Best of LessWrong: October 2012

1. LessWrong help desk - free paper downloads and more
2. Firewalling the Optimal from the Rational
3. The Useful Idea of Truth
4. Causal Diagrams and Causal Models
5. Proofs, Implications, and Models
6. How To Have Things Correctly
7. Taking "correlation does not imply causation" back from the internet
8. Skill: The Map is Not the Territory
9. Stuff That Makes Stuff Happen
10. Rationality: Appreciating Cognitive Algorithms
11. Prediction market sequence requested
12. [Link] "Fewer than X% of Americans know Y"
13. Causal Reference
14. Raising the waterline
15. Raising the forecasting waterline (part 1)

Best of LessWrong: October 2012

1. LessWrong help desk - free paper downloads and more
2. Firewalling the Optimal from the Rational
3. The Useful Idea of Truth
4. Causal Diagrams and Causal Models
5. Proofs, Implications, and Models
6. How To Have Things Correctly
7. Taking "correlation does not imply causation" back from the internet
8. Skill: The Map is Not the Territory
9. Stuff That Makes Stuff Happen
10. Rationality: Appreciating Cognitive Algorithms
11. Prediction market sequence requested
12. [Link] "Fewer than X% of Americans know Y"
13. Causal Reference
14. Raising the waterline
15. Raising the forecasting waterline (part 1)

LessWrong help desk - free paper downloads and more

Over the last year, VincentYu, gwern, myself and others have provided 132 academic papers for the LessWrong community (out of 152 requests, a 87% success rate) through the Free research, editing and articles thread. We originally intended to provide editing, research and general troubleshooting help, but article downloads are by far the most requested service.

If you're doing a LessWrong relevant project we want to help you. If you need help accessing a journal article or academic book chapter, we can get it for you. If you need some research or writing help, we can help there too.

Turnaround times for articles published in the last 20 years or so is usually less than a day. Older articles often take a couple days.

Please make new article requests in the comment section of this thread.

If you would like to help out with finding papers, please monitor this thread for requests. If you want to monitor via RSS like I do, Google Reader will give you the comment feed if you give it the URL for this thread (or use this link directly).

If you have some special skills you want to volunteer, mention them in the comment section.

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 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:

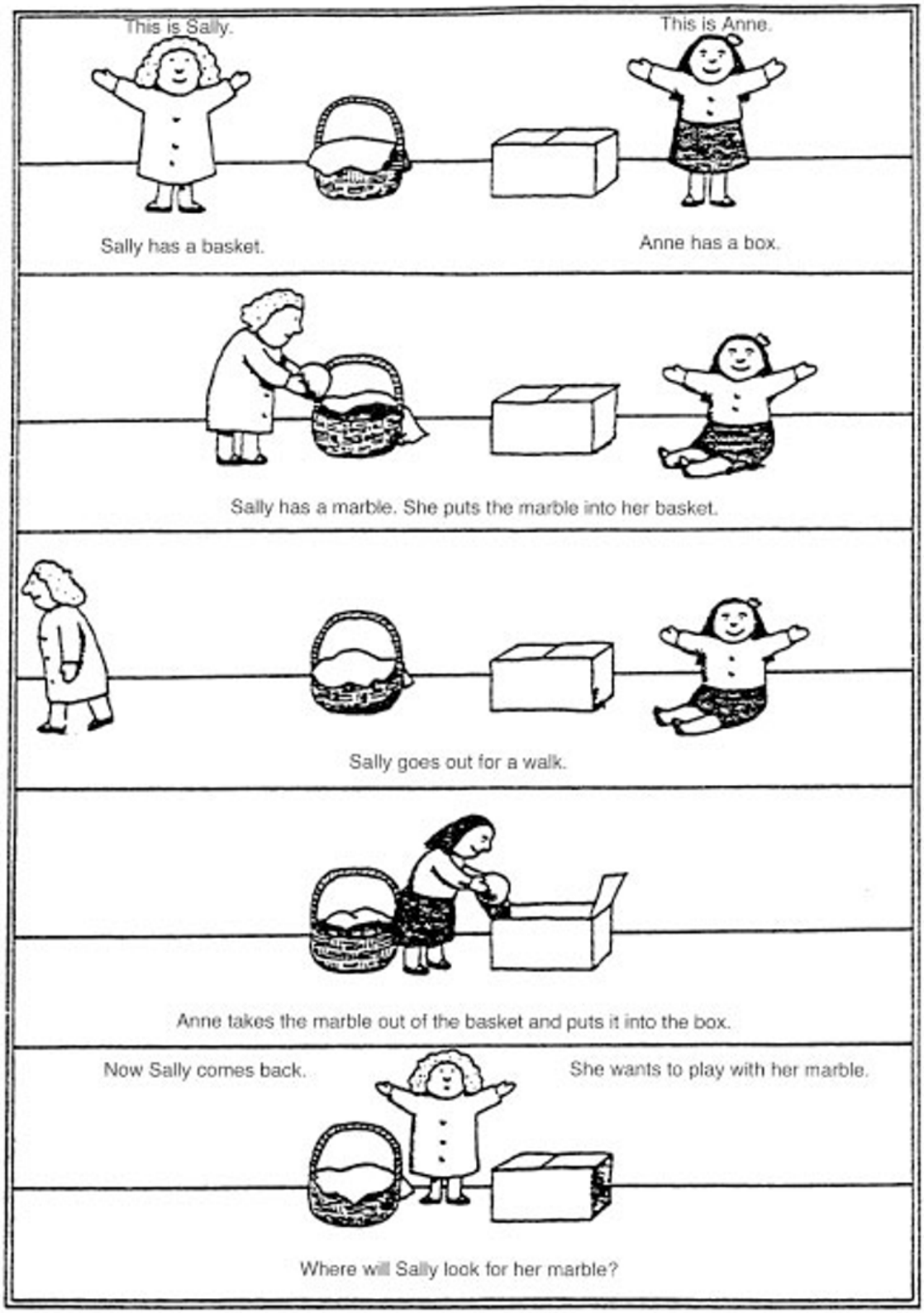
The child sees Sally hide a marble inside a covered basket, as Anne looks on.

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.

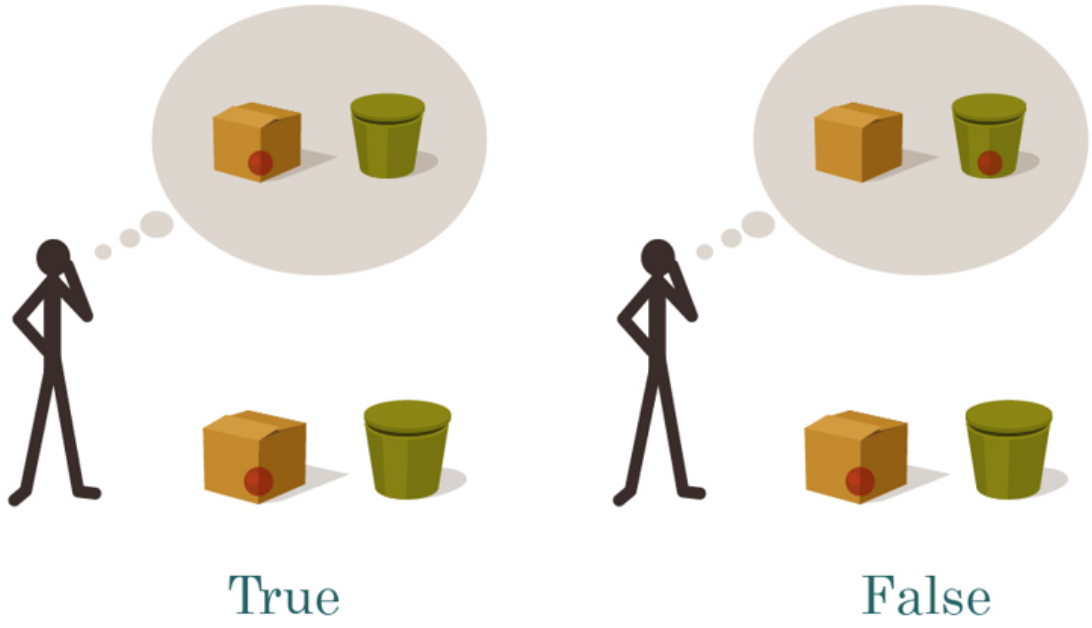
The experimenter asks the child where Sally will look for her marble.

