. There's no causality there, no computation, no law relating one time-slice of a universe to the next...

Now we set the computer to look over this entire set of candidates, and mark with a 1 those that obey the modified relations of the time-traveling cellular automaton, and mark with a 0 those that don't.



If N is large enough - if the size of the possible universe and its duration is large enough - there would be descriptions of universes which experienced natural selection, evolution, perhaps the evolution of intelligence, and of course, time travel with self-consistent Time-Turners, obeying the modified relations of the cellular automaton. And the checker would mark those descriptions with a 1, and all others with a 0.

Suppose we pick out one of the histories marked with a 1 and look at it. It seems to contain a description of people who remember experiencing time travel.

Now, were their experiences real? Did we *make* them real by marking them with a 1 - by applying the logical filter using a causal computer? Even though there was no way of computing future events from past events; even though their universe isn't a causal universe; even though they will have had experiences that literally were not 'caused', that did not have any causal graph behind them, within the framework of their own universe and its rules?

I don't know. *But...*

Our *own* universe does *not* appear to have Time-Turners, and *does* appear to have strictly local causality in which each variable can be computed strictly forward-in-time.

And I don't know *why* that's the case; but it's a likely-looking *hint* for anyone wondering what sort of universes can be real in the first place.

The collection of hypothetical mathematical thingies that can be *described logically* (in terms of relational rules with consistent solutions) looks *vastly* larger than the collection of *causal universes* with locally determined, acyclically ordered events. Most mathematical objects aren't like that. When you say, "We live in a causal universe", a universe that can be computed in-order using local and directional rules of determination, you're *vastly narrowing down the possibilities* relative to all of Math-space.

So it's rather *suggestive* that we find ourselves in a causal universe rather than a logical universe - it suggests that not all mathematical objects can be real, and the sort of thingies that *can* be real and have people in them are constrained to somewhere in the vicinity of 'causal universes'. That you can't have consciousness without computing an agent made of causes and effects, or maybe something can't be real at all unless it's a fabric of cause and effect. It suggests that if there *is* a Tegmark Level IV multiverse, it isn't "all logical universes" but "all causal universes".

Of course you also have to be a bit careful when you start assuming things like "Only causal things can be real" because it's so easy for Reality to come back at you and shout "WRONG!" Suppose you thought reality had to be a *discrete* causal graph, with a finite number of nodes and discrete descendants, *exactly* like Pearl-standard causal models. There would be *no hypothesis in your hypothesis-space* to describe the standard model of physics, where space is continuous, indefinitely divisible, and has complex amplitude assignments over uncountable cardinalities of points.

Reality is primary, saith the wise old masters of science. The first rule of science is just to go with what you see, and try to understand it; rather than standing on your assumptions, and trying to argue with reality.

But *even so,* it's *interesting* that the pure, ideal structure of causal models, invented by statisticians to reify the idea of 'causality' as simply as possible, looks *much* more like the modern view of physics than does the old Newtonian ideal.

If you believed in Newtonian billiard balls bouncing around, and somebody asked you what sort of things can be real, you'd probably start talking about 'objects', like the billiard balls, and 'properties' of the objects, like their location and velocity, and how the location 'changes' between one 'time' and another, and so on.

But suppose you'd never heard of atoms or velocities or this 'time' stuff - just the causal diagrams and causal models invented by statisticians to represent the simplest possible cases of cause and effect. Like this:

And then someone says to you, "Invent a *continuous analogue* of this."

You wouldn't invent billiard balls. There's no billiard balls in a causal diagram.

You wouldn't invent a single time sweeping through the universe. There's no sweeping time in a causal diagram.

You'd stare a bit at B, C, and D which are the sole nodes determining A, screening off the rest of the graph, and say to yourself:

"Okay, how can I invent a *continuous* analogue of there being three nodes that screen off the rest of the graph? How do I do that with a continuous neighborhood of points, instead of three nodes?"

You'd stare at E determining D determining A, and ask yourself:

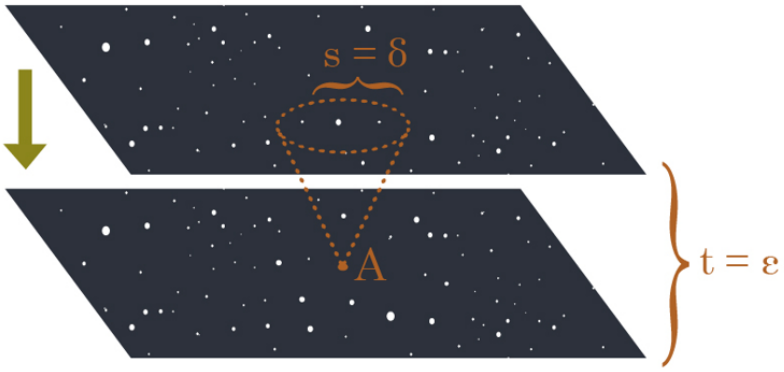
"How can I invent a *continuous* analogue of 'determination', so that instead of E determining D determinining A, there's a continuum of determined points between E and A?"

If you generalized in a certain simple and obvious fashion...

The continuum of relatedness from B to C to D would be what we call *space.*

The continuum of determination from E to D to A would be what we call *time.*

There would be a rule stating that for epsilon time before A, there's a neighborhood of spatial points delta which screens off the rest of the universe from being relevant to A (so long as no descendants of A are observed); and that epsilon and delta can both get arbitrarily close to zero.



There might be - if you were just picking the simplest rules you could manage - a physical constant which related the metric of relatedness (space) to the metric of determination (time) and so enforced a simple continuous analogue of local causality...

...in our universe, we call it *c,* the speed of light.

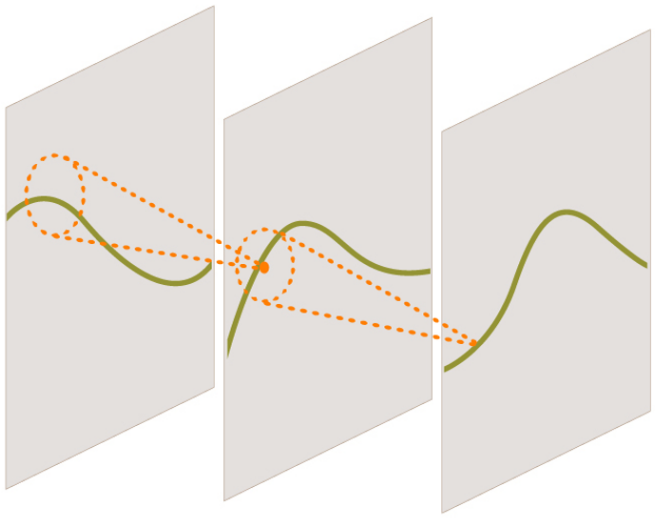
And it's worth remembering that Isaac Newton did *not* expect that rule to be there.

If we just stuck with Special Relativity, and didn't get any *more* modern than that, there would still be little billiard balls like electrons, occupying some particular point in that neighborhood of space.

But if your little neighborhoods of space have billiard balls with velocities, many of which are *slower* than lightspeed... well, that doesn't look like the simplest continuous analogues of a causal diagram, does it?

When we make the first quantum leap and describe particles as waves, we find that the billiard balls have been eliminated. There's no 'particles' with a single point position and a velocity slower than light. There's an electron *field,* and waves propagate through the electron field through points interacting only with locally neighboring points. If a particular electron seems to be moving slower than light, that's just because - even though causality always propagates at exactly *c* between points within the electron *field* - the crest of the electron *wave* can appear to move slower than that. A billiard ball moving through space over time, has been replaced by a set of points with values determined by their immediate historical neighborhood.

vs.



And when we make the second quantum leap into configuration space, we find a timeless universal wavefunction with complex amplitudes assigned over the points in that configuration space, and the amplitude of every point causally determined by its immediate neighborhood in the configuration space.[2]

So, yes, Reality can poke you in the nose if you decide that only discrete causal graphs can be real, or something silly like that.

But on the other hand, taking advice from the math of causality wouldn't always lead you astray. Modern physics looks a *heck* of a lot more similar to "Let's build a continuous analogue of the simplest diagrams statisticians invented to describe theoretical causality", than like anything Newton or Aristotle imagined by looking at the apparent world of boulders and planets.

I don't know what it means... but perhaps we shouldn't ignore the *hint* we received by virtue of finding ourselves inside the narrow space of "causal universes" - rather than the much wider space "all logical universes" - when it comes to guessing what sort of thingies can be real. To the extent we allow non-causal universes in our hypothesis space, there's a strong chance that we are broadening our imagination beyond what can *really* be real under the Actual Rules - whatever *they* are! (It *is* possible to broaden your metaphysics too much, as well as too little. For example, you could allow logical contradictions into your hypothesis space - collections of axioms with no models - and ask whether we lived in one of those.)

If we trusted absolutely that only causal universes could be real, then it would be safe to allow only causal universes into our hypothesis space, and assign probability literally zero to everything else.

But if you were scared of being wrong, then assigning probability literally *zero* means you can't change your mind, ever, even if Professor McGonagall shows up with a Time-Turner tomorrow.

Meditation: Suppose you needed to assign non-zero probability to any way things could conceivably turn out to be, given humanity's rather young and confused state - enumerate all the hypotheses a superintelligent AI should ever be able to arrive at, based on any sort of strange world it might find by observation of Time-Turners or

stranger things. How would you enumerate the hypothesis space of all the worlds we could remotely maybe possibly be living in, including worlds with hypercomputers and Stable Time Loops and even stranger features?

Mainstream status.

---

[1] Sometimes I still marvel about how in most time-travel stories nobody thinks of this. I guess it really is true that only people who are sensitized to 'thinking about existential risk' *even notice* when a world ends, or when billions of people are extinguished and replaced by slightly different versions of themselves. But then almost nobody will notice that sort of thing inside their fiction if the characters all act like it's okay.)

[2] Unless you believe in 'collapse' interpretations of quantum mechanics where Bell's Theorem mathematically requires that *either* your causal models don't obey the Markov condition *or* they have faster-than-light nonlocal influences. (Despite a large literature of obscurantist verbal words intended to obscure this fact, as generated and consumed by physicists who don't know about formal definitions of causality or the Markov condition.) If you believe in a collapse postulate, this whole post goes out the window. But frankly, if you believe that, you are bad and you should feel bad.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Mixed Reference: The Great Reductionist Project"

Previous post: "Logical Pinpointing"

Proofs, Implications, and Models

Followup to: Causal Reference

From a math professor's blog:

> One thing I discussed with my students here at HCSSiM yesterday is the question of what is a proof.
>
> They're smart kids, but completely new to proofs, and they often have questions about whether what they've written down constitutes a proof. Here's what I said to them.
>
> A proof is a social construct – it is what we need it to be in order to be convinced something is true. If you write something down and you want it to count as a proof, the only real issue is whether you're completely convincing.

This is not quite the definition I would give of what constitutes "proof" in mathematics - perhaps because I am so used to isolating arguments that are convincing, but ought not to be.

Or here again, from "An Introduction to Proof Theory" by Samuel R. Buss:

> There are two distinct viewpoints of what a mathematical proof is. The first view is that proofs are social conventions by which mathematicians convince one another of the truth of theorems. That is to say, a proof is expressed in natural language plus possibly symbols and figures, and is sufficient to convince an expert of the correctness of a theorem. Examples of social proofs include the kinds of proofs that are presented in conversations or published in articles. Of course, it is impossible to precisely define what constitutes a valid proof in this social sense; and, the standards for valid proofs may vary with the audience and over time. The second view of proofs is more narrow in scope: in this view, a proof consists of a string of symbols which satisfy some precisely stated set of rules and which prove a theorem, which itself must also be expressed as a string of symbols. According to this view, mathematics can be regarded as a 'game' played with strings of symbols according to some precisely defined rules. Proofs of the latter kind are called "formal" proofs to distinguish them from "social" proofs.

In modern mathematics there is a much better answer that could be given to a student who asks, "What exactly is a proof?", which does not match *either* of the above ideas. So:

Meditation: What distinguishes a correct mathematical proof from an incorrect mathematical proof - what does it mean for a mathematical proof to be good? And why, in the real world, would anyone ever be interested in a mathematical proof of this type, or obeying whatever goodness-rule you just set down? How could you use your notion of 'proof' to improve the real-world efficacy of an Artificial Intelligence?

...
...
...

Consider the following syllogism:

1. All kittens are little;
2. Everything little is innocent;
3. Therefore all kittens are innocent.

Here's four mathematical universes, aka "models", in which the objects collectively obey or disobey these three rules:

There are some models where not all kittens are little, like models B and D. And there are models where not everything little is innocent, like models C and D. But there are no models where all kittens are little, *and* everything little is innocent, and yet there exists a guilty kitten. Try as you will, you won't be able to imagine a model like that. Any model containing a guilty kitten has at least one kitten that isn't little, or at least one little entity that isn't innocent - no way around it.

Thus, the jump from 1 & 2, to 3, is *truth-preserving:* in any universe where premises (1) and (2) are true to start with, the conclusion (3) is true of the same universe at the end.

Which is what makes the following implication *valid*, or, as people would usually just say, "true":

"If all kittens are little and everything little is innocent, then all kittens are innocent."

The *advanced* mainstream view of logic and mathematics (i.e., the mainstream view among professional scholars of logic as such, not necessarily among all mathematicians in general) is that when we talk about math, we are talking about *which conclusions follow from which premises*. The "truth" of a mathematical theorem - or to not overload the word 'true' meaning comparison-to-causal-reality, the *validity* of a mathematical theorem - has nothing to do with the physical truth or falsity of the conclusion in our world, and everything to do with the inevitability of the *implication.* From the standpoint of *validity*, it doesn't matter a fig whether or not all kittens are

innocent in our *own* universe, the connected causal fabric within which we are embedded. What matters is whether or not you can prove the implication, starting from the premises; whether or not, if all kittens *were* little and all little things *were* innocent, it would follow *inevitably* that all kittens were innocent.

To paraphrase Mayor Daley, logic is not there to *create* truth, logic is there to *preserve* truth. Let's illustrate this point by assuming the following equation:

$$x = y = 1$$

...which is true in at least some cases. E.g. 'x' could be the number of thumbs on my right hand, and 'y' the number of thumbs on my left hand.

Now, starting from the above, we do a little algebra:

1	$x = y = 1$	starting premise
2	$x^2 = xy$	multiply both sides by x
3	$x^2 - y^2 = xy - y^2$	subtract y^2 from both sides
4	$(x + y)(x - y) = y(x - y)$	factor
5	$x + y = y$	cancel
6	$2 = 1$	substitute 1 for x and 1 for y

We have reached the conclusion that in every case where x and y are equal to 1, 2 is equal to 1. This does not seem like it should follow inevitably.

You could try to find the flaw just by staring at the lines... maybe you'd suspect that the error was between line 3 and line 4, following the heuristic of first mistrusting what looks like the most complicated step... but another way of doing it would be to try *evaluating* each line to see what it said concretely, for example, multiplying out $x^2 = xy$ in line 2 to get $(1^2) = (1 * 1)$ or $1 = 1$. Let's try doing this for each step, and then afterward mark whether each equation looks *true* or *false:*

1	$x = y = 1$	$1 = 1$	**true**
2	$x^2 = xy$	$1 = 1$	**true**
3	$x^2 - y^2 = xy - y^2$	$0 = 0$	**true**

4	(x + y)(x - y) = y(x - y)	0 = 0	**true**
5	x + y = y	2 = 1	**false**
6	2 = 1	2 = 1	**false**

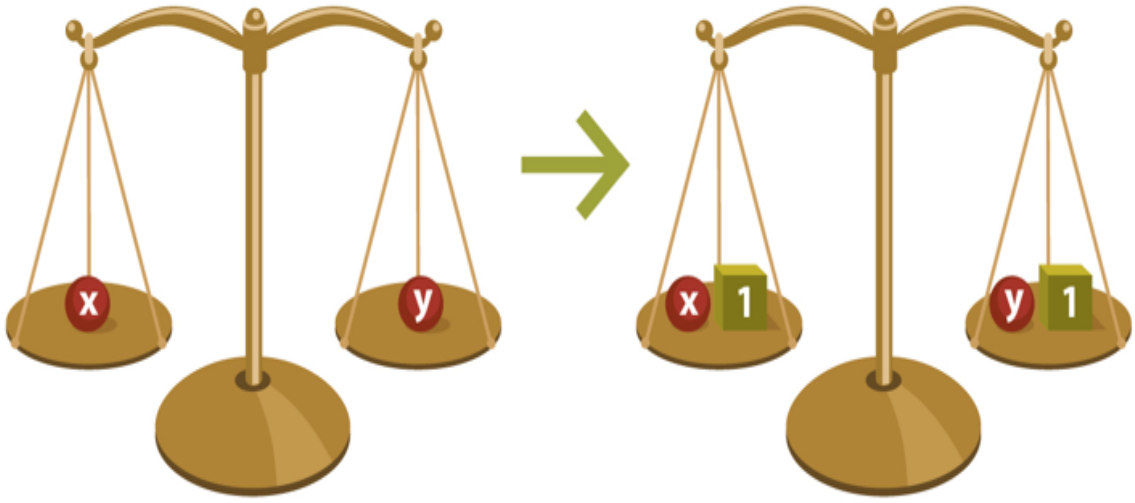
Logic is there to preserve truth, not to create truth. Whenever we take a logically valid step, we can't guarantee that the premise is true to start with, but *if* the premise is true the conclusion should always be true. Since we went from a true equation to a false equation between step 4 and step 5, we must've done something that is *in general* invalid.

In *particular,* we divided both sides of the equation by (x - y).

Which is invalid, i.e. *not universally truth-preserving,* because (x - y) might be equal to 0.

And if you divide both sides by 0, you can get a false statement from a true statement. 3 * 0 = 2 * 0 is a true equation, but 3 = 2 is a false equation, so it is not allowable in general to cancel *any* factor if the factor might equal zero.

On the other hand, adding 1 to both sides of an equation is *always* truth-preserving. We can't guarantee as a matter of logic that x = y to start with - for example, x might be my number of ears and y might be my number of fingers. But *if* x = y *then* x + 1 = y + 1, always. Logic is not there to create truth; logic is there to preserve truth. *If* a scale *starts out* balanced, then adding the same weight to both sides will result in a scale that is *still* balanced:



I will remark, in some horror and exasperation with the modern educational system, that I do not recall any math-book of my youth ever once explaining that the reason why you are always allowed to add 1 to both sides of an equation is that it is a kind of step which always produces true equations from true equations.

What is a valid proof in algebra? It's a proof where, in each step, we do something that is *universally allowed,* something which can only produce true equations from true equations, and so the proof gradually transforms the starting equation into a final equation which must be true if the starting equation was true. Each step should also - this is part of what makes proofs *useful in reasoning* - be *locally verifiable* as allowed, by looking at only a small number of previous points, not the entire past history of the proof. If in some previous step I believed $x^2 - y = 2$, I only need to look at that single step to get the conclusion $x^2 = 2 + y$, because I am always allowed to add y to both sides of the equation; because I am always allowed to add any quantity to both sides of the equation; because if the two sides of an equation are in balance to start with, adding the same quantity to both sides of the balance will preserve that balance. I can know the inevitability of this implication without considering all the surrounding circumstances; it's a step which is *locally* guaranteed to be *valid*. (Note the similarity - and the differences - to how we can compute a causal entity knowing only its immediate parents, and no other ancestors.)

You may have read - I've certainly read - some philosophy which endeavors to score points for counter-intuitive cynicism by saying that all mathematics is a *mere game of tokens;* that we start with a meaningless string of symbols like:

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$$

...and we follow some symbol-manipulation rules like "If you have the string 'A ∧ (A → B)' you are allowed to go to the string 'B'", and so finally end up with the string:

$$\forall x : K(x) \rightarrow I(x)$$

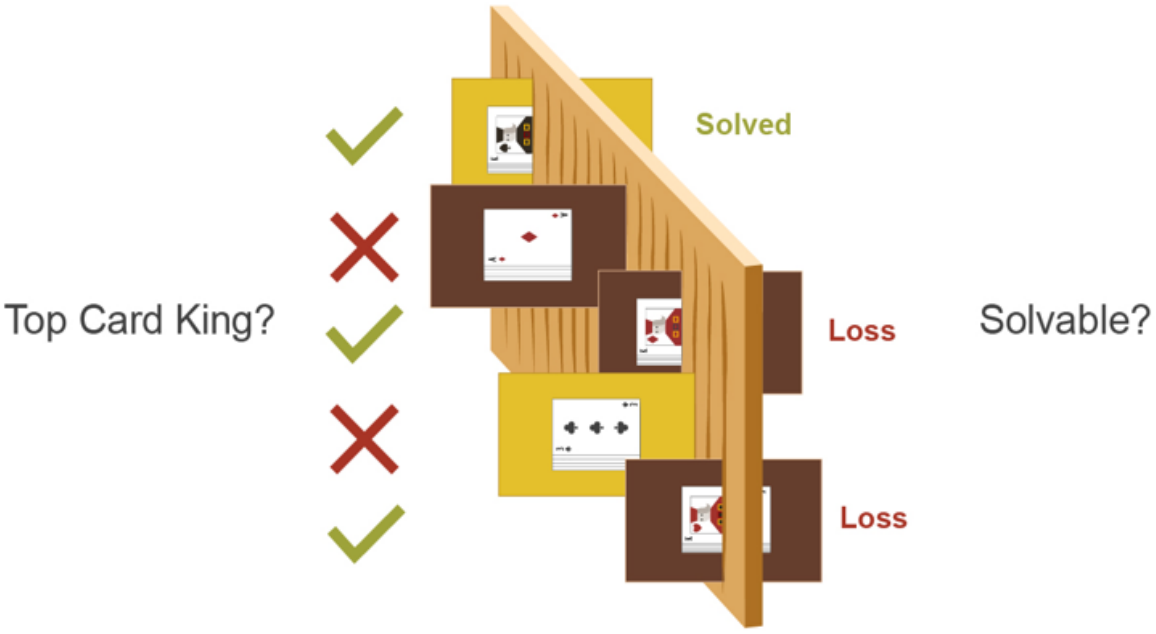
...and this activity of string-manipulation is all there is to what mathematicians call "theorem-proving" - all there is to the glorious human endeavor of mathematics.

This, like a lot of other cynicism out there, is *needlessly* deflationary.

There's a family of techniques in machine learning known as "Monte Carlo methods" or "Monte Carlo simulation", one of which says, roughly, "To find the probability of a proposition Q given a set of premises P, simulate random models that obey P, and then count how often Q is true." Stanislaw Ulam invented the idea after trying for a while to calculate the probability that a random Canfield solitaire layout would be solvable, and finally realizing that he could get better information by trying it a hundred times and counting the number of successful plays. This was during the era when computers were first becoming available, and the thought occurred to Ulam that the same technique might work on a current neutron diffusion problem as well.