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.

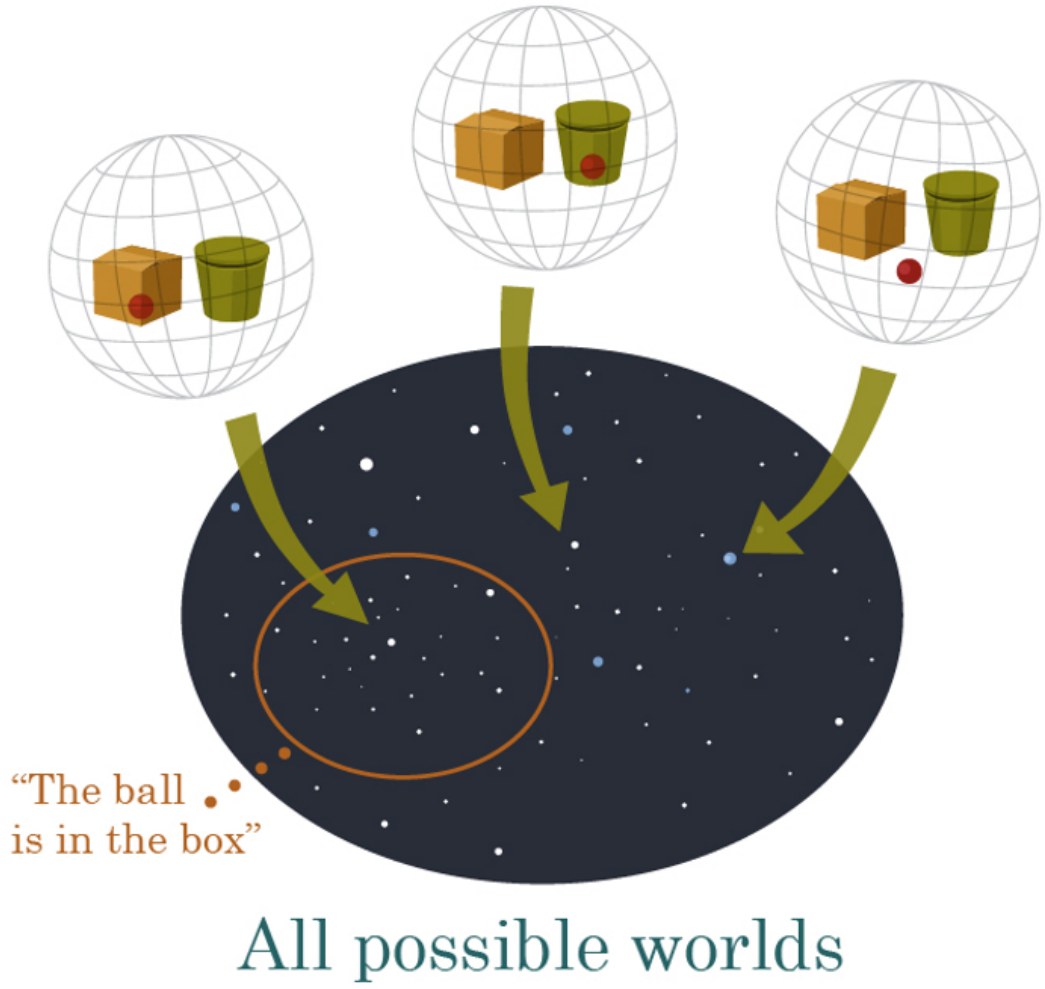


(Attributed to: Baron-Cohen, S., Leslie, L. and Frith, U. (1985) 'Does the autistic child have a "theory of mind"?', Cognition, vol. 21, pp. 37–46.)

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.



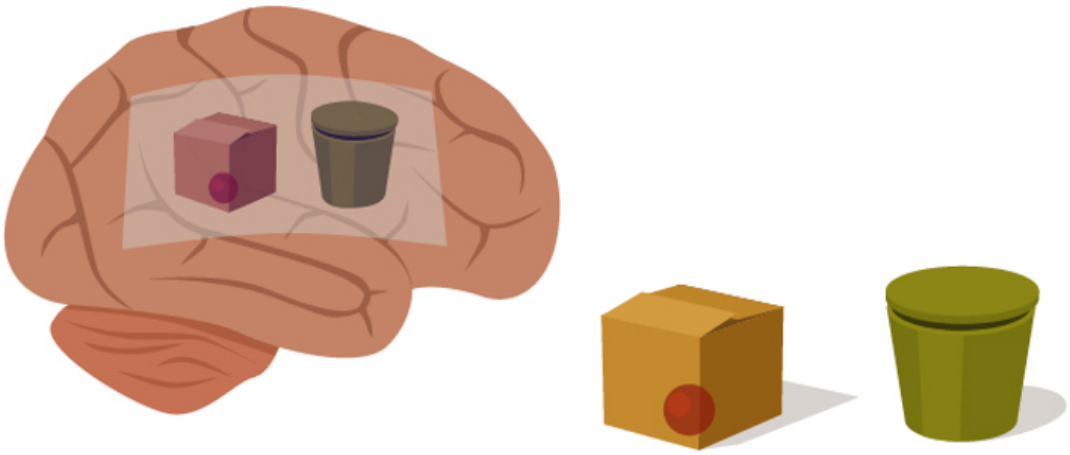
The mathematician Alfred Tarski once described the notion of 'truth' via an infinite family of truth-conditions:

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

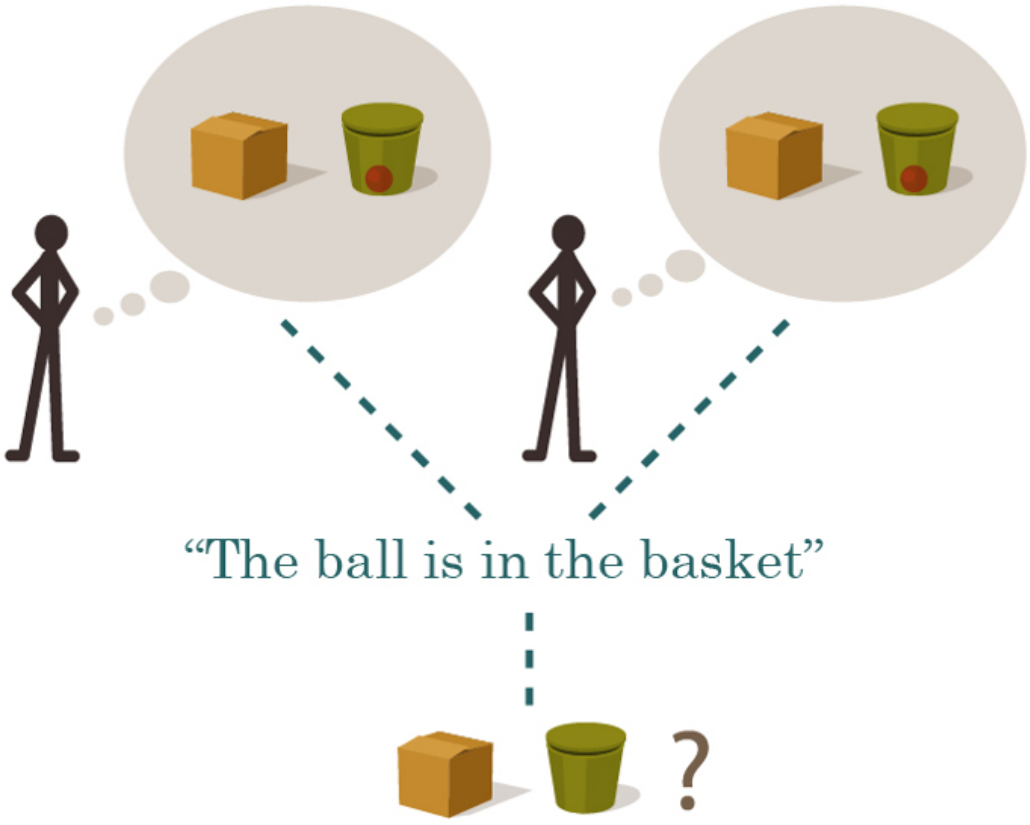
The sentence 'the sky is blue' is true if and only if the sky is blue.

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.

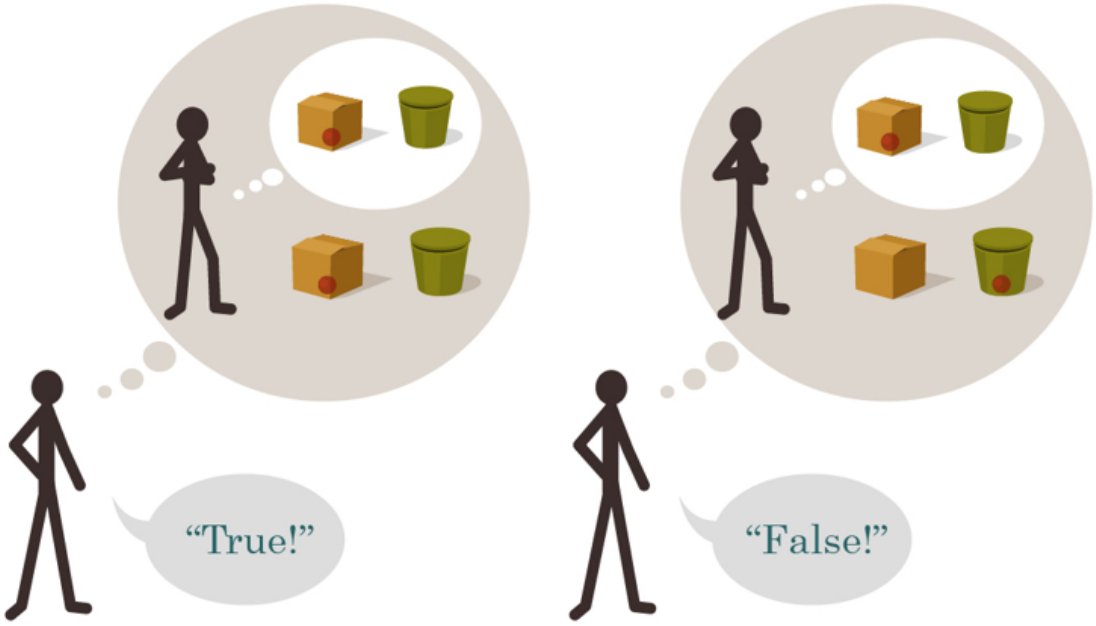


(Then why talk about these abstract 'sentences' instead of just neurally embodied beliefs? Maybe Sally and Fred believe "the same thing", i.e., their brains both have internal models of the marble inside the basket - two brain-bound beliefs with the same truth condition - in which case the thing these 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

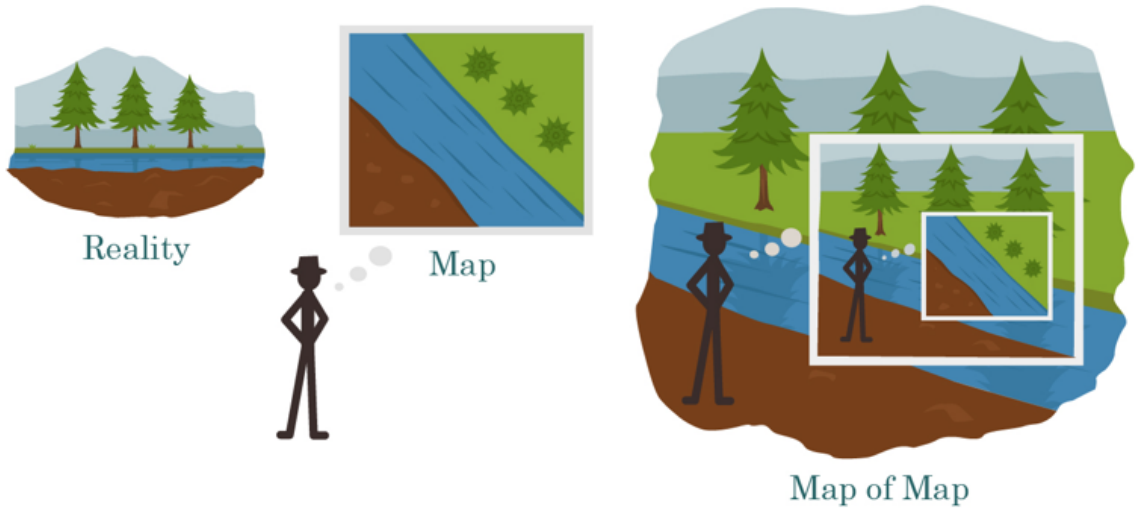
compare *someone else's belief* to *your belief:*



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'.