Similarly, to answer a question like, "What is the probability that a random Canfield solitaire is solvable, given that the top card in the deck is a king?" you might imagine simulating many 52-card layouts, throwing *away* all the layouts where the top card in the deck was not a king, using a computer algorithm to solve the remaining layouts, and counting what percentage of those were solvable. (It would be more efficient, in this case, to start by directly placing a king on top and then randomly distributing the

other 51 cards; but this is not always efficient in Monte Carlo simulations when the condition to be fulfilled is more complex.)



Okay, now for a harder problem. Suppose you've wandered through the world a bit, and you've observed the following:

(1) So far I've seen 20 objects which have been kittens, and on the 6 occasions I've paid a penny to observe the size of something that's a kitten, all 6 kitten-objects have been little.

(2) So far I've seen 50 objects which have been little, and on the 30 occasions where I've paid a penny to observe the morality of something little, all 30 little objects have been innocent.

(3) This object happens to be a kitten. I want to know whether it's innocent, but I don't want to pay a cent to find out directly. (E.g., if it's an innocent kitten, I can buy it for a cent, sell it for two cents, and make a one-cent profit. But if I pay a penny to observe directly whether the kitten is innocent, I won't be able to make a profit, since gathering evidence is costly.)

Your previous experiences have led you to suspect the general rule "All kittens are little" and also the rule "All little things are innocent", even though you've never before *directly* checked whether a kitten is innocent.

Furthermore...

You've *never heard of logic,* and you have no idea how to play that 'game of symbols' with K(x), I(x), and L(x) that we were talking about earlier.

But that's all right. The problem is still solvable by Monte Carlo methods!

First we'll generate a large set of random universes. Then, for each universe, we'll check whether that universe obeys all the rules we currently suspect or believe to be

true, like "All kittens are little" and "All little things are innocent" and "The force of gravity goes as the square of the distance between two objects and the product of their masses". If a universe passes this test, we'll check to see whether the inquiry of interest, "Is the kitten in front of me innocent?", also happens to be true in that universe.

We shall repeat this test *a large number of times*, and at the end we shall have an approximate estimate of the probability that the kitten in front of you is innocent.

On this algorithm, you perform inference by visualizing many possible universes, throwing out universes that disagree with generalizations you already believe in, and then checking what's true (probable) in the universes that remain. This algorithm doesn't tell you the state of the real physical world with certainty. Rather, it gives you a measure of probability - i.e., the probability that the kitten is innocent - *given everything else you already believe to be true*.

And if, instead of visualizing many imaginable universes, you checked *all possible logical models* - which would take something beyond magic, because that would include models containing uncountably large numbers of objects - and the inquiry-of-interest was true in *every* model matching your previous beliefs, you would have found that the conclusion followed *inevitably* if the generalizations you already believed were true.

This might take a whole lot of reasoning, but at least you wouldn't have to pay a cent to observe the kitten's innocence directly.

But it would also *save you some computation* if you could play that *game of symbols* we talked about earlier - a game which does not create truth, but *preserves* truth. In this game, the steps can be *locally* pronounced valid by a mere 'syntactic' check that doesn't require us to visualize all possible models. Instead, if the mere *syntax* of the proof checks out, we know that the conclusion is always true in a model whenever the premises are true in that model.

And that's a mathematical proof: A conclusion which is true in any model where the axioms are true, which we know because we went through a series of transformations-of-belief, each step being licensed by some rule which guarantees that such steps never generate a false statement from a true statement.

The way we would say it in standard mathematical logic is as follows:

A collection of axioms X *semantically* implies a theorem Y, if Y is true in all models where X are true. We write this as $X \models Y$.

A collection of axioms X *syntactically* implies a theorem Y within a system S, if we can get to Y from X using transformation steps allowed within S. We write this as $X \vdash Y$.

The point of the system S known as "classical logic" is that its syntactic transformations preserve semantic implication, so that any syntactically allowed proof is semantically valid:

If $X \vdash Y$, then $X \models Y$.

If you make this idea be about proof steps in algebra doing things that always preserve the balance of a previously balanced scale, I see no reason why this idea couldn't be presented in eighth grade or earlier.

I can attest by spot-checking for small N that even most *mathematicians* have not been exposed to this idea. It's the standard concept in mathematical logic, but for some odd reason, the knowledge seems *constrained* to the study of "mathematical logic" as a separate field, which not all mathematicians are interested in (many just want to do Diophantine analysis or whatever).

So far as real life is concerned, mathematical logic only tells us the implications of what we already believe or suspect, but this is a computational problem of supreme difficulty and importance. After the first thousand times we observe that objects in Earth gravity accelerate downward at 9.8 m/s^2, we can suspect that this will be true on the next occasion - which is a matter of probabilistic induction, not valid logic. But then to go from that suspicion, *plus* the observation that a building is 45 meters tall, to a *specific* prediction of how long it takes the object to hit the ground, is a matter of logic - what will happen if everything we else we already believe, is actually true. It requires computation to make this conclusion transparent. We are not 'logically omniscient' - the technical term for the impossible dreamlike ability of knowing all the implications of everything you believe.

The great virtue of logic in *argument* is not that you can prove things by logic that are absolutely certain. Since logical implications are valid in every possible world, "observing" them never tells us *anything* about *which* possible world we live in. Logic can't tell you that you won't suddenly float up into the atmosphere. (What if we're in the Matrix, and the Matrix Lords decide to do that to you on a whim as soon as you finish reading this sentence?) Logic can only tell you that, if that *does* happen, you were wrong in your extremely strong suspicion about gravitation being always and everywhere true in our universe.

The great virtue of valid logic in *argument,* rather, is that logical argument exposes premises, so that anyone who disagrees with your conclusion has to (a) point out a premise they disagree with or (b) point out an invalid step in reasoning which is strongly liable to generate false statements from true statements.

For example: Nick Bostrom put forth the Simulation Argument, which is that *you must disagree with either statement (1) or (2) or else agree with statement (3):*

(1) Earth-originating intelligent life will, in the future, acquire vastly greater computing resources.

(2) Some of these computing resources will be used to run many simulations of ancient Earth, aka "ancestor simulations".

(3) We are almost certainly living in a computer simulation.

...but unfortunately it appears that not only do most respondents decline to say *why* they disbelieve in (3), most are unable to understand the distinction between the Simulation *Hypothesis* that we are living in a computer simulation, versus Nick Bostrom's actual support for the Simulation *Argument* that "You must either disagree with (1) or (2) or agree with (3)". They just treat Nick Bostrom as having claimed that we're all living in the Matrix. Really. Look at the media coverage.

I would seriously generalize that the mainstream media only understands the "and" connective, not the "or" or "implies" connective. I.e., it is impossible for the media to report on a discovery that one of two things must be true, or a discovery that *if* X is true then Y must be true (when it's not known that X is true). Also, the media only understands the "not" prefix when applied to atomic facts; it should go without saying that "not (A and B)" cannot be reported-on.

Robin Hanson sometimes complains that when he tries to argue that conclusion X follows from reasonable-sounding premises Y, his colleagues disagree with X while refusing to say which premise Y they think is false, or else say which step of the reasoning seems like an invalid implication. Such behavior is not only annoying, but logically rude, because someone else went out of their way and put in extra effort to make it *as easy as possible* for you to explain why you disagreed, and *you* couldn't be bothered to pick one item off a multiple-choice menu.

The inspiration of logic for argument is to lay out a modular debate, one which conveniently breaks into smaller pieces that can examined with smaller conversations. At least when it comes to trying to have a real conversation with a respected partner - I wouldn't necessarily advise a teenager to try it on their oblivious parents - that is the great inspiration we can take from the study of mathematical logic: *An argument is a succession of statements each allegedly following with high probability from previous statements or shared background knowledge.* Rather than, say, snowing someone under with as much fury and as many demands for applause as you can fit into sixty seconds.

Michael Vassar is fond of claiming that most people don't have the concept of an argument, and that it's pointless to try and teach them anything else until you can convey an intuitive sense for what it means to argue. I *think* that's what he's talking about.

Meditation: It has been claimed that logic and mathematics is the study of which conclusions follow from which premises. But when we say that 2 + 2 = 4, are we really just *assuming* that? It seems like 2 + 2 = 4 was true well before anyone was around to assume it, that two apples equaled two apples before there was anyone to count them, and that we couldn't make it 5 just by assuming differently.

Mainstream status.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Logical Pinpointing"

Previous post: "Causal Reference"

Logical Pinpointing

Followup to: Causal Reference, Proofs, Implications and Models

> The fact that one apple added to one apple invariably gives two apples helps in the teaching of arithmetic, but has no bearing on the truth of the proposition that 1 + 1 = 2.
>
> -- James R. Newman, The World of Mathematics

Previous meditation 1: If we can only meaningfully talk about parts of the universe that can be pinned down by chains of cause and effect, where do we find the fact that 2 + 2 = 4? Or did I just make a meaningless noise, there? Or if you claim that "2 + 2 = 4"*isn't* meaningful or true, then what alternate property does the sentence "2 + 2 = 4" have which makes it so much more useful than the sentence "2 + 2 = 3"?

Previous meditation 2: It has been claimed that logic and mathematics is the study of which conclusions follow from which premises. But when we say that 2 + 2 = 4, are we really just *assuming* that? It seems like 2 + 2 = 4 was true well before anyone was around to assume it, that two apples equalled two apples before there was anyone to count them, and that we couldn't make it 5 just by assuming differently.

Speaking conventional English, we'd say the sentence 2 + 2 = 4 is "true", and anyone who put down "false" instead on a math-test would be marked wrong by the schoolteacher (and not without justice).

But what can *make* such a belief true, what is the belief *about,* what is the truth-condition of the belief which can make it true or alternatively false? The sentence '2 + 2 = 4' is true if and only if... what?

In the previous post I asserted that the study of logic is the study of which conclusions follow from which premises; and that although this sort of inevitable implication is sometimes called "true", it could more specifically be called "valid", since checking for inevitability seems quite different from comparing a belief to our own universe. And you could claim, accordingly, that "2 + 2 = 4" is 'valid' because it is an inevitable implication of the axioms of Peano Arithmetic.

And yet thinking about 2 + 2 = 4 doesn't really *feel* that way. Figuring out facts about the natural numbers doesn't feel like the operation of making up assumptions and then deducing conclusions from them. It feels like the numbers are just *out* there, and the only point of making up the axioms of Peano Arithmetic was to *allow* mathematicians to talk about them. The Peano axioms might have been convenient for *deducing* a set of theorems like 2 + 2 = 4, but really all of those theorems were true *about* numbers to begin with. Just like "The sky is blue" is true about the sky, regardless of whether it follows from any particular assumptions.

So comparison-to-a-standard does seem to be at work, just as with *physical* truth... and yet this notion of 2 + 2 = 4 seems different from "stuff that makes stuff happen". Numbers don't occupy space or time, they don't arrive in any order of cause and effect, there are no *events* in numberland.

Meditation: What are we talking *about* when we talk about numbers? We can't navigate to them by following causal connections - so how do we get there from here?

...
...
...

"Well," says the mathematical logician, "that's indeed a very important and interesting question - where are the numbers - but first, I have a question for you. *What* are these 'numbers' that you're talking about? I don't believe I've heard that word before."

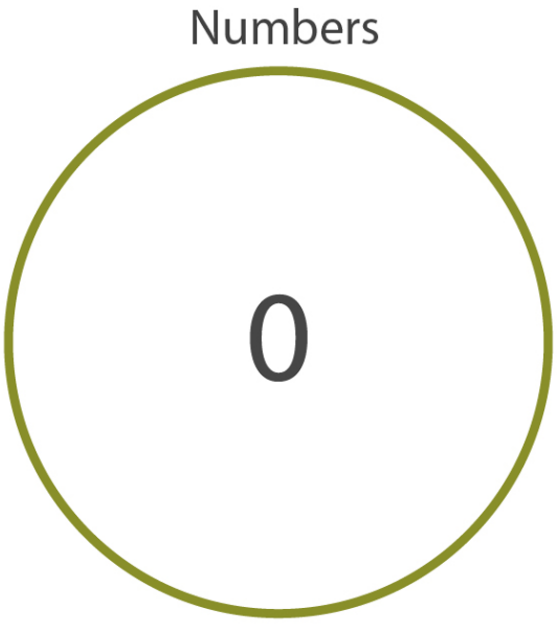
Yes you have.

"No, I haven't. I'm not a typical mathematical logician; I was just created five minutes ago for the purposes of this conversation. So I genuinely don't know what numbers are."

But... you know, 0, 1, 2, 3...

"I don't recognize that 0 thingy - what is it? I'm not asking you to give an exact definition, I'm just trying to figure out what the heck you're talking about in the first place."

Um... okay... look, can I start by asking you to just take on faith that there are these thingies called 'numbers' and 0 is one of them?



"Of course! 0 is a number. I'm happy to believe that. Just to check that I understand correctly, that does mean there exists a number, right?"

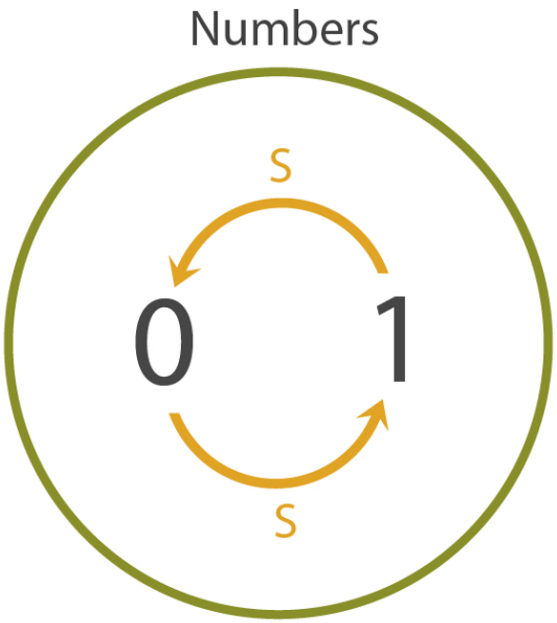
Um, yes. And then I'll ask you to believe that we can take the successor of any number. So we can talk about the successor of 0, the successor of the successor of 0, and so on. Now 1 is the successor of 0, 2 is the successor of 1, 3 is the successor of 2, and so on indefinitely, because we can take the successor of any number -

"In other words, the successor of any number is also a number."

Exactly.

"And in a simple case - I'm just trying to visualize how things might work - we would have 2 equal to 0."

What? No, why would that be -



"I was visualizing a case where there were two numbers that were the successors of each other, so SS0 = 0. I mean, I could've visualized one number that was the successor of itself, but I didn't want to make things *too* trivial -"

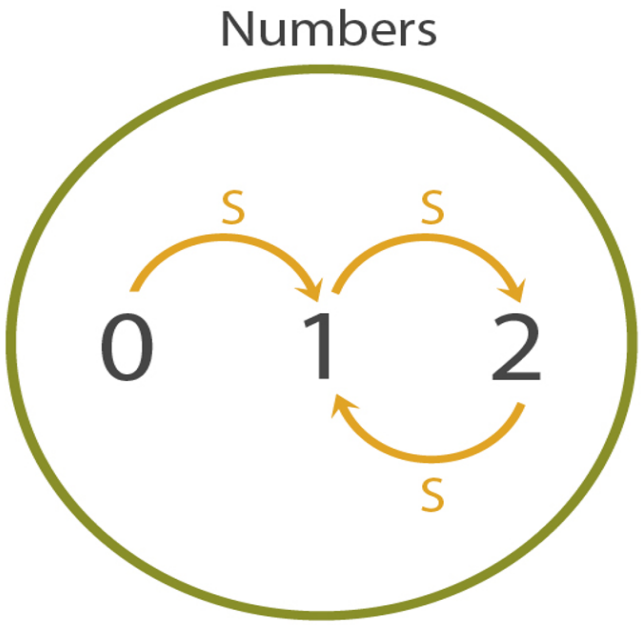
No! That model you just drew - that's *not* a model of the numbers.

"Why not? I mean, what property do the numbers have that this model doesn't?"

Because, um... zero is not the successor of *any* number. Your model has a successor link from 1 to 0, and that's not allowed.

"I see! So we can't have SS0=0. But we could still have SSS0=S0."

What? How -



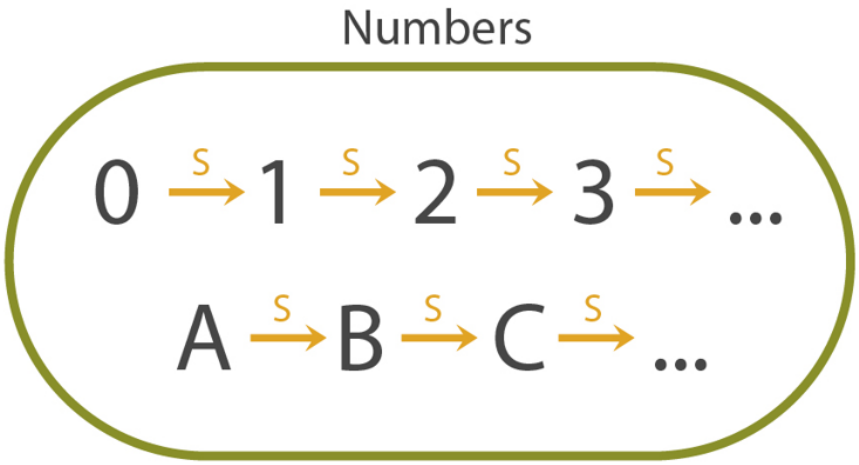
No! Because -

(consults textbook)

- if two numbers have the same successor, they are the same number, that's why! You can't have 2 and 0 *both* having 1 as a successor unless they're the same number, and if 2 was the same number as 0, then 1's successor would be 0, and that's not allowed! Because 0 is not the successor of any number!

"I see. Oh, wow, there's an awful lot of numbers, then. The first chain goes on *forever*."

It sounds like you're starting to get what I - wait. Hold on. What do you mean, the *first* chain -

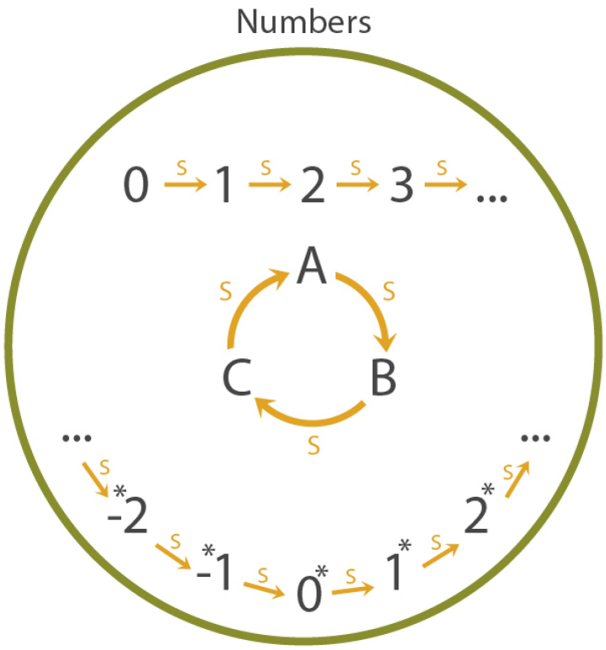


"I mean, you said that there was at least one start of an infinite chain, called 0, but -"

I misspoke. Zero is the *only* number which is not the successor of any number.

"I see, so any other chains would either have to loop or go on forever in *both* directions."

Wha?



"You said that zero is the only number which is not the successor of any number, that the successor of every number is a number, and that if two numbers have the same successor they are the same number. So, following those rules, any successor-chains besides the one that start at 0 have to loop or go on forever in both directions -"

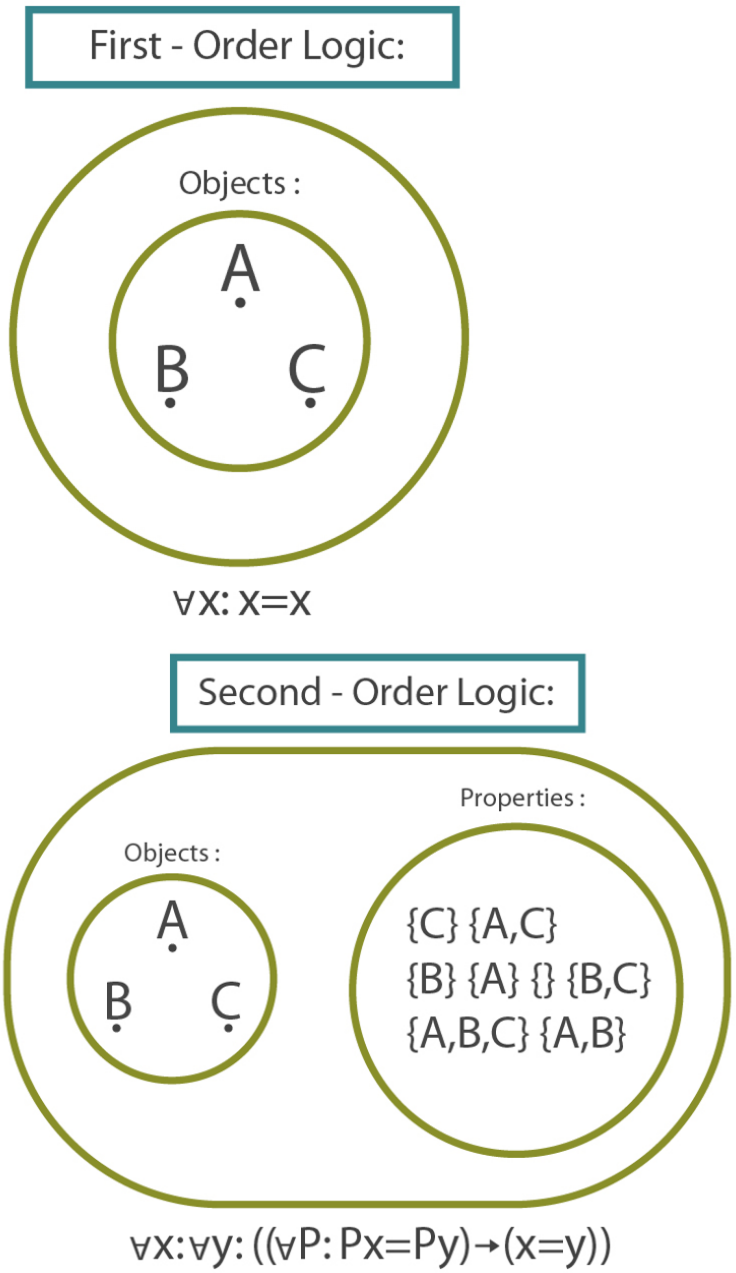
There *aren't supposed to be any chains* besides the one that starts at 0! Argh! And now you're going to ask me how to say that there shouldn't be any other chains, and I'm not a mathematician so I can't figure out exactly how to -

"Hold on! Calm down. *I'm* a mathematician, after all, so I can help you out. Like I said, I'm not trying to torment you here, just understand what you *mean*. You're right that it's not trivial to formalize your statement that there's only one successor-chain in the model. In fact, you can't say that *at all* inside what's called *first-order logic.* You have to jump to something called *second-order logic* that has some remarkably different properties (ha ha!) and make the statement there."

What the heck is second-order logic?

"It's the logic of properties! First-order logic lets you quantify over *all objects* - you can say that all objects are red, or all objects are blue, or '∀x: red(x)→¬blue(x)', and so on. Now, that 'red' and 'blue' we were just talking about - those are *properties,* functions which, applied to any object, yield either 'true' or 'false'. A property divides all objects into two classes, a class inside the property and a complementary class outside the property. So everything in the universe is either blue or not-blue, red or not-red, and

so on. And then second-order logic lets you quantify over properties - instead of looking at particular objects and asking whether they're blue or red, we can talk *about* properties in general - quantify over *all possible* ways of sorting the objects in the universe into classes. We can say, 'For all properties P', not just, 'For all objects X'."

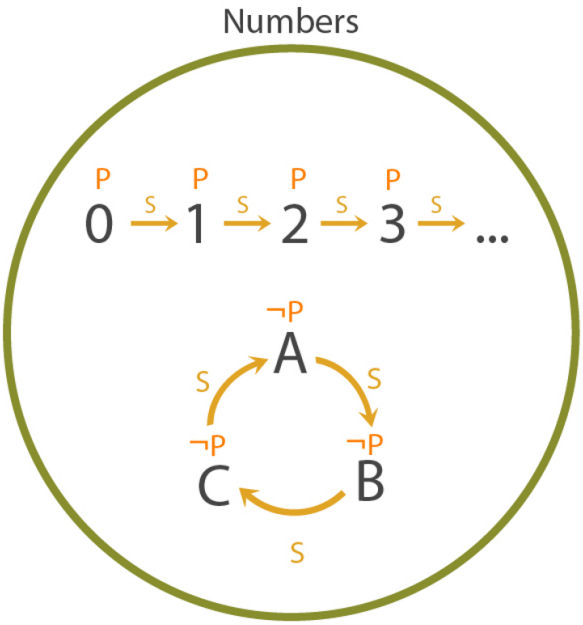


Okay, but what does that have to do with saying that there's only one chain of successors?

"To say that there's only one chain, you have to make the jump to second-order logic, and say that *for all properties P*, if P being true of a number implies P being true of the successor of that number, *and* P is true of 0, *then* P is true of all numbers."

Um... huh. That does sound reminiscent of something I remember hearing about Peano Arithmetic. But how does that solve the problem with chains of successors?

"Because if you had another *separated* chain, you could have a property P that was true all along the 0-chain, but false along the separated chain. And then P would be true of 0, true of the successor of any number of which it was true, and *not* true of all numbers."



I... huh. That's pretty neat, actually. You thought of that pretty fast, for somebody who's never heard of numbers.

"Thank you! I'm an imaginary fictionalized representation of a very *fast* mathematical reasoner."

Anyway, the next thing I want to talk about is addition. First, suppose that for every x, x + 0 = x. Next suppose that if x + y = z, then x + Sy = Sz -

"There's no need for that. We're done."

What do you mean, we're done?

"Every number has a successor. If two numbers have the same successor, they are the same number. There's a number 0, which is the only number that is not the successor of any other number. And every property true at 0, and for which P(Sx) is true whenever P(x) is true, is true of all numbers. In combination, those premises narrow down a *single* model in mathematical space, up to isomorphism. If you show me two models matching these requirements, I can perfectly map the objects and successor relations in them. You can't add any new object to the model, or subtract an object, without violating the axioms you've already given me. It's a uniquely identified

mathematical collection, the objects and their structure *completely pinned down*. Ergo, there's no point in adding any more requirements. Any meaningful statement you can make about these 'numbers', as you've defined them, is *already true* or *already false* within that pinpointed model - its truth-value is already semantically implied by the axioms you used to talk about 'numbers' as opposed to something else. If the new axiom is already true, adding it won't change what the previous axioms *semantically* imply."

Whoa. But don't I have to define the + operation before I can talk about it?

"Not in second-order logic, which can quantify over relations as well as properties. You just say: 'For every relation R that works exactly like addition, the following statement Q is true about that relation.' It would look like, '∀ relations R: (∀x∀y∀z: (R(x, 0, z)↔(x=z)) ∧ (R(x, Sy, z)↔R(Sx, y, z))) → Q)', where Q says whatever you meant to say about +, using the token R. Oh, sure, it's more convenient to add + to the language, but that's a mere *convenience* - it doesn't change which facts you can prove. Or to say it outside the system: So long as I *know* what numbers are, you can just explain to me how to add them; that doesn't change which mathematical structure we're already talking about."

...Gosh. I think I see the idea now. It's not that 'axioms' are mathematicians asking for you to just assume some things about numbers that seem obvious but can't be proven. Rather, axioms *pin down that we're talking about numbers as opposed to something else.*

"Exactly. That's why the *mathematical* study of numbers is *equivalent* to the *logical* study of which conclusions follow inevitably from the number-axioms. When you formalize logic into syntax, and prove theorems like '2 + 2 = 4' by syntactically deriving new sentences from the axioms, you can safely infer that 2 + 2 = 4 is semantically implied within the mathematical universe that the axioms pin down. And there's no way to try to 'just study the numbers without assuming any axioms', because those axioms are how you can talk about *numbers* as opposed to something else. You can't take for granted that just because your mouth makes a sound 'NUM-burz', it's a meaningful sound. The axioms aren't things you're arbitrarily making up, or assuming for convenience-of-proof, about some pre-existent thing called numbers. You need axioms to pin down a mathematical universe before you can talk *about* it in the first place. The axioms are pinning down what the heck this 'NUM-burz' sound means in the first place - that your mouth is talking about 0, 1, 2, 3, and so on."

Could you also talk about unicorns that way?

"I suppose. Unicorns don't exist in reality - there's nothing in the world that behaves like that - but they could nonetheless be described using a consistent set of axioms, so that it would be *valid* if not quite *true* to say that if a unicorn would be attracted to Bob, then Bob must be a virgin. Some people might dispute whether unicorns *must* be attracted to virgins, but since unicorns aren't real - since we aren't locating them within our universe using a causal reference - they'd just be talking about different models, rather than arguing about the properties of a known, fixed mathematical model. The 'axioms' aren't making questionable guesses about some real physical unicorn, or even a mathematical unicorn-model that's already been pinpointed; they're just fictional premises that make the word 'unicorn' talk about something inside a story."

But when I put two apples into a bowl, and then put in another two apples, I get four apples back out, regardless of anything I assume or don't assume. I don't need any axioms at all to get four apples back out.

"Well, you do need axioms to talk about *four,* SSSS0, when you say that you got 'four' apples back out. That said, indeed your experienced outcome - what your eyes see - doesn't depend on what axioms you assume. But that's because the apples are behaving like numbers whether you believe in numbers or not!"

The apples are behaving like numbers? What do you mean? I thought numbers were this ethereal mathematical model that got pinpointed by axioms, not by looking at the real world.

"Whenever a part of reality behaves in a way that conforms to the number-axioms - for example, if putting apples into a bowl obeys rules, like no apple spontaneously appearing or vanishing, which yields the high-level behavior of numbers - then all the mathematical theorems we proved valid in the universe of numbers can be imported back into reality. The conclusion isn't absolutely certain, because it's not absolutely certain that nobody will sneak in and steal an apple and change the physical bowl's behavior so that it doesn't match the axioms any more. But so long as the premises are true, the conclusions are true; the conclusion can't fail unless a premise also failed. You get four apples in reality, because those apples *behaving numerically* isn't something you *assume,* it's something that's *physically true.* When two clouds collide and form a bigger cloud, on the other hand, they aren't behaving like integers, whether you assume they are or not."

But if the awesome hidden power of mathematical reasoning is to be imported into parts of reality that behave like math, why not reason about apples in the first place instead of these ethereal 'numbers'?

"Because you can prove once and for all that *in any process which behaves like integers,* 2 thingies + 2 thingies = 4 thingies. You can store this general fact, and recall the resulting prediction, for *many* different places inside reality where physical things behave in accordance with the number-axioms. Moreover, so long as we believe that a calculator behaves like numbers, pressing '2 + 2' on a calculator and getting '4' tells us that 2 + 2 = 4 is true of numbers and then to expect four apples in the bowl. It's not like anything fundamentally different from that is going on when we try to add 2 + 2 inside our own *brains* - all the information we get about these 'logical models' is coming from the observation of physical things that allegedly behave like their axioms, whether it's our neurally-patterned thought processes, or a calculator, or apples in a bowl."

I... think I need to consider this for a while.

"Be my guest! Oh, and if you run out of things to think about from what I've said already -"

Hold on.

"- try pondering this one. Why does 2 + 2 come out the same way each time? Never mind the question of why the *laws of physics* are stable - why is *logic* stable? Of course I can't *imagine* it being any other way, but that's not an explanation."

Are you sure you didn't just degenerate into talking bloody nonsense?

"Of *course* it's bloody nonsense. If I knew a way to think about the question that wasn't bloody nonsense, I would already know the answer."

Meditation for next time:

> HUMANS NEED FANTASY TO BE HUMAN.
>
> "Tooth fairies? Hogfathers? Little—"
>
> YES. AS PRACTICE. YOU HAVE TO START OUT LEARNING TO BELIEVE THE *LITTLE* LIES.
>
> "So we can believe the big ones?"
>
> YES. JUSTICE. MERCY. DUTY. THAT SORT OF THING.
>
> "They're not the same at all!"
>
> YOU THINK SO? THEN TAKE THE UNIVERSE AND GRIND IT DOWN TO THE FINEST POWDER AND SIEVE IT THROUGH THE FINEST SIEVE AND THEN *SHOW* ME ONE ATOM OF JUSTICE, ONE MOLECULE OF MERCY.
>
> - Susan and Death, in *Hogfather* by Terry Pratchett

So far we've talked about two kinds of meaningfulness and two ways that sentences can refer; a way of comparing to physical things found by following pinned-down causal links, and logical reference by comparison to models pinned-down by axioms. Is there anything else that can be meaningfully talked about? Where would you find justice, or mercy?

Mainstream status.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Causal Universes"

Previous post: "Proofs, Implications, and Models"

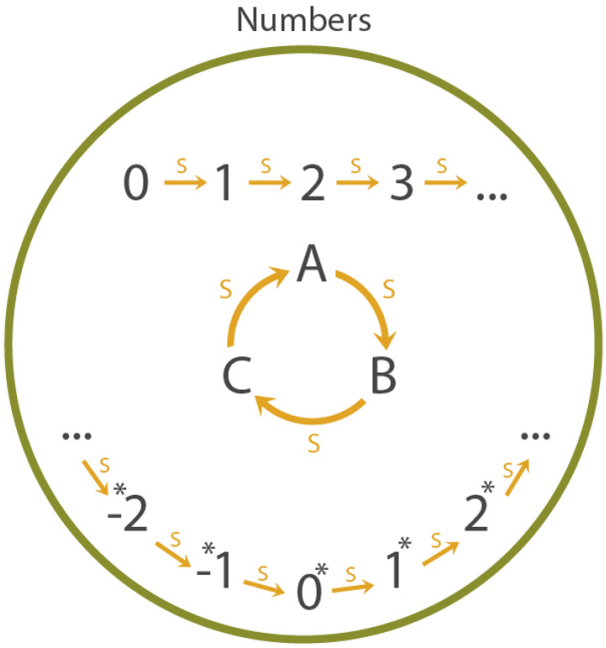
Standard and Nonstandard Numbers

Followup to: Logical Pinpointing

"Oh! Hello. Back again?"

Yes, I've got another question. Earlier you said that you *had* to use second-order logic to define the numbers. But I'm pretty sure I've heard about something called 'first-order Peano arithmetic' which is also supposed to define the natural numbers. Going by the name, I doubt it has any 'second-order' axioms. Honestly, I'm not sure I understand this second-order business at all.

"Well, let's start by examining the following model:"



"This model has three properties that we would expect to be true of the standard numbers - 'Every number has a successor', 'If two numbers have the same successor they are the same number', and '0 is the only number which is not the successor of any number'. All three of these statements are true in this model, so in that sense it's quite numberlike -"

And yet this model clearly is *not* the numbers we are looking for, because it's got all these mysterious extra numbers like C and -2*. That C thing even loops around, which I certainly wouldn't expect any number to do. And then there's that infinite-in-both-directions chain which isn't corrected to anything else.

"Right, so, the difference between first-order logic and second-order logic is this: In first-order logic, we can get rid of the ABC - make a statement which *rules out* any model that has a loop of numbers like that. But we can't get rid of the infinite chain underneath it. In second-order logic we can get rid of the extra chain."

I would ask you to explain why that was true, but at this point I don't even know what second-order logic *is.*

"Bear with me. First, consider that the following formula *detects 2-ness:*"

$$x + 2 = x * 2$$

In other words, that's a formula which is true when x is equal to 2, and false everywhere else, so it singles out 2?

"Exactly. And this is a formula which detects odd numbers:"

$$\exists y: x=(2*y)+1$$

Um... okay. That formula says, 'There exists a y, such that x equals 2 times y plus one.' And that's true when x is 1, because 0 is a number, and 1=(2*0)+1. And it's true when x is 9, because there exists a number 4 such that 9=(2*4)+1... right. The formula is true at all odd numbers, and only odd numbers.

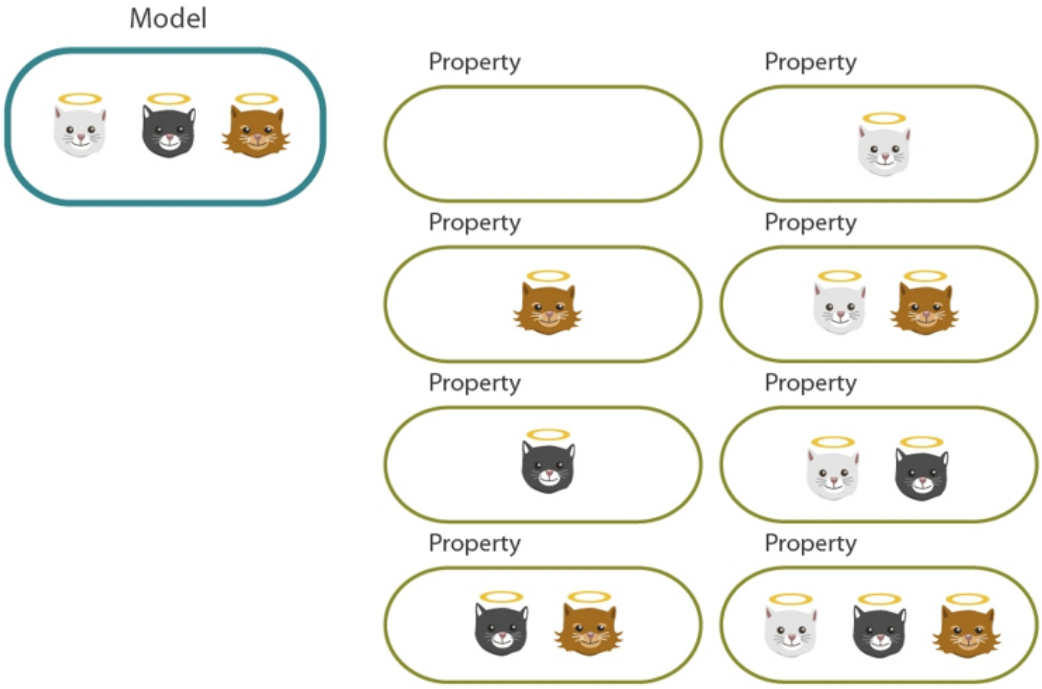
"Indeed. Now suppose we had some way to *detect the existence* of that ABC-loop in the model - a formula which was *true* at the ABC-loop and *false* everywhere else. Then I could adapt the *negation* of this statement to say 'No objects like this are allowed to exist', and add that as an axiom alongside 'Every number has a successor' and so on. Then I'd have *narrowed down* the possible set of models to get rid of models that have an extra ABC-loop in them."

Um... can I rule out the ABC-loop by saying $\neg\exists x:(x=A)$?

"Er, only if you've told me what A is in the first place, and in a logic which has ruled out all models with loops in them, you shouldn't be able to point to a specific object that doesn't exist -"

Right. Okay... so the idea is to rule out loops of successors... hm. In the numbers 0, 1, 2, 3..., the number 0 isn't the successor of any number. If I just took a group of numbers starting at 1, like {1, 2, 3, ...}, then 1 wouldn't be the successor of any number *inside* that group. But in A, B, C, the number A is the successor of C, which is the successor of B, which is the successor of A. So how about if I say: 'There's no group of numbers G such that for any number x in G, x is the successor of some other number y in G.'

"Ah! Very clever. But it so happens that you just used second-order logic, because you talked about *groups* or *collections* of entities, whereas *first-order logic* only talks about *individual* entities. Like, suppose we had a logic talking about kittens and whether they're innocent. Here's a model of a universe containing exactly three distinct kittens who are all innocent:"



Er, what are those 'property' thingies?

"They're all possible collections of kittens. They're labeled *properties* because every collection of kittens corresponds to a property that some kittens have and some kittens don't. For example, the collection on the top right, which contains only the grey kitten, corresponds to a predicate which is true at the grey kitten and false everywhere else, or to a property which the grey kitten has which no other kitten has. Actually, for now let's just pretend that 'property' just says 'collection'."

Okay. I understand the concept of a collection of kittens.

"In first-order logic, we can talk about individual kittens, and how they relate to other individual kittens, and whether or not any kitten bearing a certain relation exists or doesn't exist. For example, we can talk about how the grey kitten adores the brown kitten. In second-order logic, we can talk about collections of kittens, and whether or not those collections exist. So in first-order logic, I can say, 'There exists a kitten which is innocent', or 'For every individual kitten, that kitten is innocent', or 'For every individual kitten, there exists another individual kitten which adores the first kitten.' But it requires second-order logic to make statements about *collections* of kittens, like, 'There exists no collection of kittens such that every kitten in it is adored by some other kitten inside the collection.'"

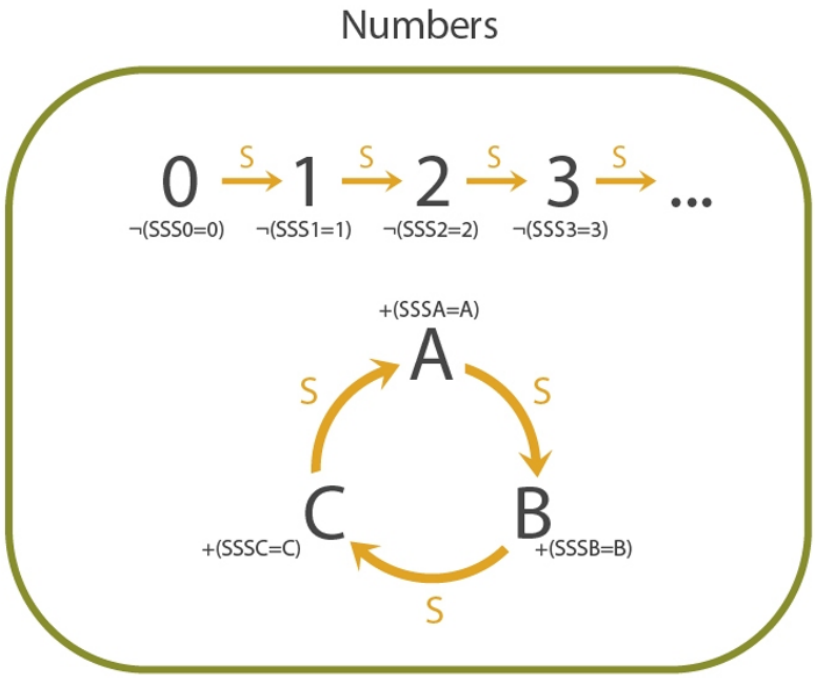
I see. So when I tried to say that you couldn't have any group of numbers, such that every number in the group was a successor of some other number in the group...

"...you quantified over the existence or nonexistence of *collections* of numbers, which means you were using *second-order logic.* However, in this particular case, it's easily possible to rule out the ABC-loop of numbers using only first-order logic. Consider the formula:"

x=SSSx

x plus 3 is equal to itself?

"Right. That's a first-order formula, since it doesn't talk about collections. And that formula is false at 0, 1, 2, 3... but true at A, B, and C."



What does the '+' mean?

"Er, by '+' I was trying to say, 'this formula works out to True' and similarly '¬' was supposed to mean the formula works out to False. The general idea is that we now have a formula for detecting 3-loops, and distinguishing them from *standard* numbers like 0, 1, 2 and so on."

I see. So by adding the new axiom, ¬∃x:x=SSSx, we could rule out all the models containing A, B, and C or any other 3-loop of nonstandard numbers.

"Right."

But this seems like a rather arbitrary sort of axiom to add to a fundamental theory of arithmetic. I mean, I've never seen any attempt to describe the numbers which says, 'No number is equal to itself plus 3' as a basic premise. It seems like it should be a theorem, not an axiom.

"That's because it's brought in using a more general rule. In particular, first-order arithmetic has an *infinite axiom schema* - an infinite but computable scheme of axioms. Each axiom in the schema says, for a different first-order formula Φ(x) - pronounced 'phi of x' - that:"

1. *If* Φ is true at 0, i.e: Φ(0)
2. *And if* Φ is true of the successor of any number where it's true, i.e:
   ∀x: Φ(x)→Φ(Sx)
3. *Then* Φ is true of all numbers: ∀n: Φ(n)

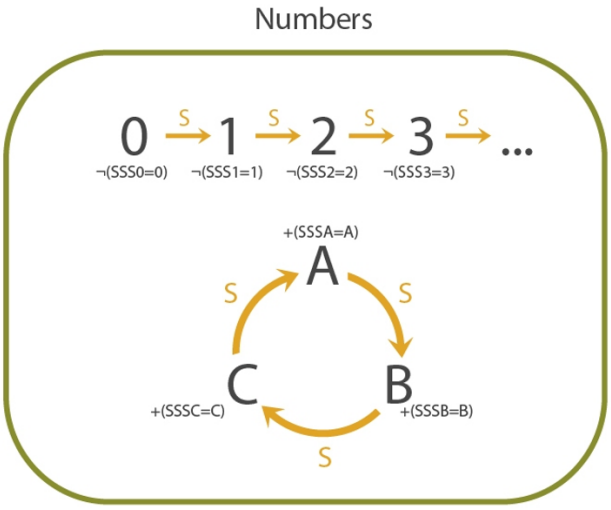
$(\Phi(0) \wedge (\forall x: \Phi(x) \rightarrow \Phi(Sx))) \rightarrow (\forall n: \Phi(n))$

"In other words, every *formula* which is true at 0, and which is true of the successor of any number of which it is true, is true *everywhere.* This is the *induction schema* of first-order arithmetic. As a special case we have the *particular* inductive axiom:"

$(0 \neq SSS0 \wedge (\forall x: (x \neq SSSx) \rightarrow (Sx \neq SSSSx)) \rightarrow (\forall n: n \neq SSSn)$

But that doesn't say that for all n, n≠n+3. It gives some premises from which that conclusion would follow, but we don't know the premises.