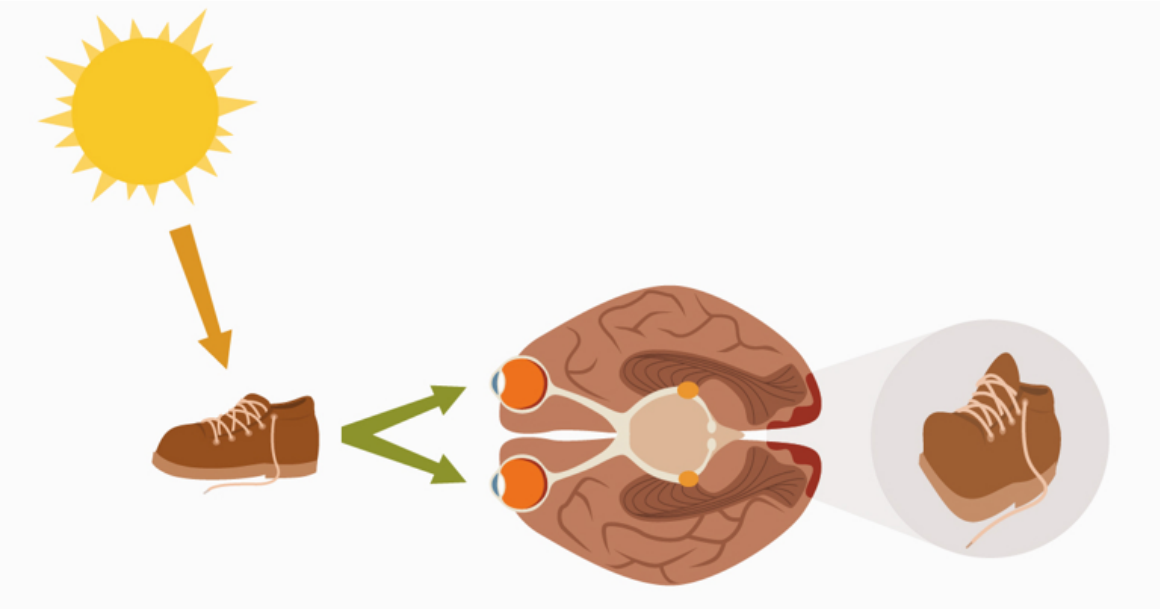
You won't get a direct collision between belief and reality - or between someone else's beliefs and reality - by sitting in your living-room with your eyes closed. But the situation is different if you open your eyes!

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.

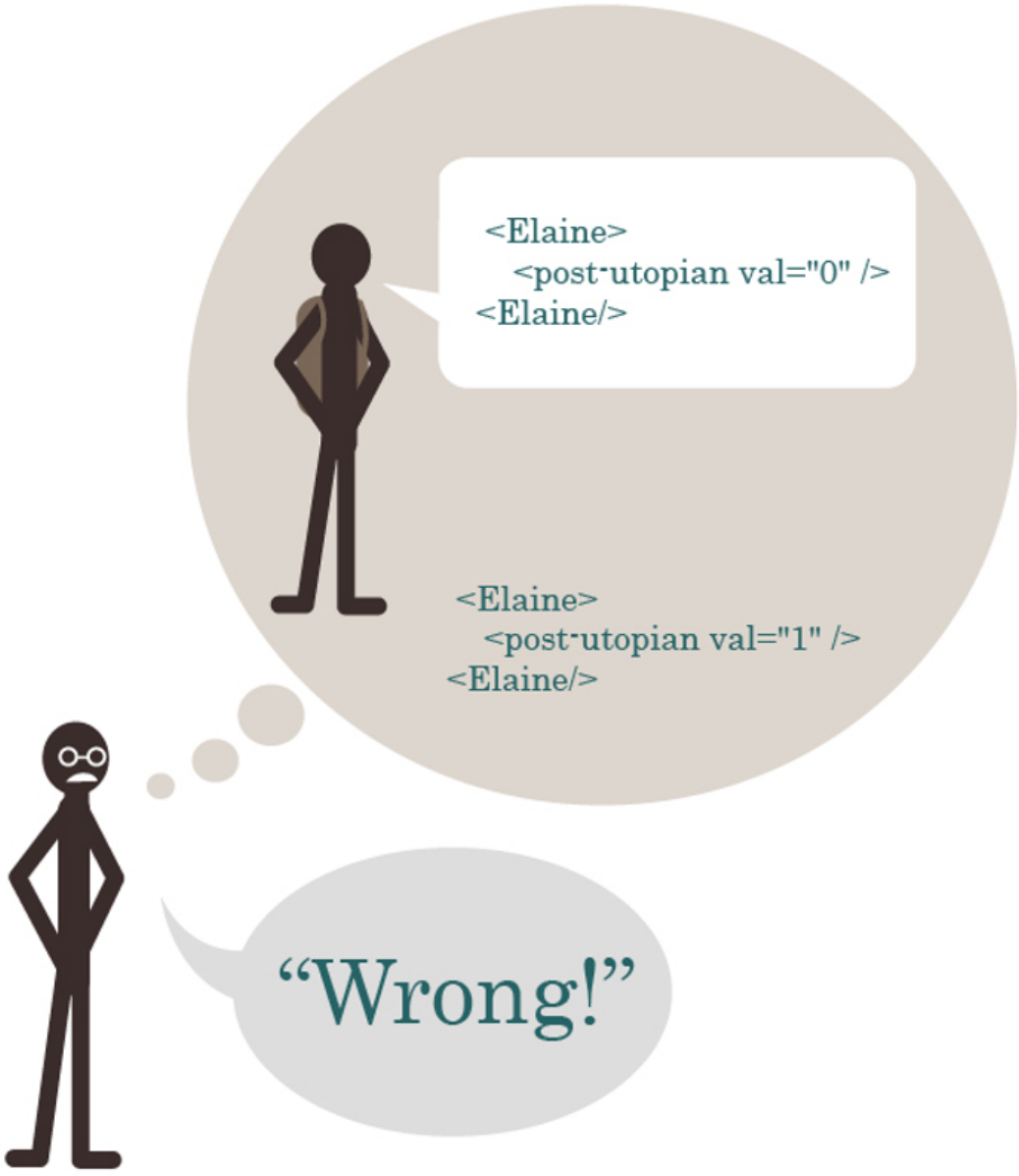
True beliefs are more likely than false beliefs to make correct experimental predictions, so if we increase our credence in hypotheses that make correct experimental predictions, our model of reality should become incrementally more true over time.

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"

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?

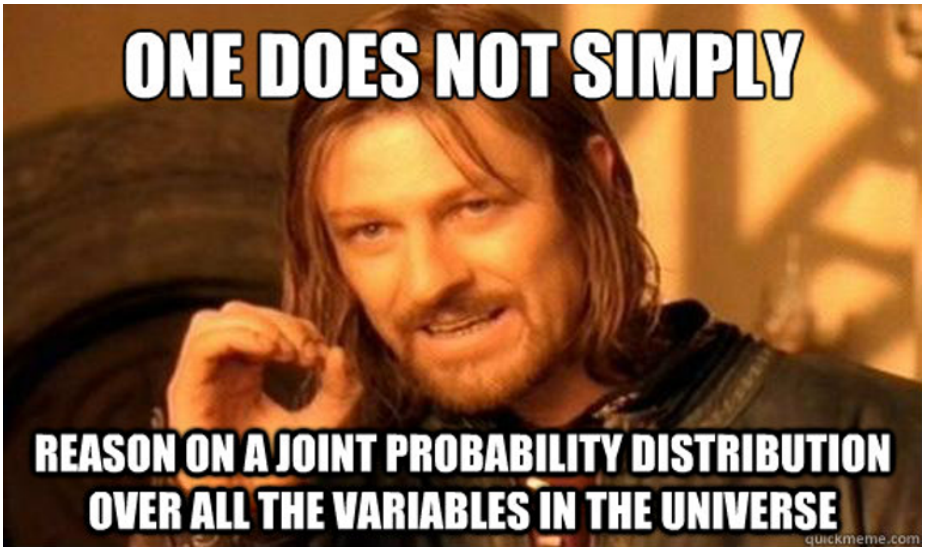
Not in real life or real Artificial Intelligence work. The problem is that this solution doesn't scale. *Boy howdy*, does it not scale! If you have a model containing *forty* binary variables - alert readers may notice that the observed physical universe contains at least forty things - and you try to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

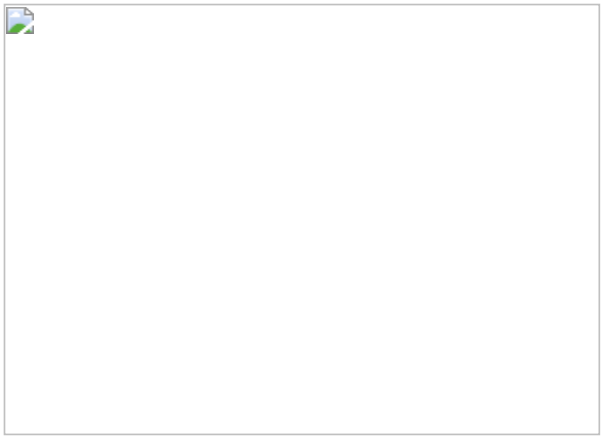
it's the difficulty of gathering enough observational data to constrain a trillion different parameters. And in a universe containing *seventy* things, things are even worse.