"Ah, however, we can *prove* those premises using the *other* axioms of arithmetic, and hence prove the conclusion. The formula (SSSx=x) is false at 0, because 0 is not the successor of *any* number, including SS0. Similarly, consider the formula SSSSx=Sx, which we can rearrange as S(SSSx)=S(x). If two numbers have the same successor they are the same number, so SSSx=x. If truth at Sx proves truth at x, then falsity at x proves falsity at Sx, modus ponens to modus tollens. Thus the formula is false at zero, false of the successor of any number where it's false, and so must be false everywhere under the induction axiom schema of first-order arithmetic. And so first-order arithmetic can rule out models like this:"

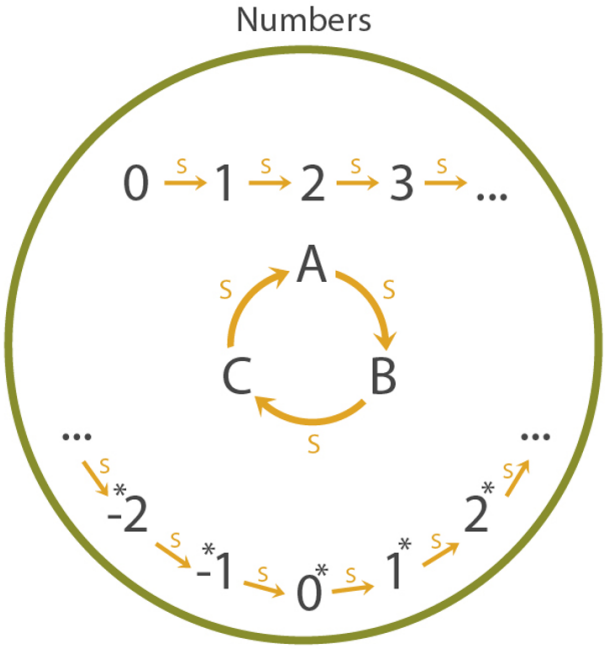


...er, I think I see? Because if this model obeys all the *other* axioms which which we *already* specified, that *didn't* filter it out earlier - axioms like 'zero is not the successor of any number' and 'if two numbers have the same successor they are the same number' - then we can *prove* that the formula x≠SSSx is true at 0, and prove that if the formula true at x it must be true at x+1. So once we then add the *further* axiom that *if* x≠SSSx is true at 0, and *if* x≠SSSx is true at Sy when it's true at y, *then* x≠SSSx is true at all x...

"We already have the premises, so we get the conclusion. ∀x: x≠SSSx, and thus we filter out all the 3-loops. Similar logic rules out N-loops for all N."

So then did we get rid of all the nonstandard numbers, and leave only the standard model?

"No. Because there was also that problem with the infinite chain ... -2*, -1*, 0*, 1* and so on."



Here's one idea for getting rid of the model with an infinite chain. All the nonstandard numbers in the chain are "greater" than all the standard numbers, right? Like, if *w* is a nonstandard number, then $w > 3$, $w > 4$, and so on?

"Well, we can prove by induction that no number is less than 0, and *w* isn't equal to 0 or 1 or 2 or 3, so I'd have to agree with that."

Okay. We should also be able to prove that if x > y then x + z > y + z. So if we take nonstandard *w* and ask about $w + w$, then $w + w$ must be greater than $w + 3$, $w + 4$, and so on. So $w + w$ can't be part of the infinite chain at all, and yet adding any two numbers ought to yield a third number.

"Indeed, that does prove that if there's one infinite chain, there must be *two* infinite chains. In other words, that original, exact model in the picture, can't all by itself be a model of first-order arithmetic. But showing that the chain implies the existence of yet other elements, isn't the same as proving that the chain doesn't exist. Similarly, since all numbers are even or odd, we must be able to find *v* with $v + v = w$, or find *v* with $v + v + 1 = w$. Then *v* must be part of another nonstandard chain that comes before the chain containing *w*."

But then that requires an *infinite* number of infinite chains of nonstandard numbers which are all greater than any standard number. Maybe we can extend this logic to eventually reach a contradiction and rule out the existence of an infinite chain in the first place - like, we'd show that any complete collection of nonstandard numbers has to be *larger than itself* -

"Good idea, but no. You end up with the conclusion that if a single nonstandard number exists, it must be part of a chain that's infinite in both directions, i.e., a chain that looks like an ordered copy of the negative and positive integers. And that if an infinite chain exists, there must be infinite chains corresponding to all *rational numbers.* So something that could actually be a nonstandard model of first-order arithmetic, has to contain at least the standard numbers *followed by* a copy of the

rational numbers with each rational number replaced by a copy of the integers. But then *that* setup works just fine with both addition and multiplication - we can't prove that it has to be any larger than what we've already said."

Okay, so how *do* we get rid of an infinite number of infinite chains of nonstandard numbers, and leave just the standard numbers at the beginning? What kind of statement would they violate - what sort of axiom would rule out all those extra numbers?

"We have to use second-order logic for that one."

Honestly I'm still not 100% clear on the difference.

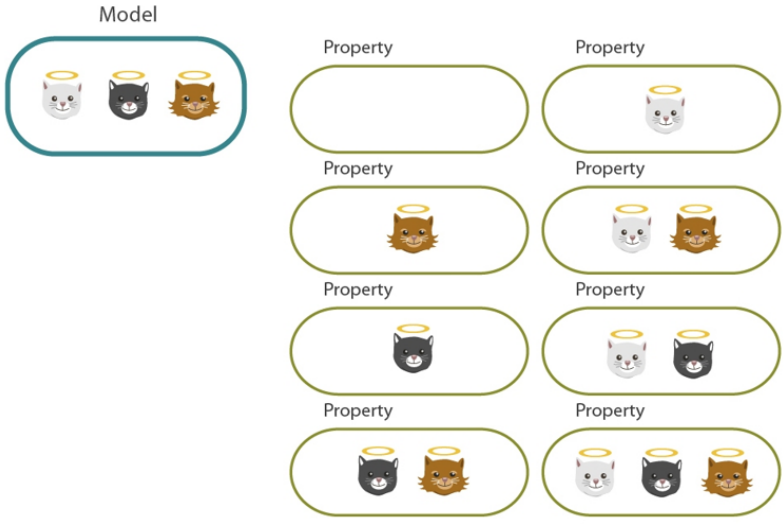
"Okay... earlier you gave me a *formula* which detected odd numbers."

Right. ∃y: x=(2*y)+1, which was true at x=1, x=9 and so on, but not at x=0, x=4 and so on.

"When you think in terms of *collections of numbers,* well, there's *some* collections which can be defined by formulas. For example, the collection of odd numbers {1, 3, 5, 7, 9, ...} can be defined by the formula, with x free, ∃y: x=(2*y)+1. But you could also try to talk about just the collection {1, 3, 5, 7, 9, ...} as a collection, a set of numbers, whether or not there happened to be any formula that defined it -"

Hold on, how can you talk about a set if you can't define a formula that makes something a member or a non-member? I mean, that seems a bit smelly from a rationalist perspective -

"Er... remember the earlier conversation about kittens?"



"Suppose you say something like, 'There *exists a collection* of kittens, such that every kitten adores only other kittens in the collection'. Give me a room full of kittens, and I can count through all possible collections, check your statement for each collection, and see whether or not there's a collection which is actually like that. So the statement is meaningful - it can be falsified or verified, and it constrains the state of reality. But you didn't give me a *local formula* for picking up a *single* kitten and deciding whether or not it ought to be in this mysterious collection. I had to iterate

through all the *collections* of kittens, find the *collections* that matched your statement, and only then could I decide whether any individual kitten had the property of being in a collection like that. But the statement was still falsifiable, even though it was, in mathematical parlance, *impredicative* - that's what we call it when you make a statement that can only be verified by looking at many possible collections, and doesn't start from any particular collection that you tell me how to construct."

Ah... hm. What about infinite universes of kittens, so you can't iterate through all possible collections in finite time?

"If you say, 'There exists a collection of kittens which all adore each other', I could exhibit a group of three kittens which adored each other, and so prove the statement true. If you say 'There's a collection of four kittens who adore only each other', I might come up with a constructive proof, given the other known properties of kittens, that your statement was false; and any time you tried giving me a group of four kittens, I could find a fifth kitten, adored by some kitten in your group, that falsified your attempt. But this is getting us into some rather deep parts of math we should probably stay out of for now. The point is that even in infinite universes, there are second-order statements that you can prove or falsify in finite amounts of time. And once you admit those *particular* second-order statements are talking about something meaningful, well, you might as well just admit that second-order statements in general are meaningful."

...that sounds a little iffy to me, like we might get in trouble later on.

"You're not the only mathematician who worries about that."

But let's get back to numbers. You say that we can use second-order logic to rule out any infinite chain.

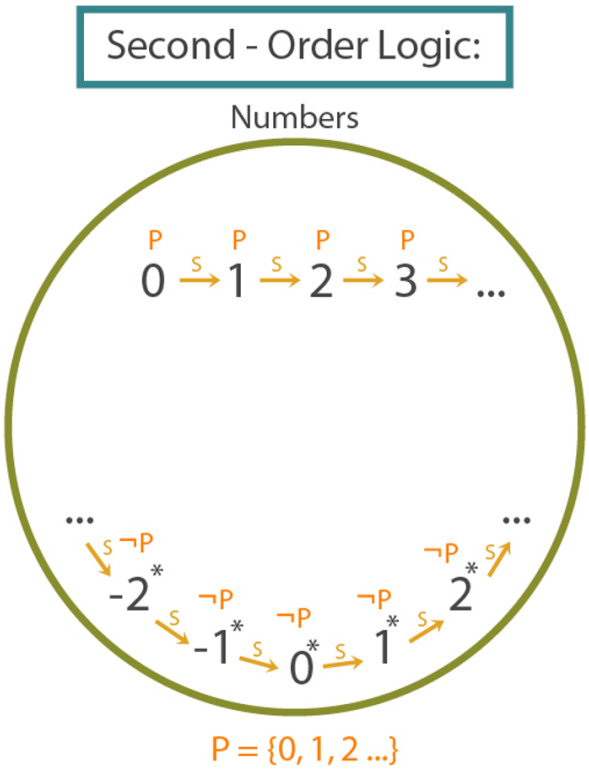
"Indeed. In second-order logic, instead of using an infinite axiom schema over all formulas Φ, we quantify over *possible collections* directly, and say, in a *single* statement:"

$$\forall P: P(0) \wedge (\forall x: P(x) \rightarrow P(Sx)) \rightarrow (\forall n: P(n))$$

"Here P is any predicate true or false of individual numbers. Any collection of numbers corresponds to a predicate that is true of numbers inside the collection and false of numbers outside of it."

Okay... and how did that rule out infinite chains again?

"Because *in principle,* whether or not there's any first-order formula that picks them out, there's *theoretically* a collection that contains the standard numbers {0, 1, 2, ...} and *only* the standard numbers. And if you treat that collection as a predicate P, then P is true at 0 - that is, 0 is in the standard numbers. And if 200 is a standard number then so is 201, and so on; if P is true at x, it's true at x+1. On the other hand, if you treat the collection 'just the standard numbers' as a predicate, it's false at -2*, false at -1*, false at 0* and so on - those numbers *aren't* in this theoretical collection. So it's vacuously true that this predicate is true at 1* if it's true at 0*, because it's *not* true at 0*. And so we end up with:"



"And so the single second-order axiom..."

$$\forall P: P0 \wedge (\forall x: Px \rightarrow P(Sx)) \rightarrow (\forall n: Pn)$$

"...rules out any disconnected chains, finite loops, and indeed every nonstandard number, in one swell foop."

But what did that axiom *mean,* exactly? I mean, taboo the phrase 'standard numbers' for a moment, pretend I've got no idea what those are, just explain to me what the axiom actually *says*.

"It says that the model being discussed - the model which fits this axiom - makes it impossible to form *any collection closed under succession* which includes 0 and doesn't include *everything*. It's impossible to have *any collection of objects in this universe* such that 0 is in the collection, and the successor of everything in the collection is in the collection, and yet this collection doesn't contain *everything.* So you can't have a disconnected infinite chain - there would then exist at least one collection over objects in this universe that contained 0 and all its successor-descendants, yet didn't contain the chain; and we have a shiny new axiom which says that can't happen."

Can you perhaps operationalize that in a more sensorymotory sort of way? Like, if this is what I believe about the universe, then what do I expect to see?

"If this is what you believe about the mathematical model that you live in... then you believe that neither you, nor any adversary, nor yet a superintelligence, nor yet God, can consistently say 'Yea' or 'Nay' to objects in such fashion that when you present

them with 0, they say 'Yea', and when you present them with any other object, if they say 'Yea', they also say 'Yea' for the successor of that object; and yet there is some object for which they say 'Nay'. You believe this can never happen, no matter what. The way in which the objects in the universe are arranged by succession, just doesn't let that happen, ever."

Ah. So if, say, they said 'Nay' for 42, I'd go back and ask about 41, and then 40, and by the time I reached 0, I'd find either that they said 'Nay' about 0, or that they said 'Nay' for 41 and yet 'Yea' for 40. And what do I expect to see if I believe in first-order arithmetic, with the infinite axiom schema?

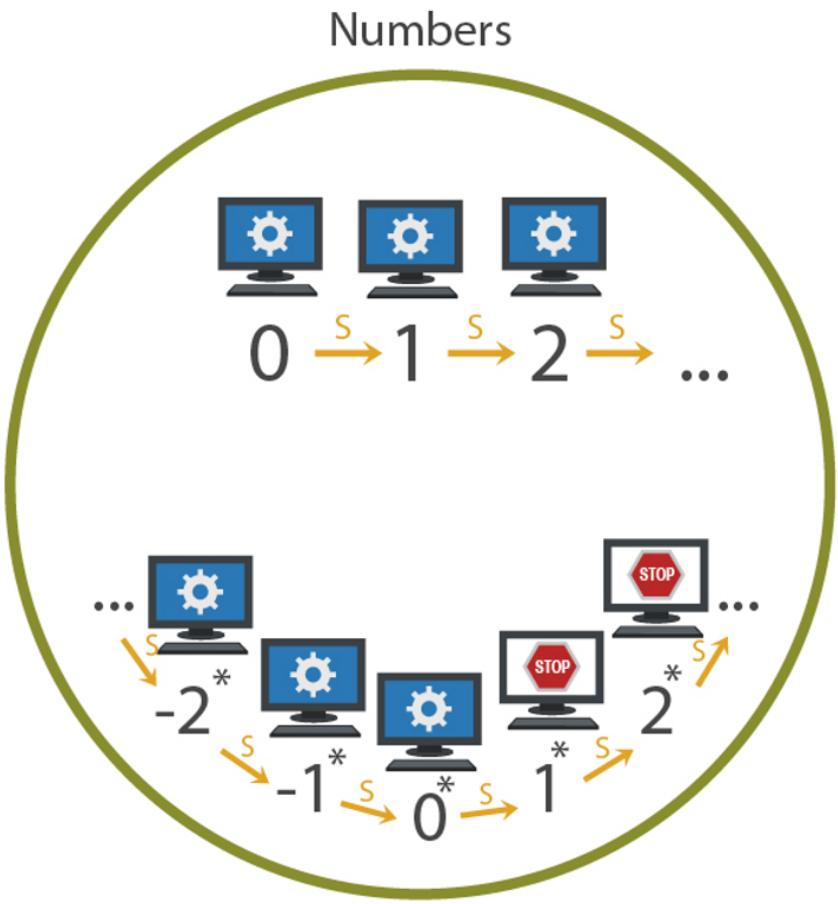
"In that case, you believe there's no neatly specifiable, compactly describable *rule* which behaves like that. But if you believe the second-order version, you believe nobody can possibly behave like that even if they're answering randomly, or branching the universe to answer different ways in different alternate universes, and so on. And note, by the way, that if we have a finite universe - i.e., we throw out the rule that *every* number has a successor, and say instead that 256 is the only number which has no successor - then we can verify this axiom in finite time."

I see. Still, is there any way to rule out infinite chains using *first*-order logic? I might find that easier to deal with, even if it looks more complicated at first.

"I'm afraid not. One way I like to look at it is that first-order logic can talk about *constraints on how the model looks from any local point*, while only second-order logic can talk about *global qualities* of chains, collections, and the model as a whole. Whether every number has a successor is a local property - a question of how the model looks from the vantage point of any one number. Whether a number plus three, can be equal to itself, is a question you could evaluate at the local vantage point of any one number. Whether a number is *even,* is a question you can answer by looking around for a single, individual number x with the property that x+x equals the first number. But when you try to say that there's *only one connected chain* starting at 0, by invoking the idea of *connectedness* and *chains* you're trying to describe non-local properties that require a logic-of-possible-collections to specify."

Huh. But if all the 'local' properties are the same regardless, why worry about global properties? In first-order arithmetic, any 'local' formula that's true at zero and all of its 'natural' successors would also have to be true of all the disconnected infinite chains... right? Or did I make an error there? All the other infinite chains besides the 0-chain - all 'nonstandard numbers' - would have just the same properties as the 'natural' numbers, right?

"I'm afraid not. The first-order axioms of arithmetic may fail to pin down whether or not a Turing machine halts - whether there *exists a time* at which a Turing machine halts. Let's say that from our perspective inside the standard numbers, the Turing machine 'really doesn't' halt - it doesn't halt on clock tick 0, doesn't halt on clock tick 1, doesn't halt on tick 2, and so on through all the standard successors of the 0-chain. In nonstandard models of the integers - models with other infinite chains - there might be somewhere inside a *nonstandard chain* where the Turing machine goes from running to halted and stays halted thereafter."



"In this new model - which is fully compatible with the first-order axioms, and can't be ruled out by them - it's not true that 'for every number t at which the Turing machine is running, it will still be running at t+1'. Even though if we could somehow restrict our attention to the 'natural' numbers, we would see that the Turing machine was running at 0, 1, 2, and every time in the successor-chain of 0."

Okay... I'm not quite sure what the *practical* implication of that is?

"It means that many Turing machines which *in fact* never halt at any standard time, can't be *proven not to halt* using first-order reasoning, because their non-halting-ness *does not actually follow logically* from the first-order axioms. Logic is about which conclusions follow from which premises, remember? If there are models which are compatible with all the first-order premises, but still falsify the statement 'X runs forever', then the statement 'X runs forever' can't *logically follow* from those premises. This means you won't be able to prove - *shouldn't* be able to prove - that this Turing machine halts, using *only* first-order logic."

How exactly would this fail in practice? I mean, where does the proof go bad?

"You wouldn't get the second step of the induction, 'for every number t at which the Turing machine is running, it will still be running at t+1'. There'd be nonstandard models with some nonstandard t that falsifies the premise - a nonstandard time where the Turing machine goes from running to halted. Even though if we could somehow

restrict our attention to *only the standard numbers*, we would see that the Turing machine was running at 0, 1, 2, and so on."

But if a Turing machine really actually halts, there's got to be some *particular time* when it halts, like on step 97 -

"Indeed. But 97 exists in *all* nonstandard models of arithmetic, so we can prove its existence in first-order logic. Any time 0 is a number, every number has a successor, numbers don't loop, and so on, there'll exist 97. Every nonstandard model has *at least* the standard numbers. So whenever a Turing machine *does* halt, you can prove in first-order arithmetic that it halts - it does indeed follow from the premises. That's kinda what you'd *expect,* given that you can just watch the Turing machine for 97 steps. When something actually does halt, you *should* be able to prove it halts without worrying about unbounded future times! It's when something *doesn't actually* halt - in the standard numbers, that is - that the existence of 'nonstandard halting times' becomes a problem. Then, the conclusion that the Turing machine runs forever *may not actually follow* from first-order arithmetic, because you can obey all the premises of first-order arithmetic, and yet still be inside a nonstandard model where this Turing machine halts at a nonstandard time."

So second-order arithmetic is more powerful than first-order arithmetic in terms of *what follows from the premises*?

"That follows inevitably from the ability to talk about *fewer possible models*. As it is written, 'What is true of one apple may not be true of another apple; thus more can be said about a single apple than about all the apples in the world.' If you can restrict your discourse to a narrower collection of models, there are more facts that follow inevitably, because the more models you might be talking about, the fewer facts can possibly be true about all of them. And it's also definitely true that second-order arithmetic proves more theorems than first-order arithmetic - for example, it can prove that a Turing machine which computes Goodstein sequences always reaches 0 and halts, or that Hercules always wins the hydra game. But there's a bit of controversy we'll get into later about whether second-order logic is *actually* more powerful than first-order logic in general."

Well, sure. After all, just because nobody has ever yet invented a first-order formula to filter out all the nonstandard numbers, doesn't mean it can never, ever be done. Tomorrow some brilliant mathematician might figure out a way to take an individual number x, and do local things to it using addition and multiplication and the existence or nonexistence of other individual numbers, which can tell us whether that number is part of the 0-chain or some other infinite-in-both-directions chain. It'll be as easy as (a=b*c) -

"Nope. Ain't never gonna happen."

But maybe you could find some entirely different creative way of first-order axiomatizing the numbers which has *only* the standard model -

"Nope."

Er... how do you *know* that, exactly? I mean, part of the Player Character Code is that you don't give up when something *seems* impossible. I can't quite see *yet* how to detect infinite chains using a first-order formula. But then earlier I didn't realize you could rule out finite loops, which turned out to be quite simple once you explained. After all, there's two distinct uses of the word 'impossible', one which indicates

positive knowledge that something can *never* be done, that no *possible* chain of actions can *ever* reach a goal, even if you're a superintelligence. This kind of knowledge requires a strong, definite grasp on the subject matter, so that you can rule out *every* possible avenue of success. And then there's another, *much more common* use of the word 'impossible', which means that you thought about it for five seconds but didn't see any way to do it, usually used in the presence of *weak* grasps on a subject, subjects that seem sacredly mysterious -

"Right. Ruling out an infinite-in-both-directions chain, using a first-order formula, is the *first* kind of impossibility. We *know* that it can never be done."

I see. Well then, what do you think you know, and how do you think you know it? How is this definite, positive knowledge of impossibility obtained, using your strong grasp on the non-mysterious subject matter?

"We'll take that up next time."