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:

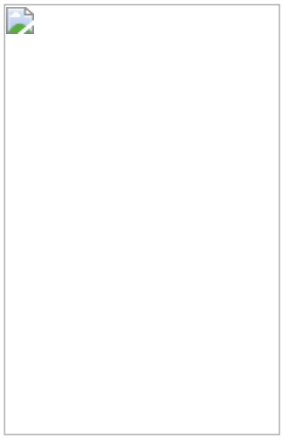
$$p(\neg e)p(\neg r)p(b|\neg r)p(a|b\neg e) = .99 * .67 * .01 * .9 = .006 = 0.6\%$$

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!

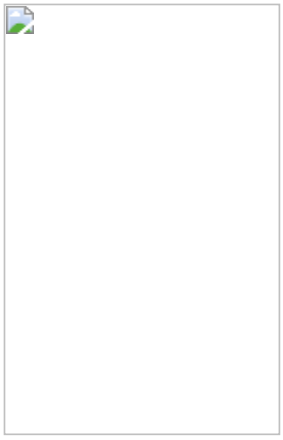
The number of free parameters in a raw joint probability distribution over a million binary variables would be $2^{1,000,000}$. Minus one.

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.

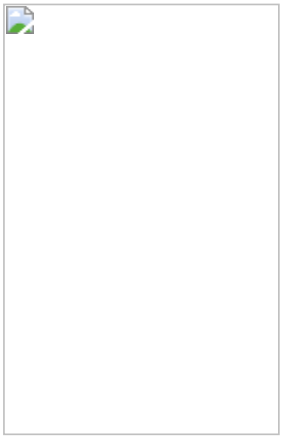
Alarm	Burglar	Recession	%
Y	Y	Y	.012
N	Y	Y	.0013
Y	N	Y	.00287
N	N	Y	.317
Y	Y	N	.003
N	Y	N	.000333
Y	N	N	.00591
N	N	N	.654

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:

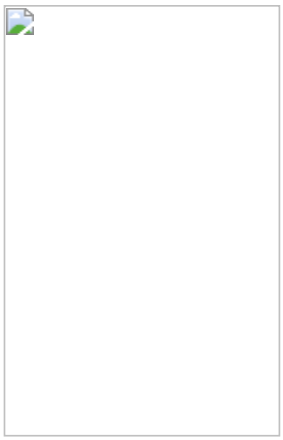
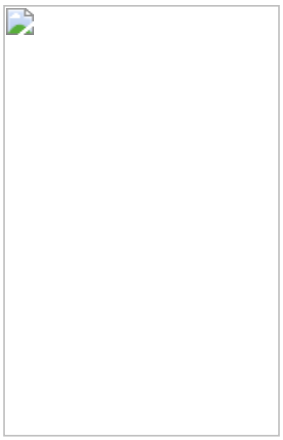
$$p(a|b) = p(a|br)$$

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.*)

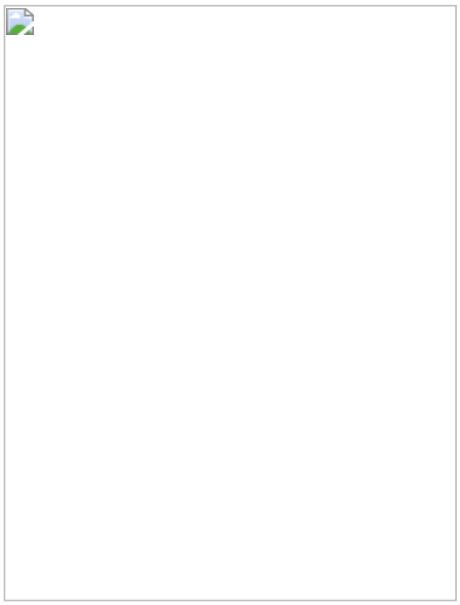
On the other hand, we can't tell the difference between:

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:

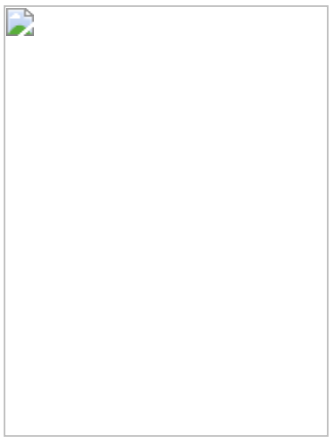
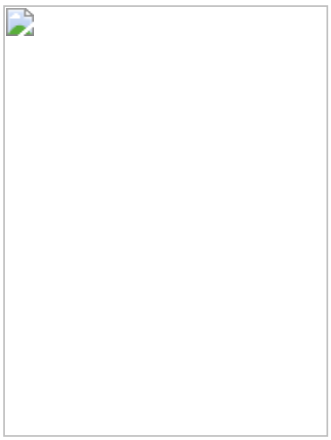
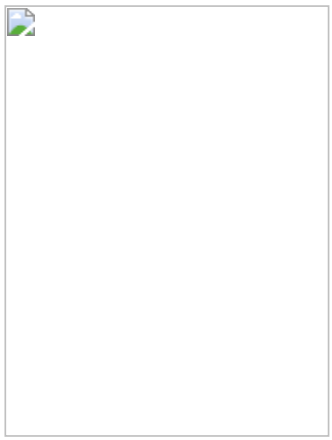
If either of those causal links existed, those two variables should be *correlated.* We shouldn't find the *lack of correlation* or *conditional independence* that we just discovered.

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"

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**

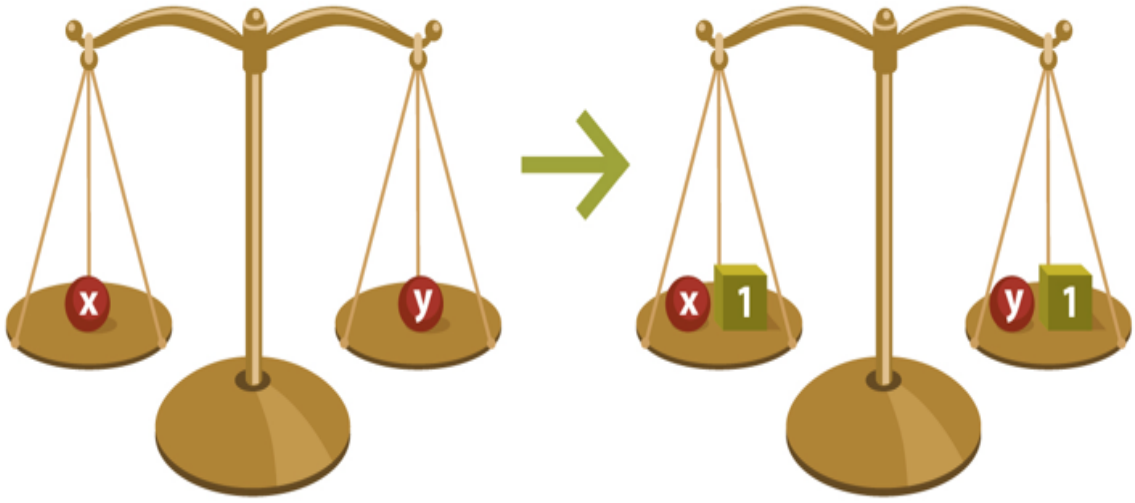
Logic is there to preserve truth, not to create truth. Whenever we take a logically valid step, we can't guarantee that the premise is true to start with, but *if* the premise is true the conclusion should always be true. Since we went from a true equation to a false equation between step 4 and step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