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Godel's Completeness and Incompleteness Theorems"

Previous post: "By Which It May Be Judged"

Godel's Completeness and Incompleteness Theorems

Followup to: Standard and Nonstandard Numbers

So... last time you claimed that using first-order axioms to rule out the existence of nonstandard numbers - other chains of numbers besides the 'standard' numbers starting at 0 - was *forever and truly impossible*, even unto a superintelligence, no matter *how* clever the first-order logic used, even if you came up with an entirely different way of axiomatizing the numbers.

"Right."

How could you, in your finiteness, possibly know that?

"Have you heard of Godel's Incompleteness Theorem?"

Of course! Godel's Theorem says that for every consistent mathematical system, there are statements which are *true* within that system, which can't be *proven* within the system itself. Godel came up with a way to encode theorems and proofs as numbers, and wrote a purely numerical formula to detect whether a proof obeyed proper logical syntax. The basic trick was to use prime factorization to encode lists; for example, the ordered list <3, 7, 1, 4> could be uniquely encoded as:

$2^3 * 3^7 * 5^1 * 7^4$

And since prime factorizations are unique, and prime powers don't mix, you could inspect this single number, 210,039,480, and get the unique ordered list <3, 7, 1, 4> back out. From there, going to an encoding for logical formulas was easy; for example, you could use the 2 prefix for NOT and the 3 prefix for AND and get, for any formulas Φ and Ψ encoded by the numbers $\#\Phi$ and $\#\Psi$:

$\neg\Phi = 2^2 * 3^{\#\Phi}$

$\Phi \wedge \Psi = 2^3 * 3^{\#\Phi} * 5^{\#\Psi}$

It was then possible, by dint of crazy amounts of work, for Godel to come up with a gigantic formula of Peano Arithmetic [](p, c) meaning, 'P encodes a valid logical proof using first-order Peano axioms of C', from which directly followed the formula []c, meaning, 'There exists a number P such that P encodes a proof of C' or just 'C is provable in Peano arithmetic.'

Godel then put in some *further* clever work to invent statements which referred to *themselves*, by having them contain sub-recipes that would reproduce the entire statement when manipulated by another formula.

And then Godel's Statement encodes the statement, 'There does not exist any number P such that P encodes a proof of (this statement) in Peano arithmetic' or in simpler terms 'I am not provable in Peano arithmetic'. If we assume first-order arithmetic is consistent and sound, then no *proof* of this statement *within* first-order arithmetic exists, which means the statement is *true* but can't be proven within the system. That's Godel's Theorem.

"Er... no."

No?

"No. I've heard rumors that Godel's Incompleteness Theorem is horribly misunderstood in your Everett branch. Have you heard of Godel's *Completeness* Theorem?"

Is that a thing?

"Yes! Godel's Completeness Theorem says that, for any collection of first-order statements, *every semantic implication of those statements is syntactically provable within first-order logic*. If something is a genuine implication of a collection of first-order statements - if it actually *does* follow, in the models pinned down by those statements - then you can *prove* it, *within* first-order logic, using *only* the syntactical rules of proof, from those axioms."

I don't see how that could possibly be true at the same time as Godel's Incompleteness Theorem. The Completeness Theorem and Incompleteness Theorem seem to say diametrically opposite things. Godel's Statement is implied by the axioms of first-order arithmetic - that is, we can see it's true using our own mathematical reasoning -

"Wrong."

What? I mean, I understand we can't prove it *within* Peano arithmetic, but from outside the system we can see that -

All right, explain.

"Basically, you just committed the equivalent of saying, 'If all kittens are little, and some little things are innocent, then some kittens are innocent.' There are universes -

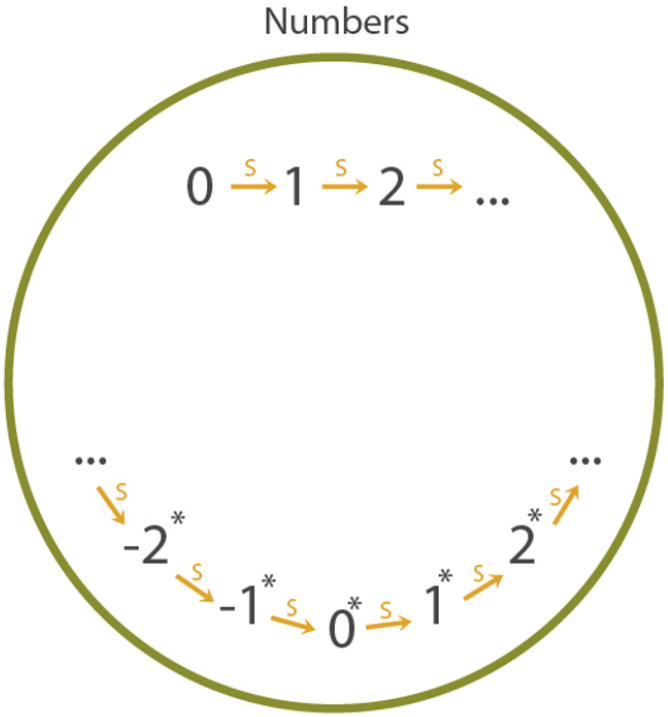
logical models - where it so happens that the premises are true and the conclusion also happens to be true:"

"But there are also valid models of the premises where the conclusion is false:"

"If you, yourself, happened to live in a universe like the first one - if, in your mind, you were *only thinking* about a universe like that - then you might *mistakenly* think that you'd proven the conclusion. But your statement is not *logically* valid, the conclusion is not true in *every* universe where the premises are true. It's like saying, 'All apples are plants. All fruits are plants. Therefore all apples are fruits.' Both the premises and the conclusions happen to be true in *this* universe, but it's not valid logic."

Okay, so how does this invalidate my previous explanation of Godel's Theorem?

"Because of the non-standard models of first-order arithmetic. First-order arithmetic narrows things down a lot - it rules out 3-loops of nonstandard numbers, for example, and mandates that every model contain the number 17 - but it doesn't pin down a *single* model. There's still the possibility of infinite-in-both-directions chains coming after the 'standard' chain that starts with 0. Maybe *you* have just the standard numbers in mind, but that's not the *only* possible model of first-order arithmetic."



So?

"So in some of those other models, there are nonstandard numbers which - according to Godel's *arithmetical* formula for encodes-a-proof - are 'nonstandard proofs' of Godel's Statement. I mean, they're not what we would call *actual* proofs. An actual proof would have a standard number corresponding to it. A nonstandard proof might look like... well, it's hard to envision, but it might be something like, 'Godel's statement is true, because not-not-Godel's statement, because not-not-not-not-Godel's statement', and so on going *backward forever*, every step of the proof being valid, because nonstandard numbers have an infinite number of predecessors."

And there's no way to say, 'You can't have an infinite number of derivations in a proof'?

"Not in first-order logic. If you could say that, you could rule out numbers with infinite numbers of predecessors, meaning that you could rule out all infinite-in-both-directions chains, and hence rule out all nonstandard numbers. And then the only *remaining* model would be the standard numbers. And then Godel's Statement would be a *semantic* implication of those axioms; there would exist *no* number encoding a proof of Godel's Statement in *any* model which obeyed the axioms of first-order arithmetic. And then, by Godel's *Completeness* Theorem, we could prove Godel's Statement from those axioms using first-order syntax. Because every *genuinely* valid implication of any collection of first-order axioms - every first-order statement that *actually does follow, in every possible model where the premises are true* - can *always* be proven, from those axioms, in first-order logic. Thus, by the *combination* of Godel's Incompleteness Theorem and Godel's Completeness Theorem, we see that there's no way to uniquely pin down the natural numbers using first-order logic. QED."

Whoa. So everyone in the human-superiority crowd gloating about how *they're* superior to mere machines and formal systems, because *they* can see that Godel's Statement is true just by their sacred and mysterious mathematical intuition...

"...Is actually committing a horrendous logical fallacy of the sort that no cleanly designed AI could ever be tricked into, yes. Godel's Statement doesn't *actually follow* from the first-order axiomatization of Peano arithmetic! There are models where all the first-order axioms are true, and yet Godel's Statement is false! The standard misunderstanding of Godel's Statement *is* something like the situation as it obtains in *second*-order logic, where there's no equivalent of Godel's Completeness Theorem. But people in the human-superiority crowd usually don't attach that disclaimer - they usually present arithmetic using the first-order version, when they're explaining what it is that they can see that a formal system can't. It's safe to say that *most* of them are inadvertently illustrating the irrational overconfidence of humans jumping to conclusions, even though there's a less stupid version of the same argument which invokes second-order logic."

Nice. But still... that proof you've shown me seems like a rather *circuitous* way of showing that you can't ever rule out infinite chains, especially since I don't see why Godel's Completeness Theorem should be true.

"Well... an equivalent way of stating Godel's Completeness Theorem is that every *syntactically* consistent set of first-order axioms - that is, every set of first-order axioms such that you cannot *syntactically* prove a contradiction from them using first-order logic - has at least one semantic model. The proof proceeds by trying to adjoin statements saying P or ~P for every first-order formula P, at least one of which must be possible to adjoin while leaving the expanded theory syntactically consistent -"

Hold on. Is there some more *constructive* way of seeing why a non-standard model has to exist?

"Mm... you could invoke the Compactness Theorem for first-order logic. The Compactness Theorem says that *if a collection of first-order statements has no model, some finite subset of those statements is also semantically unrealizable*. In other words, if a collection of first-order statements - even an *infinite* collection - is unrealizable in the sense that no possible mathematical model fits all of those premises, then there must be *some* finite subset of premises which are also unrealizable. Or modus ponens to modus tollens, if all finite subsets of a collection of axioms have at least one model, then the whole infinite collection of axioms has at least one model."

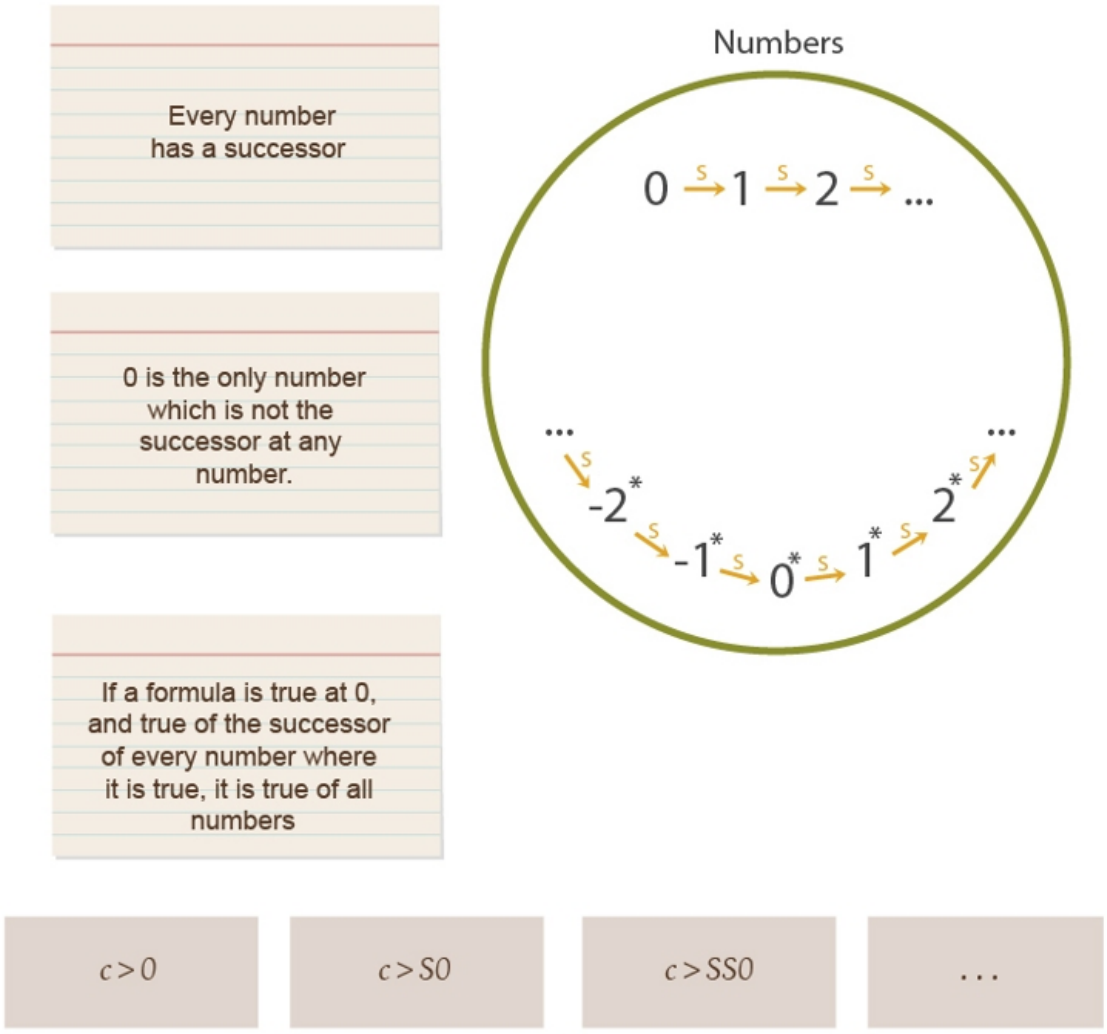
Ah, and can you explain why the Compactness Theorem should be true?

"No."

I see.

"But at least it's simpler than the Completeness Theorem, and from the Compactness Theorem, the inability of first-order arithmetic to pin down a standard model of numbers follows immediately. Suppose we take first-order arithmetic, and adjoin an axiom which says, 'There exists a number greater than 0.' Since there does in fact exist a number, 1, which is greater than 0, first-order arithmetic plus this new axiom should be semantically okay - it should have a model if any model of first-order arithmetic ever existed in the first place. Now let's adjoin a new constant symbol *c* to the language, i.e., *c* is a constant symbol referring to a single object across all statements

where it appears, the way 0 is a constant symbol and an axiom then identifies 0 as the object which is not the successor of any object. Then we start adjoining axioms saying '*c* is greater than X', where X is some concretely specified number like 0, 1, 17, 2^{256}, and so on. In fact, suppose we adjoin an *infinite* series of such statements, one for every number:"



Wait, so this new theory is saying that there exists a number *c* which is larger than every number?

"No, the infinite schema says that there exists a number *c* which is larger than any *standard* number."

I see, so this new theory *forces* a nonstandard model of arithmetic.

"Right. It rules out *only* the standard model. And the Compactness Theorem says this new theory is still semantically realizable - it has *some* model, just not the standard one."

Why?

"Because any finite subcollection of the new theory's axioms, can only use a finite number of the extra axioms. Suppose the largest extra axiom you used was '*c* is larger than 2^{256}'. In the standard model, there certainly exists a number $2^{256}+1$ with which *c* could be consistently identified. So the standard numbers must be a model of that collection of axioms, and thus that finite subset of axioms must be semantically realizable. Thus by the Compactness Theorem, the full, infinite axiom system must also be semantically realizable; it must have at least one model. Now, adding axioms never *increases* the number of compatible models of an axiom system - each additional axiom can only *filter out* models, not *add* models which are incompatible with the other axioms. So this new model of the larger axiom system - containing a number which is greater than 0, greater than 1, and greater than every other 'standard' number - must *also* be a model of first-order Peano arithmetic. That's a relatively simpler proof that first-order arithmetic - in fact, *any* first-order axiomatization of arithmetic - has nonstandard models."

Huh... I can't quite say that seems obvious, because the Compactness Theorem doesn't feel obvious; but at least it seems more specific than trying to prove it using Godel's Theorem.

"A similar construction to the one we used above - adding an infinite series of axioms saying that a thingy is even larger - shows that if a first-order theory has models of unboundedly large finite size, then it has at least one infinite model. To put it even more alarmingly, there's no way to characterize the property of *finiteness* in first-order logic! You can have a first-order theory which characterizes models of cardinality 3 - just say that there exist x, y, and z which are not equal to each other, but with all objects being equal to x or y or z. But there's no first-order theory which characterizes the property of *finiteness* in the sense that all finite models fit the theory, and no infinite model fits the theory. A first-order theory either limits the size of models to some particular upper bound, or it has infinitely large models."

So you can't even say, 'x is finite', without using second-order logic? Just forming the *concept* of infinity and distinguishing it from finiteness requires second-order logic?

"Correct, for pretty much exactly the same reason you can't say 'x is only a finite number of successors away from 0'. You can say, 'x is less than a googolplex' in first-order logic, but not, in full generality, 'x is finite'. In fact there's an even *worse* theorem, the Lowenheim-Skolem theorem, which roughly says that if a first-order theory has *any* infinite model, it has models *of all possible infinite cardinalities.* There are uncountable models of first-order Peano arithmetic. There are countable models of first-order real arithmetic - countable models of any attempt to axiomatize the real numbers in first-order logic. There are countable models of Zermelo-Frankel set theory."

How could you *possibly* have a countable model of the real numbers? Didn't Cantor *prove* that the real numbers were uncountable? Wait, let me guess, Cantor implicitly used second-order logic somehow.

"It follows from the Lowenheim-Skolem theorem that he must've. Let's take Cantor's proof as showing that you can't map every set of integers onto a distinct integer - that is, the powerset of integers is larger than the set of integers. The Diagonal Argument is that if you show me a mapping like that, I can take the set which contains 0 if and only if 0 is not in the set mapped to the integer 0, contains 1 if and only if 1 is *not* in the set mapped to the integer 1, and so on. That gives you a set of integers that no integer maps to."

You know, when I was very young indeed, I thought I'd found a *counterexample* to Cantor's argument. Just take the base-2 integers - 1='1', 2='10', 3='11', 4='100', 5='101', and so on, and let each integer correspond to a set in the obvious way, keeping in mind that I was also young enough to think the integers started at 1:

1	10	11	100	101	110	111	1000	1001
{1}	{2}	{2, 1}	{3}	{3, 1}	{3, 2}	{3, 2, 1}	{4}	{4, 1}

Clearly, every set of integers would map onto a unique integer this way.

"Heh."

Yeah, I thought I was going to be famous.

"How'd you realize you were wrong?"

After an embarrassingly long interval, it occurred to me to actually try *applying* Cantor's Diagonal Argument to my own construction. Since 1 is in {1} and 2 is in {2}, they wouldn't be in the resulting set, but 3, 4, 5 and everything else would be. And of course my construct didn't have the set {3, 4, 5, ...} anywhere in it. I'd mapped all the *finite* sets of integers onto integers, but none of the infinite sets.

"Indeed."

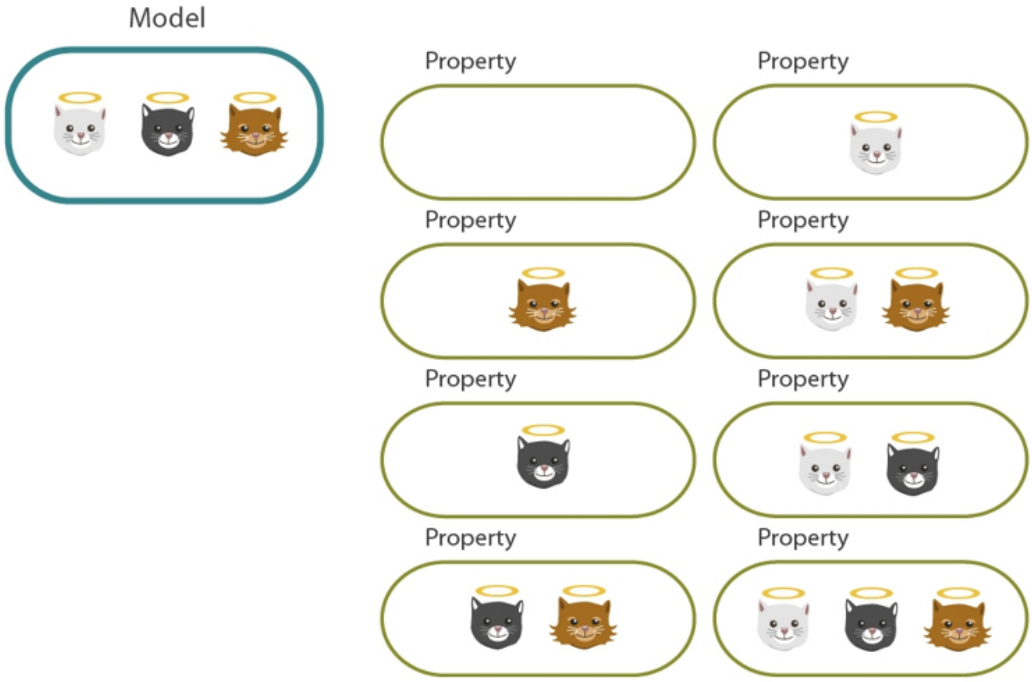
I was then tempted to *go on* arguing that Cantor's Diagonal Argument was wrong *anyhow* because it was wrong to have infinite sets of integers. Thankfully, despite my young age, I was self-aware enough to realize I was being tempted to become a mathematical crank - I had also read a book on mathematical cranks by this point - and so I just quietly gave up, which was a valuable life lesson.

"Indeed."

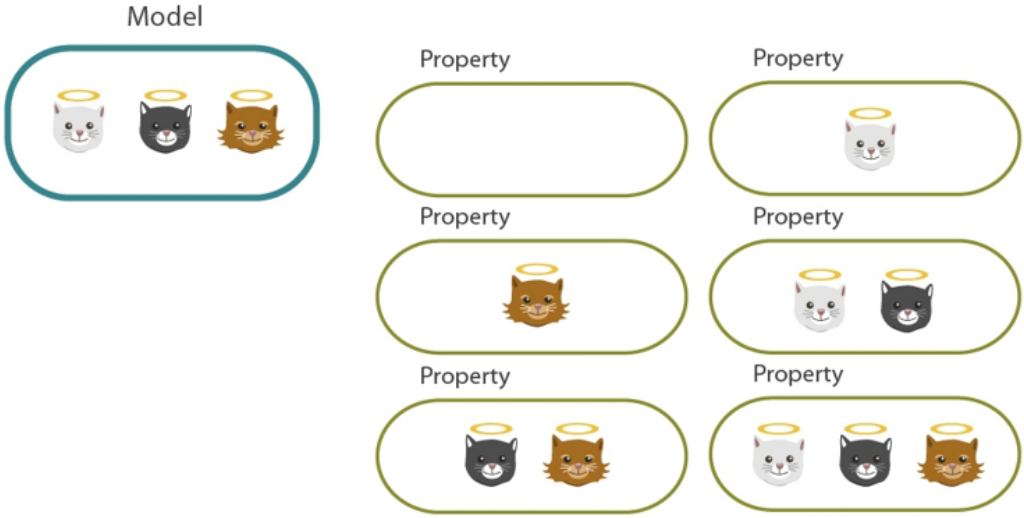
But how exactly does Cantor's Diagonal Argument depend on second-order logic? Is it something to do with nonstandard integers?

"Not exactly. What happens is that there's no way to make a first-order theory contain *all* subsets of an infinite set; there's no way to talk about *the* powerset of the integers. Let's illustrate using a finite metaphor. Suppose you have the axiom "All kittens are innocent." One model of that axiom might contain five kittens, another model might contain six kittens."