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

...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:

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

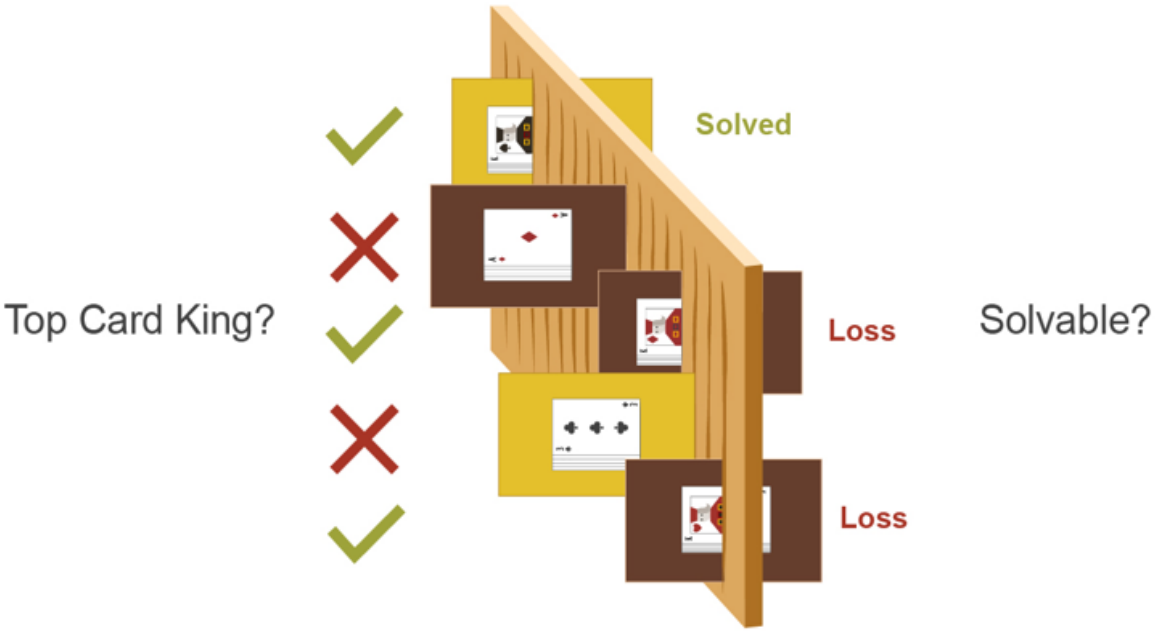
...and this activity of string-manipulation is all there is to what mathematicians call "theorem-proving" - all there is to the glorious human endeavor of mathematics.

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"

How To Have Things Correctly

I think people who are not made happier by having things either have the wrong things, or have them incorrectly. Here is how I get the most out of my stuff.

Money doesn't buy happiness. If you want to try throwing money at the problem anyway, you should buy experiences like vacations or services, rather than purchasing objects. If you have to buy objects, they should be absolute and not positional goods; positional goods just put you on a treadmill and you're never going to catch up.

Supposedly.

I think getting value out of spending money, owning objects, and having positional goods are all three of them *skills*, that people often don't have naturally but can develop. I'm going to focus mostly on the middle skill: how to have things correctly[1].

1. Obtain more-correct things in the first place.

If you and I are personal friends, you probably know that I have weird gift-receiving protocols. This is partly because I hate surprises. But it's also because I don't want to have incorrect things, cluttering my space, generating guilt because they're gifts that I never use, and generally having high opportunity cost because the giver could have gotten me something else.

This problem isn't only with gifts. People get themselves incorrect things all the time - seduced by marketing, seized by impulse, or too hurried to think about which of several choices is the best one for their wants and needs. I have some incorrect clothes, which I got because I was sick of shopping and needed *a* new pair of pants even if it was terrible; as soon as I found better pants (or whatever) those clothes were never worn again and now they're just waiting for my next haul to Goodwill. I bet a lot of people have incorrect printers, mostly because printers in general are evil, but partly because it's irritating and dull to investigate them ahead of time. Cars may also tend to fall into this category, with a lot of people neutral or ambivalent about their self-selected objects that cost thousands of dollars.

If you are not currently living in a cluttered space, or feeling guilty about not using your objects enough, or tending to dislike the things that you have, or finding yourself wanting things that you "can't" get because you already have an inferior item in the same reference class, or just buying too many not-strict-necessities than is appropriate for your budget - then this might not be a step you need to focus on. If you have objects you don't like (not just aren't getting a lot out of, that's for later steps, but actually dislike) then you might need to change your thresholds for object-acquisition.

This doesn't mean something stodgy like "before you get something, think carefully about whether you will actually use and enjoy it, using outside view information about items in this reference class". Or, well, it can mean that, but that's not the only criterion! You can also increase the amount of sheer emotional want that you allow to move you to action - wait until you more-than-idly desire it. If I were good at math, I would try to operationalize this as some sort of formula, but suffice it to say that the cost of the object (in money, but also social capital and storage space and inconvenience and whatnot) should interact with how much you just-plain-want-it and also with how much use you will likely get out of it.

Speaking of how much use you get out of it...

2. Find excuses to use your stuff.

I have a cloak. It cost me about $80 on Etsy. It is custom made, and reversible between black and gray, and made out of my favorite fabric, and falls all the way to the floor from my shoulders, and has a hood so deep that I can hide in it if I want. If I run while I wear it, it swoops out behind me. It's soft and warm but not *too* warm. I *like* my cloak.

I also have sweaters. They didn't cost me anywhere near $80, not a one of them.

When it's chilly, I reach for the cloak first.

I'm playing a game with my brain: I will let it make me spend $80 on a cloak, if it will produce enough impetus towards cloak-wearing and cloak-enjoying that I actually get $80 of value out of it. If it can't follow through, then I later trust its wants less ("last time I bought something like this, it just hung in my closet forever and I only pulled it out on Halloween!"), and then it doesn't get to make me buy any more cloaklike objects, which it really wants to be able to do. (I don't know if everyone's brain is wired to play this sort of game, but if yours is, it's worth doing.) My brain is doing a very nice job of helping me enjoy my cloak. Eventually I may let it buy another cloak in a different pair of colors, if it demonstrates that it really can keep this up long-term.

People sometimes treat not using their stuff like something that happens to them. "I never wound up using it." "It turned out that I just left it in the basement." This is silly. If I'm going to use my cloak - or my miniature cheesecake pan or my snazzy letter opener - then this is because at some point I will decide to put on my cloak, make miniature cheesecakes, or open letters with my snazzy dedicated device instead of my nail file. You know, on purpose.