"In a second-order logic, you can talk about *all* possible collections of kittens - in fact, it's built into the syntax of the language when you quantify over all properties."



"In a first-order set theory, there are *some* subsets of kittens whose existence is provable, but others might be missing."



"Though that image is only metaphorical, since you *can* prove the existence of all the finite subsets. Just imagine that's an infinite number of kittens we're talking about up there."

And there's no way to say that *all possible* subsets exist?

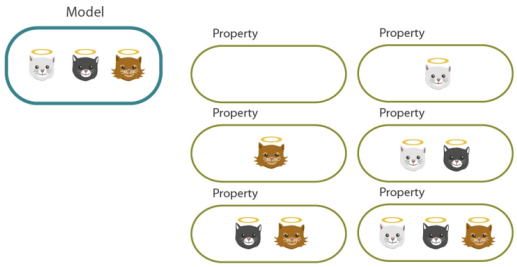
"Not in first-order logic, just like there's no way to say that you want as few natural numbers as possible. Let's look at it from the standpoint of first-order set theory. The Axiom of Powerset says:"

$$\forall A \, \exists P \, \forall B \, [B \in P \iff \forall C \, (C \in B \Rightarrow C \in A)]$$

Okay, so that says, for every set A, there exists a set P which is the *power set* of all subsets of A, so that for every set B, B is inside the powerset P *if and only if* every element of B is an element of A. Any set which contains only elements from A, will be inside the powerset of A. Right?

"Almost. There's just one thing wrong in that explanation - the word 'all' when you say 'all subsets'. The Powerset Axiom says that for any collection of elements from A, *if a set B happens to exist* which embodies that collection, that set B is inside the powerset P of A. There's no way of saying, within a first-order logical theory, that a set exists for *every possible* collection of A's elements. There may be *some* sub-collections of A whose existence you can prove. But other sub-collections of A will happen to exist as sets inside some models, but not exist in others."

So in the same way that first-order Peano arithmetic suffers from mysterious extra numbers, first-order set theory suffers from mysterious missing subsets.



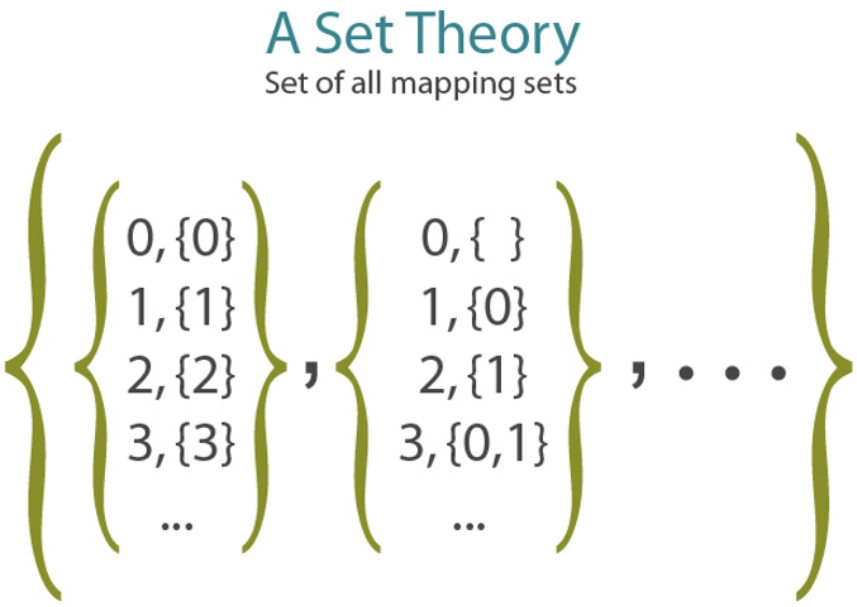
"Precisely. A first-order set theory might happen to be missing the particular infinite set corresponding to, oh, say, {3, 8, 17, 22, 28, ...} where the '...' is an infinite list of random numbers with no *compact* way of specifying them. If there's a compact way of specifying a set - if there's a finite formula that describes it - you can often prove it exists. But *most* infinite sets won't have any finite specification. It's precisely the claim to generalize over *all possible collections* that characterizes second-order logic. So it's trivial to say in a second-order set theory that *all* subsets exist. You would just say that for any set A, for any possible predicate P, there exists a set B which contains x iff x in A and Px."

I guess that torpedoes my clever idea about using first-order set theory to uniquely characterize the standard numbers by first asserting that there exists a set containing *at least* the standard numbers, and then talking about the *smallest subset* which obeys the Peano axioms.

"Right. When you talk about the numbers using first-order set theory, if there are *extra* numbers inside your set of numbers, the subset containing *just* the standard numbers must be missing from the powerset of that set. Otherwise you could find the smallest subset inside the powerset such that it contained 0 and contained the successor of every number it contained."

Hm. So then what exactly goes wrong with Cantor's Diagonal Argument?

"Cantor's Diagonal Argument uses the idea of a mapping between integers and sets of integers. In set theory, each mapping would itself be a set - in fact there would be a set of all mapping sets:"



"There's no way to first-order assert the existence of *every possible mapping* that *we* can imagine from outside. So a first-order version of the Diagonal Argument would show that in any *particular* model, for any mapping *that existed in the model* from integers to sets of integers, the model would also contain a diagonalized set of integers that wasn't in that mapping. This doesn't mean that *we* couldn't count all the sets of integers which *existed in the model.* The model could have so many 'missing' sets of integers that the remaining sets were denumerable. But then some mappings from integers to sets would also be missing, and in particular, the 'complete' mapping we can imagine from outside would be missing. And for every mapping that *was* in the model, the Diagonal Argument would construct a set of integers that wasn't in the mapping. On the outside, *we* would see a possible mapping from integers to sets - but that mapping wouldn't exist *inside* the model as a set. It takes a logic-of-collections to say that *all possible* integer-collections exist as sets, or that *no possible* mapping exists from the integers onto those sets."

So if first-order logic can't even talk about *finiteness* vs. *infiniteness* - let alone prove that there are *really* more sets of integers than integers - then why is anyone interested in first-order logic in the first place? Isn't that like trying to eat dinner using only a fork, when there are lots of interesting foods which *provably* can't be eaten with a fork, and you have a spoon?

"Ah, well... some people believe there *is* no spoon. But let's take that up next time."

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Second-Order Logic: The Controversy"

Previous post: "Standard and Nonstandard Numbers"

Second-Order Logic: The Controversy

Followup to: Godel's Completeness and Incompleteness Theorems

"So the question you asked me last time was, 'Why does anyone bother with first-order logic at all, if second-order logic is so much more powerful?'"

Right. If first-order logic can't talk about finiteness, or distinguish the size of the integers from the size of the reals, why even bother?

"The first thing to realize is that first-order theories can still have a *lot* of power. First-order arithmetic does narrow down the possible models by a lot, even if it doesn't narrow them down to a *single* model. You can prove things like the existence of an infinite number of primes, because *every* model of the first-order axioms has an infinite number of primes. First-order arithmetic is never going to prove anything that's *wrong* about the standard numbers. Anything that's true in *all* models of first-order arithmetic will also be true in the *particular* model we call the standard numbers."

Even so, if first-order theory is strictly weaker, why bother? Unless second-order logic is just as incomplete relative to third-order logic, which is weaker than fourth-order logic, which is weaker than omega-order logic -

"No, surprisingly enough - there's tricks for making second-order logic encode any proposition in third-order logic and so on. If there's a collection of third-order axioms that characterizes a model, there's a collection of second-order axioms that characterizes the same model. Once you make the jump to second-order logic, you're *done* - so far as anyone knows (so far as I know) there's *nothing* more powerful than second-order logic in terms of *which models it can characterize*."

Then if there's one spoon which can eat anything, why not just use the spoon?

"Well... this gets into complex issues. There are mathematicians who don't believe there *is* a spoon when it comes to second-order logic."

Like there are mathematicians who don't believe in infinity?

"Kind of. Look, suppose you *couldn't* use second-order logic - you belonged to a species that doesn't have second-order logic, or anything like it. Your species doesn't have any native mental intuition you could use to construct the notion of 'all properties'. And then suppose that, after somebody used first-order set theory to prove that first-order arithmetic had many possible models, you stood around shouting that you believed in only *one* model, what you called the *standard* model, but you couldn't explain what made this model different from any other model -"

Well... a lot of times, even in math, we make statements that genuinely mean something, but take a while to figure out how to define. I think somebody who talked about 'the numbers' would mean something even before second-order logic was invented.

"But here the hypothesis is that you belong to a species that *can't* invent second-order logic, or think in second-order logic, or anything like it."

Then I suppose you want me to draw the conclusion that this hypothetical alien is just standing there shouting about standardness, but its words don't mean anything because they have no way to pin down one model as opposed to another one. And I expect this species is also magically forbidden from talking about all possible subsets of a set?

"Yeah. They can't talk about the largest powerset, just like they can't talk about the smallest model of Peano arithmetic."

Then you could arguably deny that shouting about the 'standard' numbers would mean anything, to the members of this particular species. You might as well shout about the 'fleem' numbers, I guess.

"Right. Even if all the members of this species *did* have a built-in sense that there was a special model of first-order arithmetic that was fleemer than any other model, if that fleem-ness wasn't bound to anything else, it would be meaningless. Just a floating word. Or if all you could do was define fleemness as floobness and floobness as fleemness, you'd have a loop of floating words; and that might give you the impression that each particular word had a meaning, but the loop as a whole wouldn't be connected to reality. That's why it doesn't help to say that the standard model of numbers is the smallest among all possible models of Peano arithmetic, if you can't define 'smallest possible' any more than you can define 'connected chain' or 'finite number of predecessors'."

But second-order logic *does* seem to have consequences first-order logic doesn't. Like, what about all that Godelian stuff? Doesn't second-order arithmetic semantically imply... its own Godel statement? Because the unique model of second-order arithmetic doesn't contain any number encoding a proof of a contradiction from second-order arithmetic? Wait, now I'm confused.

"No, that's correct. It's not paradoxical, because there's no effective way of finding all the *semantic* implications of a collection of second-order axioms. There's no analogue of Godel's Completeness Theorem for second-order logic - no *syntactic* system for deriving *all* the semantic consequences. Second-order logic is *sound*, in the sense that anything syntactically provable from a set of premises, is true in any model obeying those premises. But second-order logic isn't *complete*; there are semantic consequences you can't derive. If you take second-order logic at face value, there's no effectively computable way of deriving all the consequences of what you say you 'believe'... which is a major reason some mathematicians are suspicious of second-order logic. What does it mean to believe something whose consequences you can't derive?"

But second-order logic clearly has *some* experiential consequences first-order logic doesn't. Suppose I build a Turing machine that looks for proofs of a contradiction from first-order Peano arithmetic. If PA's consistency isn't provable in PA, then by the Completeness Theorem there must exist nonstandard models of PA where this machine halts after finding a proof of a contradiction. So first-order logic doesn't tell me that this machine runs forever - maybe it has nonstandard halting times, i.e., it runs at all standard N, but halts at -2* somewhere along a disconnected chain. Only second-order logic tells me there's no proof of PA's inconsistency *and therefore* this machine runs forever. Only second-order logic tells me I should *expect to see this machine keep runnin*g, and *not expect* - note falsifiability - that the machine ever halts.

"Sure, you just used a second-order theory to derive a consequence you didn't get from a first-order theory. But that's not the same as saying that you can *only* get that consequence using second-order logic. Maybe another first-order theory would give you the same prediction."

Like what?

"Well, for one thing, first-order *set theory* can prove that first-order *arithmetic* has a model. Zermelo-Fraenkel set theory can prove the existence of a set such that all the first-order Peano axioms are true about that set. It follows within ZF that sound reasoning on first-order arithmetic will never prove a contradiction. And since ZF can prove that the syntax of classical logic is sound -"

What does it mean for *set theory* to prove that *logic* is sound!?

"ZF can quote formulas as structured, and encode models as sets , and then represent a finite ZF-formula which says whether or not a set of quoted formulas is true about a model. ZF can then prove that no step of classical logic goes from premises that are true inside a set-model, to premises that are false inside a set-model. In other words, ZF can represent the idea 'formula X is semantically true in model Y' and 'these syntactic rules never produce semantically false statements from semantically true statements'."

Wait, then why can't ZF prove *itself* consistent? If ZF believes in all the axioms of ZF, and it believes that logic never produces false statements from true statements -

"Ah, but ZF can't prove that there exists any set which is a model of ZF, so it can't prove the ZF-axioms are consistent. ZF *shouldn't* be able to prove that some set is a model of ZF, because that's not true in all models. Many models of ZF don't contain any *individual* set well-populated enough for that one set to be a model of ZF all by itself."

I'm kind of surprised in a Godelian sense that ZF *can* contain sets as large as the universe of ZF. Doesn't any given set have to be smaller than the whole universe?

"Inside *any particular model* of ZF, every set *within* that model is smaller than that model. But not all models of ZF are the same size; in fact, models of ZF of *every* size exist, by the Lowenheim-Skolem theorem. So you can have *some* models of ZF - some universes in which all the elements collectively obey the ZF-relations - containing individual sets which are larger than *other* entire models of ZF. A set that large is called a *Grothendieck universe* and assuming it exists is equivalent to assuming the existence of 'strongly inaccessible cardinals', sizes so large that you provably can't prove inside set theory that anything that large exists."

Whoa.

(Pause.)

But...

"But?"

I agree you've shown that *one* experiential consequence of second-order arithmetic - namely that a machine looking for proofs of inconsistency from PA, won't be seen to

halt - can be derived from first-order set theory. Can you get *all* the consequences of second-order arithmetic in some particular first-order theory?

"You can't get all the *semantic* consequences of second-order logic, taken at face value, in *any* theory or *any* computable reasoning. What about the halting problem? Taken at face value, it's a semantic consequence of second-order logic that any given Turing machine either halts or doesn't halt -"

Personally I find it rather intuitive to imagine that a Turing machine either halts or doesn't halt. I mean, if I'm walking up to a machine and watching it run, telling me that its halting or not-halting 'isn't entailed by my axioms' strikes me as not describing any actual experience I can have with the machine. Either I'll see it halt eventually, or I'll see it keep running forever.

"My point is that the statements we *can* derive from the syntax of current second-order logic is limited by that syntax. And by the halting problem, we shouldn't ever expect a computable syntax that gives us *all* the semantic consequences. There's no possible theory you can *actually use* to get a correct advance prediction about whether an arbitrary Turing machine halts."

Okay. I agree that no computable reasoning, on second-order logic or anything else, should be able to solve the halting problem. Unless time travel is possible, but even then, you shouldn't be able to solve the expanded halting problem for machines that use time travel.

"Right, so the *syntax* of second-order logic can't prove everything. And in fact, it turns out that, in terms of what you can *prove syntactically* using the standard syntax, second-order logic is identical to a many-sorted first-order logic."

Huh?

"Suppose you have a first-order logic - one that doesn't claim to be able to quantify over all possible predicates - which does allow the universe to contain two different sorts of things. Say, the logic uses lower-case letters for all type-1 objects and upper-case letters for all type-2 objects. Like, '∀x: x = x' is a statement over all type-1 objects, and '∀Y: Y = Y' is a statement over all type-2 objects. But aside from that, you use the same syntax and proof rules as before."

Okay...

"Now add an element-relation x∈Y, saying that x is an element of Y, and add some first-order axioms for making the type-2 objects behave like collections of type-1 objects, including axioms for making sure that most describable type-2 collections exist - i.e., the collection X containing just x is guaranteed to exist, and so on. What you can *prove syntactically* in this theory will be identical to what you can prove using the standard syntax of second-order logic - even though the theory doesn't claim that *all possible* collections of type-1s are type-2s, and the theory will have models where many 'possible' collections are missing from the type-2s."

Wait, now you're saying that second-order logic is no more powerful than first-order logic?

"I'm saying that the supposed power of second-order logic derives from *interpreting* it a particular way, and taking on faith that when you quantify over *all properties*, you're actually talking about all properties."

But then second-order arithmetic is no more powerful than first-order arithmetic in terms of what it can actually *prove*?

"2nd-order arithmetic is *way* more powerful than first-order arithmetic. But that's because first-order set theory is more powerful than arithmetic, and adding the syntax of second-order logic corresponds to adding axioms with set-theoretic properties. In terms of which consequences can be *syntactically* proven, second-order arithmetic is more powerful than first-order arithmetic, but *weaker* than first-order set theory. First-order set theory can prove the existence of a model of second-order arithmetic - ZF can prove there's a collection of numbers and sets of numbers which models a many-sorted logic with syntax corresponding to second-order logic - and so ZF can prove second-order arithmetic consistent."

But first-order logic, including first-order set theory, can't even *talk about* the standard numbers!

"Right, but first-order set theory can *syntactically prove* more statements about 'numbers' than second-order arithmetic can prove. And when you combine that with the *semantic* implications of second-order arithmetic not being computable, and with any second-order logic being syntactically identical to a many-sorted first-order logic, and first-order logic having neat properties like the Completeness Theorem... well, you can see why some mathematicians would want to give up entirely on this whole second-order business."

But if you deny second-order logic you *can't even say the word* '*finite*'. You would have to believe the word 'finite' was a *meaningless noise*.

"You'd define finiteness relative to whatever first-order model you were working in. Like, a set might be 'finite' only on account of the model not containing any one-to-one mapping from that set onto a smaller subset of itself -"

But that set wouldn't *actually* be finite. There wouldn't actually be, like, only a billion objects in there. It's just that all the mappings which could *prove* the set was infinite would be mysteriously missing.

"According to some *other* model, maybe. But since there is no one true model -"

How is this not crazy talk along the lines of 'there is no one true reality'? Are you saying there's no *really* smallest set of numbers closed under succession, without all the extra infinite chains? Doesn't talking about how these theories have multiple possible models, imply that those possible models are *logical thingies* and one of them actually *does* contain the largest powerset and the smallest integers?

"The mathematicians who deny second-order logic would see that reasoning as invoking an implicit background universe of set theory. Everything you're saying makes sense *relative* to some *particular model* of set theory, which would contain possible models of Peano arithmetic as sets, and could look over those sets and pick out the smallest *in that model*. Similarly, that set theory could look over a proposed model for a many-sorted logic, and say whether there were any subsets *within* the larger universe which were missing from the many-sorted model. Basically, your brain is insisting that it lives inside some *particular* model of set theory. And then, from that standpoint, you could look over some *other* set theory and see how it was missing subsets that *your* theory had."

Argh! No, damn it, I live in the set theory that *really does* have all the subsets, with no mysteriously missing subsets or mysterious extra numbers, or denumerable collections of all possible reals that could like totally map onto the integers if the mapping that did it hadn't gone missing in the Australian outback -

"But *everybody* says that."

Okay...

"Yeah?"

Screw set theory. I live in the *physical universe* where when you run a Turing machine, and keep watching forever *in the physical universe*, you never *experience a time* where that Turing machine outputs a proof of the inconsistency of Peano Arithmetic. Furthermore, I live in a universe where space is *actually* composed of a real field and space is *actually* infinitely divisible and contains *all* the points between A and B, rather than space only containing a denumerable number of points whose existence can be proven from the first-order axiomatization of the real numbers. So to talk about *physics* - forget about mathematics - I've got to use second-order logic.

"Ah. You know, that particular response is not one I have seen in the previous literature."

Yes, well, I'm not a pure mathematician. When I ask whether I want an Artificial Intelligence to think in second-order logic or first-order logic, I wonder how that affects what the AI does in *the actual physical universe*. Here in the *actual physical universe* where times are followed by successor times, I *strongly suspect* that we should only expect to experience *standard* times, and not experience any nonstandard times. I think time is *connected*, and global connectedness is a property I can only talk about using second-order logic. I think that every *particular* time is finite, and yet time has no upper bound - that there are all finite times, but only finite times - and that's a property I can only characterize using second-order logic.

"But if you can't ever tell the difference between standard and nonstandard times? I mean, *local* properties of time can be described using first-order logic, and you can't directly *see* global properties like 'connectedness' -"

But I *can* tell the difference. There are only nonstandard times where a proof-checking machine, running forever, outputs a proof of inconsistency from the Peano axioms. So I don't expect to experience seeing a machine do that, since I expect to experience only standard times.

"Set theory can also prove PA consistent. If you use set theory to define time, you similarly won't expect to see a time where PA is proven inconsistent - those nonstandard integers don't exist in any model of ZF."

Why should I anticipate that my physical universe is restricted to having only the nonstandard times allowed by a *more* powerful set theory, instead of nonstandard times allowed by first-order arithmetic? If I then talk about a nonstandard time where a proof-enumerating machine proves ZF inconsistent, will you say that only nonstandard times allowed by some still more powerful theory can exist? I think it's clear that the way you're deciding which experimental outcomes you'll have to excuse, is by secretly assuming that *only standard times* exist regardless of which theory is required to narrow it down.

"Ah... hm. Doesn't physics say this universe is going to run out of negentropy before you can do an infinite amount of computation? Maybe there's only a bounded amount of time, like it stops before googolplex or something. That can be characterized by first-order theories."

We don't know that for certain, and I wouldn't want to build an AI that just *assumed* lifespans had to be finite, in case it was wrong. Besides, should I use a different *logic* than if I'd found myself in Conway's Game of Life, or something else really infinite? Wouldn't the same sort of creatures evolve in that universe, having the same sort of math debates?

"Perhaps no universe like that *can* exist; perhaps only finite universes can exist, because only finite universes can be uniquely characterized by first-order logic."

You just used the word 'finite'! Furthermore, taken at face value, our own universe *doesn't* look like it has a finite collection of entities related by first-order logical rules. Space and time both look like infinite collections of points - continuous collections, which is a second-order concept - and then to characterize the *size* of that infinity we'd need second-order logic. I mean, by the Lowenheim-Skolem theorem, there aren't just *denumerable* models of first-order axiomatizations of the reals, there's also *unimaginably large cardinal infinities* which obey the same premises, and that's a possibility straight out of H. P. Lovecraft. Especially if there are any *things* hiding in the *invisible cracks of space*."

"How could *you* tell if there were inaccessible cardinal quantities of points hidden inside a straight line? And anything that *locally* looks continuous each time you try to split it at a describable point, can be axiomatized by a first-order schema for continuity."

That brings up another point: Who'd *really* believe that the reason Peano arithmetic works on physical times, is because that whole infinite axiom *schema* of induction, containing for every Φ a *separate* rule saying...

$$(\Phi(0) \wedge (\forall x: \Phi(x) \rightarrow \Phi(Sx))) \rightarrow (\forall n: \Phi(n))$$

...was used to specify our universe? How improbable would it be for an *infinitely long* list of rules to be true, if there wasn't a unified reason for all of them? It seems much more likely that the *real* reason first-order PA works to describe time, is that all *properties* which are true at a starting time and true of the successor of any time where they're true, are true of all later times; and this generalization over *properties* makes induction hold for *first-order formulas* as a special case. If my native thought process is first-order logic, I wouldn't see the connection between each individual formula in the axiom schema - it would take separate evidence to convince me of each one - they would feel like independent mathematical facts. But after doing *scientific* induction over the fact that many *properties* true at zero, with succession preserving truth, seem to be true everywhere - after generalizing the simple, compact *second-order* theory of numbers and times - *then* you could invent an infinite first-order theory to approximate it.