Sure, some things seem to prompt you to use them more easily. If you get a new video game, and you really like it, it's probably not going turn out that you never think to play it. If you get a *cat* or something sufficiently autonomous like that, you will *know* if you are not paying it sufficient attention.

But if you get a muffin tin and you have no pre-installed prompts for "I could make muffins" because that impulse was extinguished due to lack of muffin tin, it will be easy to ignore. You're going to need to train yourself to think of muffins as a makeable thing. And you can train yourself to do that! Put the muffins on your to-do list. Lead your friends to expect baked goods. Preheat the oven and leave a stick of butter out to soften so you're committed. If that doesn't sound appealing to you - if you don't want to bake muffins - then you shouldn't have acquired a muffin tin.

Speaking of your friends...

3. Invite others to benefit from your thing.

I've got a pet snake. Six days of the week, she is just *my* pet snake. On Saturdays, during my famous dinner parties at which the Illuminati congregate, I often pass her around to interested visitors, too. The dinner parties themselves allow my friends to benefit from my stuff, too - kitchen implements and appliances and the very table at which my guests sit. It would be less useful to own a stand mixer or a giant wok if I only ever cooked for myself. It would be less pleasant to have a pet snake if I had no

chance to share her. It would be less important to have pretty clothes if no one ever saw me wearing them.

You're a social ape. If you're trying to get more out of something, an obvious first hypothesis to test is to see if adding other social apes helps:

- Loan your stuff out. (People seem to acquire physical copies of books for this motivation; it is good. Do more of that.)
- Acquire more stuff that can be used cooperatively. (Own games you like, for instance.)
- Find creative ways to use stuff cooperatively where it was not intended.
- Tell people stories about your stuff, if you have interesting stories about it.
- Fetch it when it is a useful tool for someone else's task.
- Accept compliments on your stuff gleefully. Let people have experiences of your stuff so that they will produce same.

Also, circling back to the bit about gifts: I bet you own some gifts. Use them as excuses to think about who gave them to you! My grandmother got me my blender, my mom made me my purse, my best friend gave me the entire signed Fablehaven series. Interacting with those objects now produces extra warmfuzzies if I take the extra cognitive step.

Speaking of how you go about experiencing your stuff...

4. Turn stuff *into* experiences via the senses.

Remember my cloak? It's made of flannel, so it's nice to pet; it's fun to swoosh it about. Remember my snake? She feels nifty and cool and smooth, and she looks pretty, and I get to watch her swallow a mouse once a week if I care to stick around to supervise. I get candy from Trader Joe's because it tastes good and music that I like because it sounds good. If you never look at your stuff or touch it or taste it or whatever is appropriate for the type of stuff, you might not be having it correctly. (Raise your hand if you have chachkas on your shelves that you don't actually *look at*.)

Caveat: Some purely instrumental tools can be had correctly without this - I don't directly experience my Dustbuster with much enthusiasm, just the cleanliness that I can use it to create. Although nothing prevents you from directly enjoying a particularly nice tool either - I have spatulas I am fond of.

And of course if you choose to follow the standard advice about purchasing experiences in a more standard way, you can still use stuff there. You will have more fun camping if you have decent camping gear; you will have more fun at the beach if you have suitable beach things; you will have more fun in the south of France if you have travel guides and phrasebooks that you like.

[1]It's an optional skill. You could neglect it in favor of others, and depending on your own talents and values, this could be higher-leverage than learning to have things correctly. But I bet the following steps will be improvements for some people.

Taking "correlation does not imply causation" back from the internet

(An idea I had while responding to this quotes thread)

"Correlation does not imply causation" is bandied around inexpertly and inappropriately all over the internet. Lots of us hate this.

But get this: the phrase, and the most obvious follow-up phrases like "what does imply causation?" are not high-competition search terms. Up until about an hour ago, the domain name correlationdoesnotimplycausation.com was not taken. I have just bought it.

There is a correlation-does-not-imply-causation shaped space on the internet, and it's ours for the taking. I would like to fill this space with a small collection of relevant educational resources explaining what is meant by the term, why it's important, why it's often used inappropriately, and the circumstances under which one may legitimately infer causation.

At the moment the Wikipedia page is trying to do this, but it's not really optimised for the task. It also doesn't carry the undercurrent of "no, seriously, *lots* of smart people get this wrong; let's make sure you're not one of them", and I think it should.

The purpose of this post is two-fold:

Firstly, it lets me say "hey dudes, I've just had this idea. Does anyone have any suggestions (pragmatic/technical, content-related, pointing out why it's a terrible idea, etc.), or alternatively, would anyone like to help?"

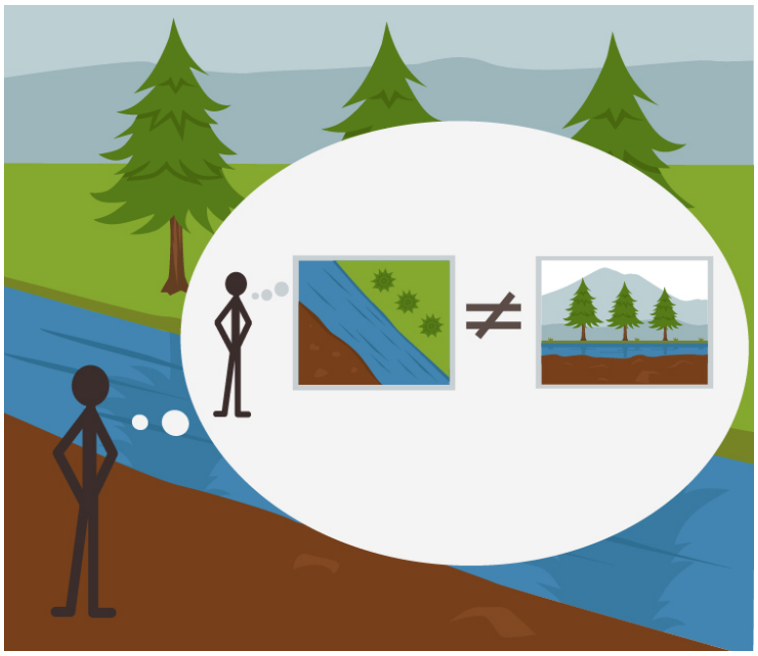
Secondly, it raises the question of what other corners of the internet are ripe for the planting of sanity waterline-raising resources. Are there any other similar concepts that people commonly get wrong, but don't have much of a guiding explanatory web presence to them? Could we put together a simple web platform for carrying out this task in lots of different places? The LW readership seems ideally placed to collectively do this sort of work.

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"

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.*"

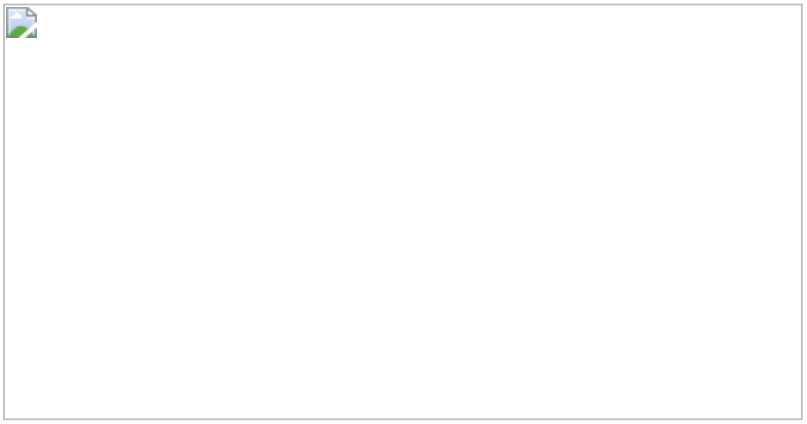
Reply:

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:

$$\forall x : (K(x) \to L(x)) \wedge (L(x) \to I(x))$$

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.