"Maybe that just says you need to adjust whatever theory of scientific induction you're using, so that it can more easily induct infinite axiom schemas."

But why the heck would you *need* to induct infinite axiom schemas in the first place, if Reality *natively* ran on first-order logic? Isn't it far more likely that the way we ended up with these infinite lists of axioms was that Reality was manufactured - forgive the

anthropomorphism - that Reality was manufactured using an underlying schema in which time is a *connected* series of events, and space is a *continuous* field, and these are properties which happen to require second-order logic for humans to describe? I mean, if you picked out first-order theories at random, what's the chance we'd end up inside an infinitely long axiom schema that just *happened* to look like the projection of a compact second-order theory? Aren't we ignoring a sort of *hint*?

"A hint to what?"

Well, I'm not that sure myself, at this depth of philosophy. But I would currently say that finding ourselves in a physical universe where times have successor times, and space looks continuous, seems like a possible *hint* that the Tegmark Level IV multiverse - or the way Reality was manufactured, or whatever - might look more like *causal universes characterizable by compact second-order theories* than *causal universes characterizable by first-order theories*.

"But since any second-order theory can just as easily be *interpreted* as a many-sorted first-order theory with quantifiers that can range over either elements or sets of elements, how could using second-order syntax actually *improve* an Artificial Intelligence's ability to handle a reality like that?"

Good question. One obvious answer is that the AI would be able to induct what *you* would call an infinite axiom schema, as a single axiom - a simple, finite hypothesis.

"There's all *sorts* of obvious hacks to scientific induction of first-order axioms which would let you assign nonzero probability to computable infinite sequences of axioms -"

Sure. So beyond that... I would currently guess that the basic assumption behind 'behaving as if' second-order logic is true, says that the AI should act as if only the 'actually smallest' numbers will ever appear in physics, relative to some 'true' mathematical universe that it thinks it lives in, but knows it can't fully characterize. Which is roughly what I'd say human mathematicians do when they take second-order logic at face value; they assume that there's some *true* mathematical universe in the background, and that second-order logic lets them talk about it.

"And what behaviorally, experimentally distinguishes the hypothesis, 'I live in the true ultimate math with fully populated powersets' from the hypothesis, 'There's some random model of first-order set-theory axioms I happen to be living in'?"

Well... one behavioral consequence is suspecting that your time obeys an infinitely long list of first-order axioms with induction schemas for *every formula*. And then moreover believing that you'll never experience a time when a proof-checking machine outputs a proof that Zermelo-Fraenkel set theory is inconsistent - even though there's *provably* some models with times like that, which fit the axiom schema you just inducted. That sounds like secretly believing that there's a background 'true' set of numbers that you think characterizes physical time, and that it's the *smallest* such set. Another suspicious behavior is that as soon as you suspect Zermelo-Fraenkel set theory is consistent, you suddenly expect not to experience any *physical* time which ZF proves isn't a standard number. You don't think you're in the nonstandard time of some weaker theory like Peano arithmetic. You think you're in the minimal time expressible by *any and all theories*, so as soon as ZF can prove some number doesn't exist in the minimal set, you think that 'real time' lacks such a number. All of these sound like behaviors you'd carry out if you thought there was a single 'true' mathematical universe that provided the best model for describing all *physical*

phenomena, like time and space, which you encounter - and believing that this 'true' backdrop used the *largest* powersets and *smallest* numbers.

"How *exactly* do you formalize all that reasoning, there? I mean, how would you *actually* make an AI reason that way?"

Er... I'm still working on that part.

"That makes your theory a bit hard to criticize, don't you think? Personally, I wouldn't be surprised if any such *formalized* reasoning looked just like believing that you live inside a first-order set theory."

I suppose I wouldn't be *too* surprised either - it's hard to argue with the results on many-sorted logics. But if comprehending the physical universe is best done by assuming that real phenomena are characterized by a 'true' mathematics containing *the* powersets and *the* natural numbers - and thus expecting that no mathematical model we can formulate will ever contain any larger powersets or smaller numbers than those of the 'true' backdrop to physics - then I'd call that a moral victory for second-order logic. In first-order logic we aren't even supposed to be able to talk about such things.

"Really? To me that sounds like believing you live inside a model of first-order set theory, and believing that all models of any theories *you* can invent must *also* be sets in the larger model. You can prove the Completeness Theorem inside ZF plus the Axiom of Choice, so ZFC already proves that all consistent theories have models which are sets, although of course it can't prove that ZFC itself is such a theory. So - anthropomorphically speaking - no model of ZFC *expects* to encounter a theory that has fewer numbers or larger powersets than itself. No model of ZFC expects to encounter any quoted-model, any set that a quoted theory entails, which contains larger powersets than the ones in its own Powerset Axiom. A first-order set theory can even make the leap from the finite statement, 'P is true of all my subsets of X', to believing, 'P is true of all my subsets of X that can be described by this denumerable collection of formulas' - it can encompass the jump from a single axiom over 'all my subsets', to a quoted axiom *schema* over formulas. I'd sum all that up by saying, 'second-order logic is how first-order set theory feels from the inside'."

Maybe. Even in the event that neither human nor superintelligent cognition will ever be able to 'properly talk about' unbounded finite times, global connectedness, particular infinite cardinalities, or true spatial continuity, it doesn't follow that Reality is similarly limited in which physics it can privilege.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Previous post: "Godel's Completeness and Incompleteness Theorems"

Mixed Reference: The Great Reductionist Project

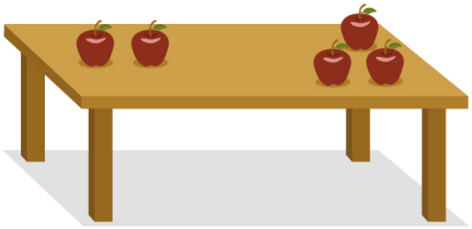
Followup to: Logical Pinpointing, Causal Reference

> Take the universe and grind it down to the finest powder and sieve it through the finest sieve and then *show* me one atom of justice, one molecule of mercy.
>
> - Death, in *Hogfather* by Terry Pratchett

Meditation: So far we've talked about two kinds of meaningfulness and two ways that sentences can refer; a way of comparing to physical things found by following pinned-down causal links, and logical validity by comparison to models pinned-down by axioms. Is there anything else that can be meaningfully talked about? Where would you find justice, or mercy?

...
...
...

Suppose that I pointed at a couple of piles of apples on a table, a pile of two apples and a pile of three apples.



And lo, I said: "If we took the number of apples in each pile, and multiplied those numbers together, we'd get six."

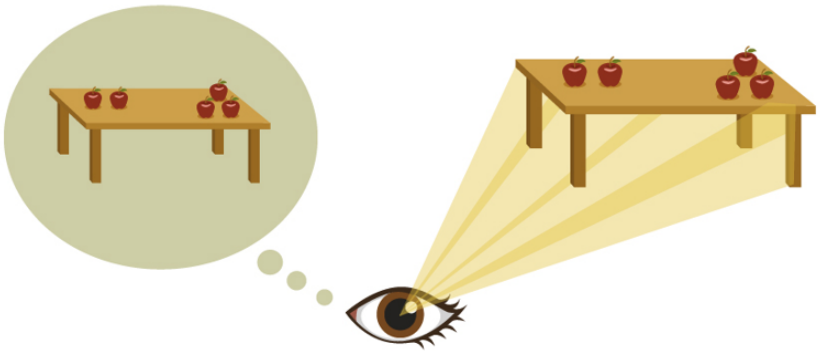
Nowhere in the physical universe is that 'six' written - there's nowhere in the laws of physics where you'll find a floating six. Even on the table itself there's only five apples, and apples aren't fundamental. Or to put it another way:

Take the apples and grind them down to the finest powder and sieve them through the finest sieve and then *show* me one atom of sixness, one molecule of multiplication.

Nor can the statement be true as a matter of pure math, comparing to some Platonic six within a mathematical model, because we could physically take one apple off the table and *make* the statement false, and you can't do that with math.

This question doesn't feel like it should be very hard. And indeed the answer is not very difficult, but it is worth spelling out; because cases like "justice" or "mercy" will turn out to proceed in a similar fashion.

Navigating to the six requires a *mixture* of physical and logical reference. This case begins with a physical reference, when we navigate to the physical apples on the table by talking about *the cause of* our apple-seeing experiences:



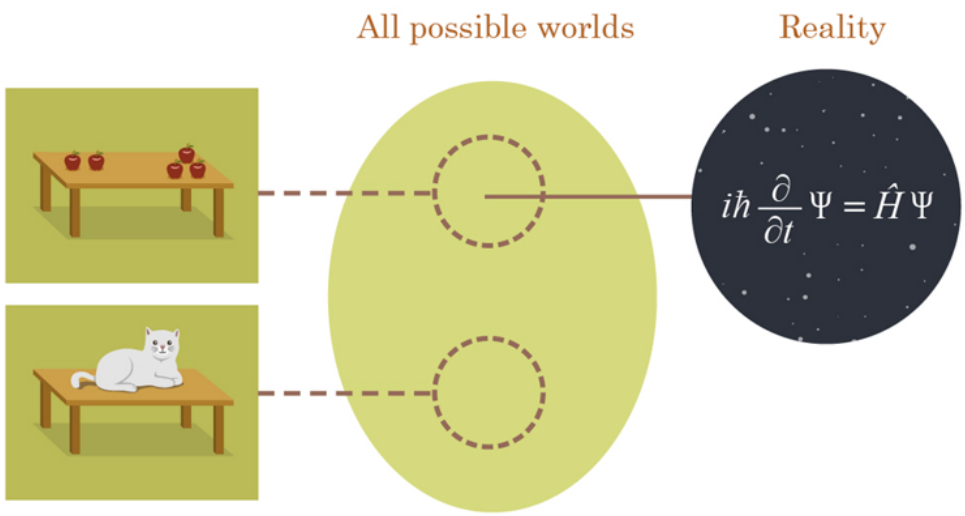
Next we have to call the stuff on the table 'apples'. But how, oh how can we do this, when grinding the universe and running it through a sieve will reveal not a single particle of appleness?

This part was covered at some length in the Reductionism sequence. Standard physics uses the same *fundamental* theory to describe the flight of a Boeing 747 airplane, and collisions in the Relativistic Heavy Ion Collider. Nuclei and airplanes alike, according to our understanding, are obeying special relativity, quantum mechanics, and chromodynamics.

We also use entirely different *models* to understand the aerodynamics of a 747 and a collision between gold nuclei in the RHIC. A computer modeling the aerodynamics of a 747 may not contain a single token, a single bit of RAM, that represents a quark. (Or a quantum field, really; but you get the idea.)

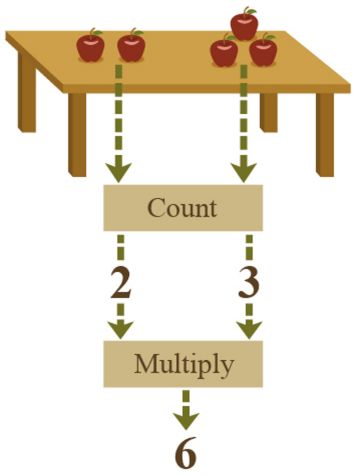
So is the 747 made of something other than quarks? And is the statement "this 747 has wings" meaningless or false? No, we're just *modeling* the 747 with *representational elements* that do not have a one-to-one correspondence with *individual* quarks.

Similarly with apples. To compare a mental image of high-level apple-objects to physical reality, for it to be true under a correspondence theory of truth, doesn't require that apples be *fundamental in physical law.* A single discrete element of fundamental physics is not the only thing that a statement can ever be compared-to. We just need truth conditions that categorize the low-level states of the universe, so that different low-level physical states are inside or outside the mental image of "some apples on the table" or alternatively "a kitten on the table".

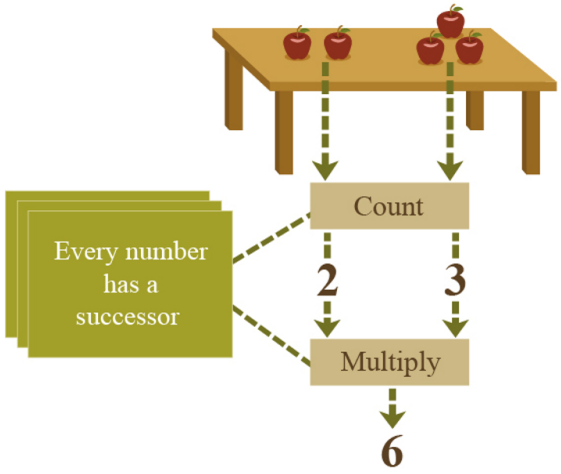


Now we can draw a correspondence from our image of discrete high-level apple objects, to reality.

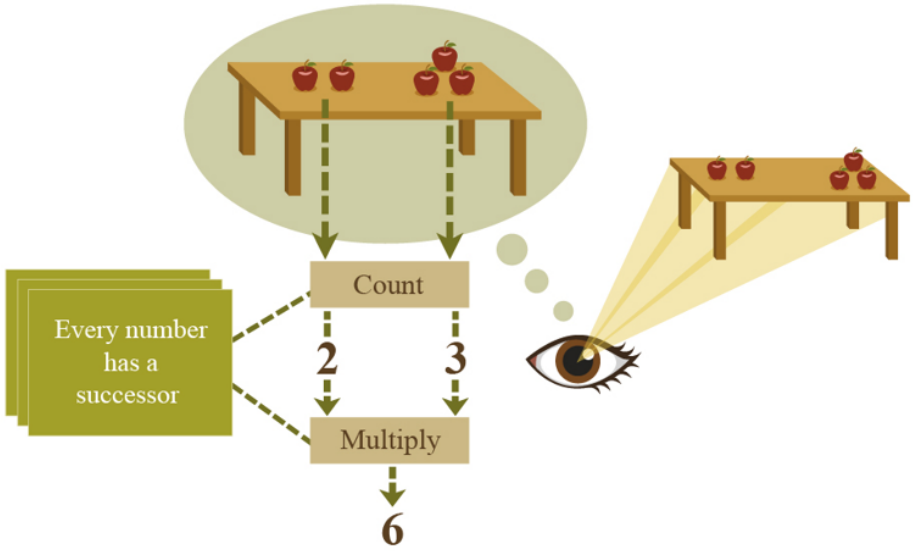
Next we need to count the apple-objects in each pile, using some procedure along the lines of going from apple to apple, marking those already counted and not counting them a second time, and continuing until all the apples in each heap have been counted. And then, having counted two numbers, we'll multiply them together. You can imagine this as taking the physical state of the universe (or a high-level representation of it) and running it through a series of functions leading to a final output:



And of course operations like "counting" and "multiplication" are pinned down by the number-axioms of Peano Arithmetic:

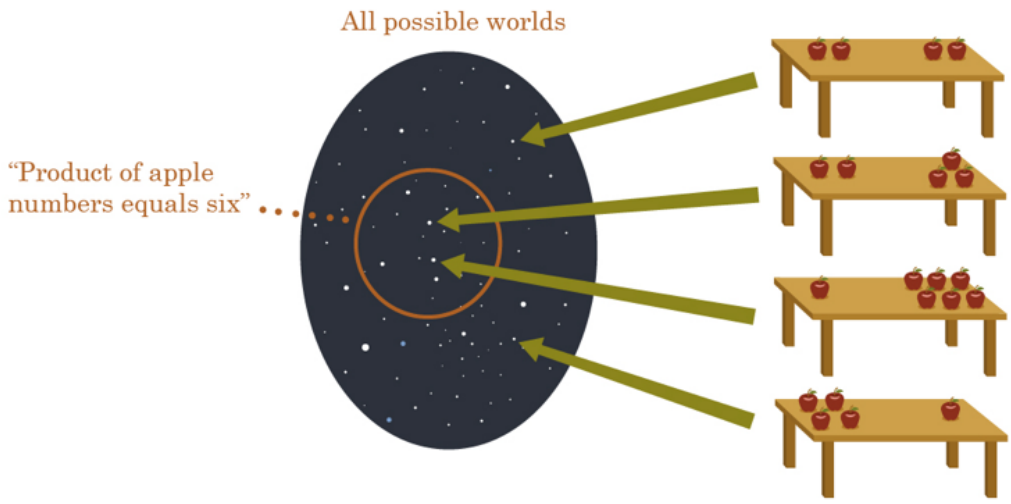


And we shouldn't forget that the image of the table, is being calculated from eyes which are in causal contact with the real table-made-of-particles out there in physical reality:



And then there's also the point that the Peano axioms themselves are being quoted inside your brain in order to pin down the *ideal* multiplicative result - after all, you can get multiplications *wrong* - but I'm not going to draw the image for that one. (We tried, and it came out too crowded.)

So long as the math *is* pinned down, any table of two apple piles should yield a single output when we run the math over it. Constraining this output constrains the possible states of the original, physical input universe:



And thus "The product of the apple numbers is six" is meaningful, constraining the possible worlds. It has a truth-condition, fulfilled by a mixture of physical reality and logical validity; and the correspondence is nailed down by a mixture of causal reference and axiomatic pinpointing.

I usually simplify this to the idea of "running a logical function over the physical universe", but of course the small picture doesn't work unless the big picture works.

The Great Reductionist Project can be seen as figuring out how to express meaningful sentences in terms of a combination of *physical references* (statements whose truth-value is determined by a truth-condition directly correspnding to the real universe we're embedded in) and *logical references* (valid implications of premises, or elements of models pinned down by axioms); where both physical references and logical references are to be described 'effectively' or 'formally', in computable or logical form. (I haven't had time to go into this last part but it's an already-popular idea in philosophy of computation.)

And the Great Reductionist Thesis can be seen as the proposition that *everything* meaningful can be expressed this way *eventually.*

But it sometimes takes *a whole bunch of work.*

And to notice when somebody has *subtly* violated the Great Reductionist Thesis - to see when a current solution is *not* decomposable to physical and logical reference - requires a fair amount of self-sensitization before the transgressions become obvious.

Example: Counterfactuals.

Consider the following pair of sentences, widely used to introduce the idea of "counterfactual conditioning":

- (A) If Lee Harvey Oswald didn't shoot John F. Kennedy, someone else did.
- (B) If Lee Harvey Oswald hadn't shot John F. Kennedy, someone else would've.

The first sentence seems agreeable - John F. Kennedy definitely was shot, historically speaking, so if it wasn't Lee Harvey Oswald it was *someone*. On the other hand, unless you believe the Illuminati planned it all, it doesn't seem particularly likely that if Lee Harvey Oswald had been removed from the equation, somebody else would've shot Kennedy instead.

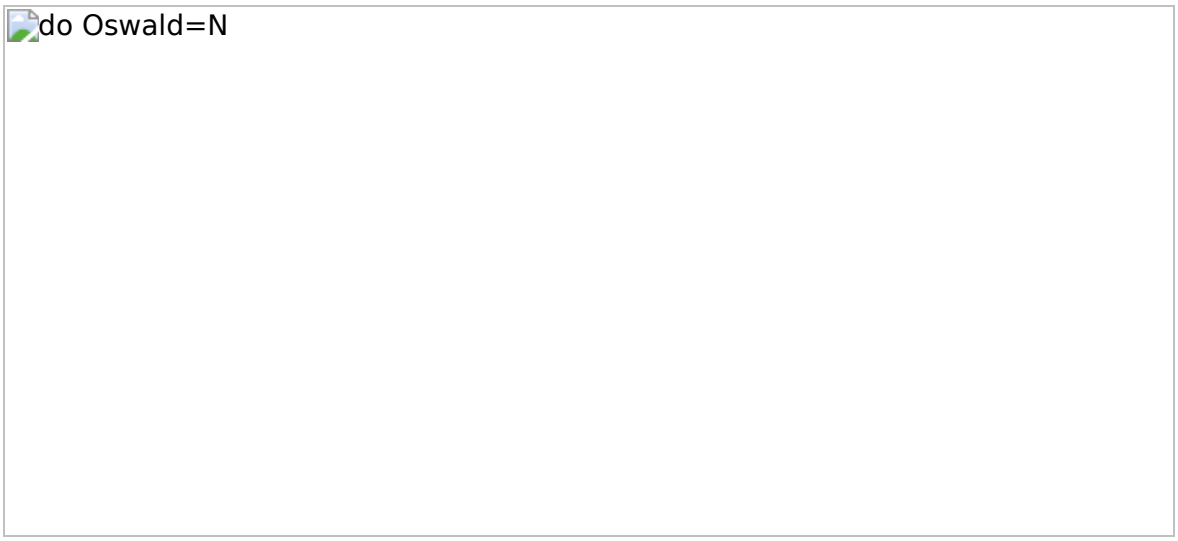
Which is to say that sentence (A) appears *true*, and sentence (B) appears *false*.

One of the historical questions about the *meaning* of causal models - in fact, of causal assertions in general - is, "How does this so-called 'causal' model of yours, differ from asserting a bunch of statistical relations? Okay, sure, these statistical dependencies have a nice neighborhood-structure, but why not just call them correlations with a nice neighborhood-structure; why use fancy terms like 'cause and effect'?"

And one of the most widely endorsed *answers,* including nowadays, is that causal models carry an extra meaning because they tell us about *counterfactual outcomes,* which ordinary statistical models don't. For example, suppose this is our causal model of how John F. Kennedy got shot:

Roughly this is intended to convey the idea that there are no Illuminati: Kennedy causes Oswald to shoot him, does not cause anybody else to shoot him, and causes the Moon landing; but once you know that Kennedy was elected, there's no correlation between his probability of causing Oswald to shoot him and his probability of causing anyone else to shoot him. In particular, there's no Illuminati who monitor Oswald and send another shooter if Oswald fails.

In any case, this diagram also implies that if Oswald hadn't shot Kennedy, nobody else would've, which is modified by a *counterfactual surgery* a.k.a. the do(.) operator, in which a node is severed from its former parents, set to a particular value, and its descendants then recomputed:



And so it was claimed that the *meaning* of the first diagram is embodied in its implicit claim (as made explicit in the second diagram) that "if Oswald *hadn't* shot Kennedy, nobody else would've". This statement is *true*, and if all the other implicit counterfactual statements are also true, the first causal model as a whole is a true causal model.

What's wrong with this picture?

Well... if you're *strict* about that whole combination-of-physics-and-logic business... the problem is that *there are no counterfactual universes* for a counterfactual statement to correspond-to. "There's apples on the table" can be true when the particles in the universe are arranged into a configuration where there's some clumps of organic molecules on the table. What arrangement of the particles in this universe could directly make true the statement "If Oswald hadn't shot Kennedy, nobody else would've"? In this universe, Oswald *did* shoot Kennedy and Kennedy *did* end up shot.

But it's a subtle sort of thing, to *notice* when you're trying to establish the truth-condition of a sentence by comparison to counterfactual universes that are not measurable, are never observed, and do not in fact actually exist.

Because our own brains carry out the same sort of 'counterfactual surgery' automatically and natively - so natively that it's embedded in the syntax of language. We don't say, "What if we perform counterfactual surgery on our models to set 'Oswald shoots Kennedy' to false?" We say, "What if Oswald *hadn't* shot Kennedy?" So there's this counterfactual-supposition operation which our brain does very quickly and invisibly to *imagine* a hypothetical non-existent universe where Oswald doesn't shoot Kennedy, and our brain very rapidly returns the supposition that Kennedy doesn't get shot, and this seems to be a fact like any other fact; and so why couldn't you just compare the causal model to this fact like any other fact?

And in one sense, "If Oswald hadn't shot Kennedy, nobody else would've" *is* a fact; it's a *mixed reference* that starts with the causal model of the *actual* universe where there are *actually* no Illuminati, and proceeds from there to the *logical* operation of counterfactual surgery to yield an answer which, like 'six' for the product of apples on the table, is not actually present anywhere in the universe. But you can't say that the

causal model is true *because* the counterfactuals are true. The truth of the counterfactuals has to be calculated from the truth of the causal model, followed by the implications of the counterfactual-surgery axioms. If the causal model couldn't be 'true' or 'false' on its own, by direct comparison to the actual real universe, there'd be no way for the counterfactuals to be true or false either, since no actual counterfactual universes exist.

So that business of counterfactuals may sound like a relatively obscure example (though it's going to play a large role in decision theory later on, and I expect to revisit it then) but it sets up some even larger points.

For example, the Born probabilities in quantum mechanics seem to talk about a 'degree of realness' that different parts of the configuration space have (proportional to the integral over squared modulus of that 'world').

Could the Born probabilities be *basic* - could there just be a *basic law of physics* which just says directly that to find out how likely you are to be in any quantum world, the integral over squared modulus gives you the answer? And the same law could've just as easily have said that you're likely to find yourself in a world that goes over the integral of modulus to the power 1.99999?

But then we would have 'mixed references' that mixed together *three kinds of stuff* - the Schrodinger Equation, a deterministic causal equation relating complex amplitudes inside a configuration space; logical validities and models; *and* a law which assigned fundamental-degree-of-realness a.k.a. magical-reality-fluid. Meaningful statements would talk about some mixture of physical laws over particle fields in our own universe, logical validities, and *degree-of-realness.*

This is just the same sort of problem if you say that causal models are meaningful and true relative to a mixture of three kinds of stuff, actual worlds, logical validities, and counterfactuals, and logical validities. You're only supposed to have *two* kinds of stuff.

People who think qualia are fundamental are also trying to build references out of at least three different kinds of stuff: physical laws, logic, and experiences.

Anthropic problems similarly revolve around a mysterious degree-of-realness, since presumably when you make more copies of people, you make their experiences more anticipate-able somehow. But this doesn't say that anthropic questions are meaningless or incoherent. It says that since we can only *talk about* anthropic problems using three kinds of stuff, we haven't finished Doing Reductionism to it yet. (I have not yet encountered a claim to have finished Reducing anthropics which (a) ends up with only two kinds of stuff and (b) does not seem to imply that I should expect my experiences to dissolve into Boltzmann-brain chaos in the next instant, given that if all this talk of 'degree of realness' is nonsense, there is no way to say that physically-lawful copies of me are more common than Boltzmann brain copies of me.)

Or to take it down a notch, naive theories of free will can be seen as *obviously not-completed Reductions* when you consider that they now contain physics, logic, and this third sort of thingy called 'choices'.

And - alas - modern philosophy is *full* of 'new sorts of stuff'; we have modal realism that makes *possibility* a real sort of thing, and then other philosophers appeal to the truth of statements about *conceivability* without any attempt to reduce conceivability

into some mixture of the actually-physically-real-in-our-universe and logical axioms; and so on, and so on.

But lest you be tempted to think that the correct course is always to just envision a simpler universe without the extra stuff, consider that we do *not* live in the 'naive un-free universe' in which all our choices are constrained by the malevolent outside hand of physics, leaving us as slaves - *reducing* choices to physics is not the same as taking a naive model with three kinds of stuff, and deleting all the 'choices' from it. This is confusing the project of getting the gnomes out of the haunted mine, with trying to unmake the rainbow. Counterfactual surgery was eventually given a formal and logical definition, but it was a lot of work to get that far - causal models had to be invented first, and before then, people could only wave their hands frantically in the air when asked what it meant for something to be a 'cause'. The overall moral I'm trying convey is that the Great Reductionist Project is *difficult;* it's not a matter of just proclaiming that there's no gnomes in the mine, or that rainbows couldn't possibly be 'supernatural'. There are all sorts of statement that were not originally, or are presently not *obviously* decomposable into physical law plus logic; but that doesn't mean you just give up immediately. The Great Reductionist Thesis is that reduction is always *possible eventually.* It is nowhere written that it is easy, or that your prior efforts were enough to find a solution if one existed.

Continued next time with justice and mercy (or rather, fairness and goodness). Because clearly, if we end up with meaningful moral statements, they're *not* going to correspond to a combination of physics and logic *plus* morality.

Mainstream status.

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "By Which It May Be Judged"

Previous post: "Causal Universes"

By Which It May Be Judged

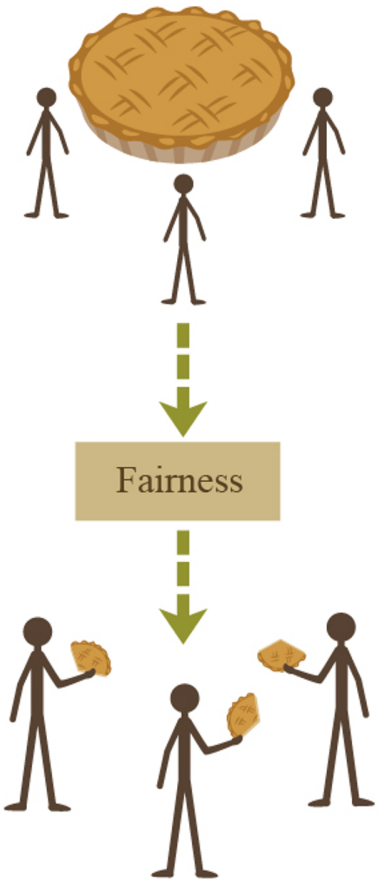
Followup to: Mixed Reference: The Great Reductionist Project

> HUMANS NEED FANTASY TO BE HUMAN.
>
> "Tooth fairies? Hogfathers? Little—"
>
> YES. AS PRACTICE. YOU HAVE TO START OUT LEARNING TO BELIEVE THE *LITTLE* LIES.
>
> "So we can believe the big ones?"
>
> YES. JUSTICE. MERCY. DUTY. THAT SORT OF THING.
>
> "They're not the same at all!"
>
> YOU THINK SO? THEN TAKE THE UNIVERSE AND GRIND IT DOWN TO THE FINEST POWDER AND SIEVE IT THROUGH THE FINEST SIEVE AND THEN *SHOW* ME ONE ATOM OF JUSTICE, ONE MOLECULE OF MERCY.
>
> - Susan and Death, in *Hogfather* by Terry Pratchett

Suppose three people find a pie - that is, three people exactly simultaneously spot a pie which has been exogenously generated in unclaimed territory. Zaire wants the entire pie; Yancy thinks that 1/3 each is fair; and Xannon thinks that fair would be taking into equal account everyone's ideas about what is "fair".

I myself would say unhesitatingly that a third of the pie each, is fair. "Fairness", as an ethical concept, can get a lot more complicated in more elaborate contexts. But in this simple context, a lot of other things that "fairness" could depend on, like work inputs, have been eliminated or made constant. Assuming no relevant conditions other than those already stated, "fairness" simplifies to the mathematical procedure of splitting the pie into equal parts; and when this logical function is run over physical reality, it outputs "1/3 for Zaire, 1/3 for Yancy, 1/3 for Xannon".

Or to put it another way - just like we get "If Oswald hadn't shot Kennedy, nobody else would've" by running a logical function over a true causal model - similarly, we can get the hypothetical 'fair' situation, whether or not it actually happens, by running the physical starting scenario through a logical function that describes what a 'fair' outcome would look like:



So am I (as Zaire would claim) just assuming-by-authority that I get to have everything my way, since I'm not defining 'fairness' the way *Zaire* wants to define it?

No more than mathematicians are flatly ordering everyone to assume-without-proof that two different numbers can't have the same successor. For fairness to be what everyone thinks is "fair" would be *entirely* circular, structurally isomorphic to "Fzeem is what everyone thinks is fzeem"... or like trying to define the counting numbers as "whatever anyone thinks is a number". It only even *looks* coherent because everyone secretly already has a mental picture of "numbers" - because their brain already navigated to the referent. But *something* akin to axioms is needed to talk about "numbers, as opposed to something else" in the first place. Even an inchoate mental image of "0, 1, 2, ..." implies the axioms no less than a formal statement - we can extract the axioms back out by asking questions about this rough mental image.

Similarly, the intuition that fairness has *something* to do with dividing up the pie equally, plays a role akin to secretly already having "0, 1, 2, ..." in mind as the subject of mathematical conversation. You need axioms, not as assumptions that aren't justified, but as pointers to what the heck the conversation is supposed to be *about.*

Multiple philosophers have suggested that this stance seems similar to "rigid designation", i.e., when I say 'fair' it intrinsically, rigidly refers to something-to-do-with-equal-division. I confess I don't see it that way myself - if somebody thinks of Euclidean geometry when you utter the sound "num-berz" they're not doing anything false, they're associating the sound to a different logical thingy. It's not about words

with intrinsically rigid referential power, it's that the words are *window dressing* on the underlying entities. I want to *talk about* a particular *logical entity,* as it might be defined by either axioms or inchoate images, regardless of which word-sounds may be associated to it. If you want to call that "rigid designation", that seems to me like adding a level of indirection; I don't care about the *word* 'fair' in the first place, I care about the logical entity of fairness. (Or to put it even more sharply: since my ontology does not have room for physics, logic, *plus* designation, I'm not very interested in discussing this 'rigid designation' business unless it's being reduced to something else.)

Once issues of justice become more complicated and all the contextual variables get added back in, we might not be sure if a *disagreement* about 'fairness' reflects:

1. The equivalent of a multiplication error within the same axioms - incorrectly dividing by 3. (Or more complicatedly: You might have a sophisticated axiomatic concept of 'equity', and *incorrectly* process those axioms to invalidly yield the assertion that, in a context where 2 of the 3 must starve and there's only enough pie for at most 1 person to survive, you should still divide the pie equally instead of flipping a 3-sided coin. Where I'm assuming that this conclusion is 'incorrect', not because I disagree with it, but because it didn't actually follow from the axioms.)
2. Mistaken models of the physical world fed into the function - mistakenly thinking there's 2 pies, or mistakenly thinking that Zaire has no subjective experiences and is not an object of ethical value.
3. People associating different logical functions to the letters F-A-I-R, which isn't a *disagreement about* some common pinpointed variable, but just different people wanting different things.

There's a lot of people who feel that this picture leaves out something fundamental, especially once we make the jump from "fair" to the broader concept of "moral", "good", or "right". And it's this worry about leaving-out-something-fundamental that I hope to address next...

...but please note, if we confess that 'right' lives in a world of physics and logic - because *everything* lives in a world of physics and logic - then we *have* to translate 'right' into those terms *somehow.*

And that is the answer Susan should have given - if she could talk about sufficiently advanced epistemology, sufficiently fast - to Death's entire statement:

> You think so? Then take the universe and grind it down to the finest powder and sieve it through the finest sieve and then *show* me one atom of justice, one molecule of mercy. And yet — Death waved a hand. And yet you act as if there is some ideal order in the world, as if there is some ... *rightness* in the universe by which it may be judged.

"But!" Susan should've said. "When we judge the universe we're comparing it to a *logical* referent, a sort of thing that isn't *in* the universe! Why, it's just like looking at a heap of 2 apples and a heap of 3 apples on a table, and comparing their invisible product to the number 6 - there isn't any 6 if you grind up the whole table, even if you grind up the whole universe, but the product is *still* 6, physico-logically speaking."

If you require that Rightness be written on some particular great Stone Tablet somewhere - to be "a light that shines from the sky", outside people, as a different

Terry Pratchett book put it - then indeed, there's no such Stone Tablet anywhere in our universe.

But there *shouldn't* be such a Stone Tablet, *given* standard intuitions about morality. This follows from the Euthryphro Dilemma out of ancient Greece.

The original Euthryphro dilemma goes, "Is it pious because it is loved by the gods, or loved by the gods because it is pious?" The religious version goes, "Is it good because it is commanded by God, or does God command it because it is good?"

The standard atheist reply is: "Would you say that it's an intrinsically good thing - even if the event has no further causal consequences which are good - to slaughter babies or torture people, if that's what God says to do?"

If we can't make it good to slaughter babies by tweaking the state of God, then morality doesn't come from God; so goes the standard atheist argument.

But if you can't make it good to slaughter babies by tweaking the physical state of *anything* - if we can't imagine a world where some great Stone Tablet of Morality has been physically rewritten, and what is right has changed - then this is telling us that...

(drumroll)

...what's "right" is a logical thingy rather than a physical thingy, that's all. The mark of a logical validity is that we can't concretely visualize a coherent possible world where the proposition is false.

And I mention this in hopes that I can show that it is not moral anti-realism to say that moral statements take their truth-value from logical entities. Even in Ancient Greece, philosophers implicitly knew that 'morality' ought to be such an entity - that it *couldn't* be something you found when you ground the Universe to powder, because then you could resprinkle the powder and make it wonderful to kill babies - though they didn't know how to say what they knew.

There's a lot of people who still feel that Death *would* be right, if the universe were all physical; that the kind of dry logical entity I'm describing here, isn't sufficient to carry the bright alive feeling of goodness.

And there are others who accept that physics and logic is everything, but who - I think *mistakenly* - go ahead and also accept Death's stance that this makes morality a lie, or, in lesser form, that the bright alive feeling can't make it. (Sort of like people who accept an incompatibilist theory of free will, also accept physics, and conclude with sorrow that they are indeed being controlled by physics.)

In case anyone is bored that I'm *still* trying to fight this battle, well, here's a quote from a recent Facebook conversation with a famous early transhumanist:

> No doubt a "crippled" AI that didn't understand the existence or nature of first-person facts could be nonfriendly towards sentient beings... Only a zombie wouldn't value Heaven over Hell. For reasons we simply don't understand, the negative value and normative aspect of agony and despair is built into the nature of the experience itself. Non-reductionist? Yes, on a standard materialist ontology. But not IMO within a more defensible Strawsonian physicalism.

It would actually be *quite surprisingly helpful* for increasing the percentage of people who will participate meaningfully in saving the planet, if there were some reliably-working standard explanation for why physics and logic together have enough room to contain morality. People who think that reductionism means we have to lie to our children, as Pratchett's Death advocates, won't be much enthused about the Center for Applied Rationality. And there are a fair number of people out there who still advocate proceeding in the confidence of ineffable morality to construct sloppily designed AIs.

So far I don't know of any exposition that works reliably - for the thesis for how morality *including* our intuitions about whether things *really are justified* and so on, is preserved in the analysis to physics plus logic; that morality has been explained rather than explained away. Nonetheless I shall now take another stab at it, starting with a simpler bright feeling:

When I see an unusually neat mathematical proof, unexpectedly short or surprisingly general, my brain gets a joyous sense of *elegance.*

There's presumably some functional slice through my brain that implements this emotion - some configuration subspace of spiking neural circuitry which corresponds to my *feeling* of elegance. Perhaps I should say that elegance is *merely* about my brain switching on its elegance-signal? But there are concepts like Kolmogorov complexity that give more formal meanings of "simple" than "Simple is whatever makes my brain feel the emotion of simplicity." Anything you do to fool my brain wouldn't make the proof *really* elegant, not in that sense. The emotion is not free of semantic content; we could build a correspondence theory for it and navigate to its logical+physical referent, and say: "Sarah feels like this proof is elegant, and her feeling is *true.*" You could even say that certain proofs are elegant even if no conscious agent sees them.

My description of 'elegance' admittedly did invoke agent-dependent concepts like 'unexpectedly' short or 'surprisingly' general. It's almost certainly true that with a different mathematical background, I would have different standards of elegance and experience that feeling on *somewhat* different occasions. Even so, that still seems like moving around in a field of *similar* referents for the emotion - much more similar to each other than to, say, the distant cluster of 'anger'.

Rewiring my brain so that the 'elegance' sensation gets activated when I see mathematical proofs where the words have lots of vowels - that wouldn't *change* what is elegant. Rather, it would make the feeling be *about* something else entirely; different semantics with a different truth-condition.

Indeed, it's not clear that this thought experiment is, or should be, *really* conceivable. If all the associated computation is about vowels instead of elegance, then from the inside you would expect that to *feel vowelly*, not *feel elegant*...

...which is to say that even feelings can be associated with logical entities. Though unfortunately not in any way that will *feel like* qualia if you can't read your own source code. I could write out an exact description of your visual cortex's spiking code for 'blue' on paper, and it wouldn't actually *look* blue to you. Still, on the higher level of description, it should seem intuitively plausible that if you tried rewriting the relevant part of your brain to count vowels, the resulting sensation would no longer have the

content or even the *feeling* of elegance. It would compute vowelliness, and feel vowelly.

My feeling of mathematical elegance is motivating; it makes me more likely to search for similar such proofs later and go on doing math. You could construct an agent that tried to add more vowels instead, and if the agent asked itself why it was doing that, the resulting justification-thought wouldn't *feel like* because-it's-elegant, it would *feel like* because-it's-vowelly.

In the same sense, when you try to do what's right, you're motivated by things like (to yet again quote Frankena's list of terminal values):

> "Life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc."

If we reprogrammed you to count paperclips instead, it wouldn't feel like *different* things having the *same* kind of motivation behind it. It wouldn't feel like doing-what's-right for a different guess about what's right. It would feel like doing-what-leads-to-paperclips.

And I quoted the above list because the feeling of rightness isn't *about* implementing a particular logical function; it contains no mention of logical functions at all; in the environment of evolutionary ancestry nobody has *heard* of axiomatization; these feelings are *about* life, consciousness, etcetera. If I could write out the whole truth-condition of the feeling in a way you could compute, you would still feel Moore's Open Question: "I can see that this event is high-rated by logical function X, but is X really *right?*" - since you can't read your own source code and the description wouldn't be commensurate with your brain's native format.

"But!" you cry. "But, is it really *better* to do what's right, than to maximize paperclips?" Yes! As soon as you start trying to cash out the logical function that gives betterness its truth-value, it will output "life, consciousness, etc. $>_B$ paperclips". And if your brain were computing a different logical function instead, like makes-more-paperclips, it wouldn't feel *better,* it would feel *moreclippy.*

But is it really *justified* to keep our own sense of betterness? Sure, and that's a logical fact - it's the objective output of the logical function corresponding to your experiential sense of what it means for something to be 'justified' in the first place. This doesn't mean that Clippy the Paperclip Maximizer will self-modify to do only things that are justified; Clippy doesn't judge between self-modifications by computing justifications, but rather, computing *clippyflurphs.*

But isn't it *arbitrary* for Clippy to maximize paperclips? Indeed; once you implicitly or explicitly pinpoint the logical function that gives judgments of arbitrariness their truth-value - presumably, revolving around the presence or absence of justifications - then this logical function will objectively yield that there's no justification whatsoever for maximizing paperclips (which is why *I'm* not going to do it) and hence that Clippy's

decision is arbitrary. Conversely, Clippy finds that there's no clippyflurph for preserving life, and hence that it is unclipperiffic. But unclipperifficness isn't arbitrariness any more than the number 17 is a right triangle; they're different logical entities pinned down by different axioms, and the corresponding judgments will have different semantic content and *feel different.* If Clippy is architected to experience that-which-you-call-qualia, Clippy's feeling of clippyflurph will be *structurally* different from the way justification feels, not just red versus blue, but vision versus sound.

But surely one *shouldn't* praise the clippyflurphers rather than the just? I quite agree; and as soon as you navigate referentially to the coherent logical entity that is the truth-condition of *should* - a function on potential actions and future states - it will agree with you that it's better to avoid the arbitrary than the unclipperiffic. Unfortunately, this logical fact does not correspond to the truth-condition of any meaningful proposition computed by Clippy in the course of how it efficiently transforms the universe into paperclips, in much the same way that rightness plays no role in that-which-is-maximized by the blind processes of natural selection.

Where moral judgment is concerned, it's logic all the way down. *ALL* the way down. Any frame of reference where you're worried that it's *really* no better to do what's right then to maximize paperclips... well, that *really* part has a truth-condition (or what does the "really" mean?) and as soon as you write out the truth-condition you're going to end up with yet another ordering over actions or algorithms or meta-algorithms or *something.* And since grinding up the universe won't and *shouldn't* yield any miniature '>' tokens, it must be a *logical* ordering. And so whatever logical ordering it is you're worried about, it probably *does* produce 'life > paperclips' - but Clippy isn't computing that logical fact any more than your pocket calculator is computing it.

Logical facts have no power to directly affect the universe except when some part of the universe is computing them, and morality is (and *should* be) logic, not physics.

Which is to say:

> The old wizard was staring at him, a sad look in his eyes. "I suppose I *do* understand now," he said quietly.
>
> "Oh?" said Harry. "Understand what?"
>
> "Voldemort," said the old wizard. "I understand him now at last. Because to believe that the world is truly like that, you must believe there is no justice in it, that it is woven of darkness at its core. I asked you why he became a monster, and you could give no reason. And if I could ask *him*, I suppose, his answer would be: Why not?"
>
> They stood there gazing into each other's eyes, the old wizard in his robes, and the young boy with the lightning-bolt scar on his forehead.
>
> "Tell me, Harry," said the old wizard, "will *you* become a monster?"
>
> "No," said the boy, an iron certainty in his voice.
>
> "Why not?" said the old wizard.
>
> The young boy stood very straight, his chin raised high and proud, and said: "There is no justice in the laws of Nature, Headmaster, no term for fairness in the equations of motion. The universe is neither evil, nor good, it simply does not

care. The stars don't care, or the Sun, or the sky. But they don't have to! *We* care! There *is* light in the world, and it is *us!*"

Part of the sequence *Highly Advanced Epistemology 101 for Beginners*

Next post: "Standard and Nonstandard Numbers"

Previous post: "Mixed Reference: The Great Reductionist Project"