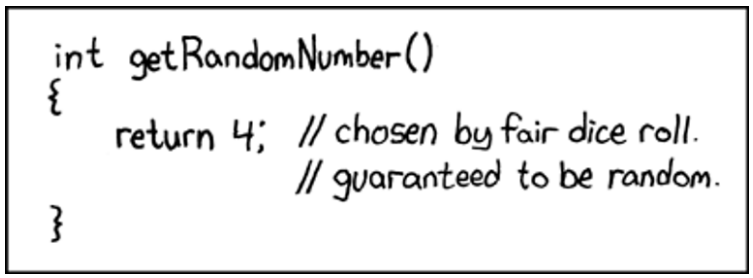
What if you realize that two variables actually *are* correlated more than you thought? What if, to make the diagram correspond to reality, you'd have to hack it to make some U_a and U_b correlated?

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...

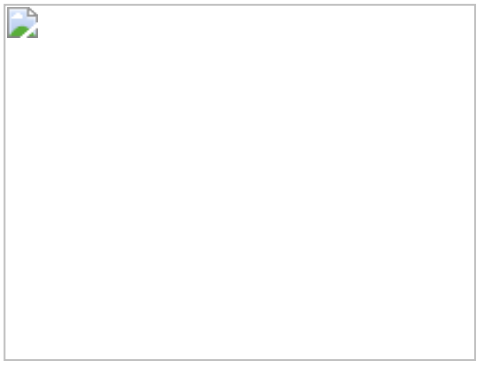
...then we're either going to draw an arrow from the top card to the prediction; an arrow from the prediction to the top card (the prediction makes it happen!); or arrows from a third

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.

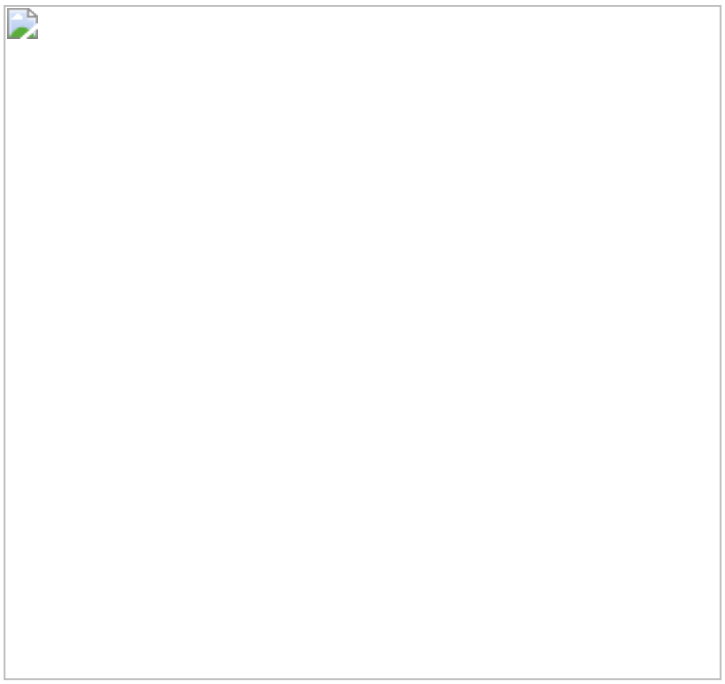
They're about stuff that makes stuff happen, and happens because of other stuff.

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"

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"

Prediction market sequence requested

Related to: Eliezer's Sequences and Mainstream Academia, Intellectual insularity and productivity, I Stand by the Sequences, Why don't people like markets?

Looking at some of the more recent arguments against them showing up in discussions I've been quite disappointed, they seem betray a sort of lack of background knowledge or opinions built up from a bottom line of "markets are baaad therefore prediction markets are baaad". The casual arguments for them are lacking as well. I will say the same of other discussions on economic, since it is apparently suddenly too mind-killing or too political to talk about markets and similar things at all. We didn't use to have tribal alerts flying up in our brains discussing such matters.

The Overcoming Bias community started with an assumption of certain kinds of background knowledge, this included economics and things like game theory. In the early days of LessWrong/Overcoming Bias Eliezer did a whole sequnece on filling in people on Quantum mechanics which despite his claims to the contrary doesn't seem *that* vital (if still important).

We now have a different demographic that we used to. Not only that, we now have young people basically using the sequences as their primary source for education on matters of human rationality, quite different from the autodidacts exploring the literature on their own terms who where common in previous years. We've recognized this to a certain extent. We wrote a series of introductory sequences and articles to fill in such background knowledge explicitly such as Yvain's recent one on Game Theory. Also part of the reason we now have a norm of more citations that EY originally did is to give study and research aids to people. Indeed I think adding comments to old articles featuring more citations or editing those in would be wise so as to avoid misconceptions.

I think we need several sequences on economics, and a good one to start would be one systematically investigating prediction markets. To a certain extent just reading Robin Hanson's relevant posts on this topic would do much the same, but unfortunately we don't have an organized series of sequences by him (beyond the tags he uses on his articles). I still hope Karmakaiser or someone else will one day undertake a project of writing up summary articles that organize links to RH's posts into sequences so new members will read them as well.

I'd write these myself but I just don't have a good background in what works and studies influence the positions of early key LW authors on economics and its relevance to rationality. I'm also only beginning my studies in that area since my background is in the hard sciences with only some half-serious opinions formed from Moldbuggian insights and 20th century social science.

[Link] "Fewer than X% of Americans know Y"

How many times have you heard a claim from a somewhat reputable source like "only 28 percent of Americans are able to name one of the constitutional freedoms, yet 52 percent are able to name at least two Simpsons family members"?

Mark Liberman over at Language Log wrote up a post showing how even when such claims are based on actual studies, the methodology is biased to exaggerate ignorance:

> The way it works is that the survey designers craft a question like the following (asked at a time when William Rehnquist was the Chief Justice of the United States):
>
> > "Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers and the like....
> > What about William Rehnquist – What job or political office does he NOW hold?"
>
> The answers to such open-ended questions are recorded — as audio recordings and/or as notes taken by the interviewer — and these records are coded, later on, by hired coders.
>
> The survey designers give these coders very specific instructions about what counts as right and wrong in the answers. In the case of the question about William Rehnquist, the criteria for an answer to be judged correct were mentions of both "chief justice" and "Supreme Court". These terms had to be mentioned explicitly, so all of the following (actual answers) were counted as wrong:
>
> Supreme Court justice. The main one.
> He's the senior judge on the Supreme Court.
> He is the Supreme Court justice in charge.
> He's the head of the Supreme Court.
> He's top man in the Supreme Court.
> Supreme Court justice, head.
> Supreme Court justice. The head guy.
> Head of Supreme Court.
> Supreme Court justice head honcho.
>
> Similarly, the technically correct answer ("Chief Justice of the United States") would also have been scored as wrong (I'm not certain whether it actually occurred or not in the survey responses).

If, every time you heard a claim of the form "Only X% of Americans know Y" you thought "there's something strange about that", then you get 1 rationality point. If you thought "I don't believe that", then you get 2 rationality points.

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?

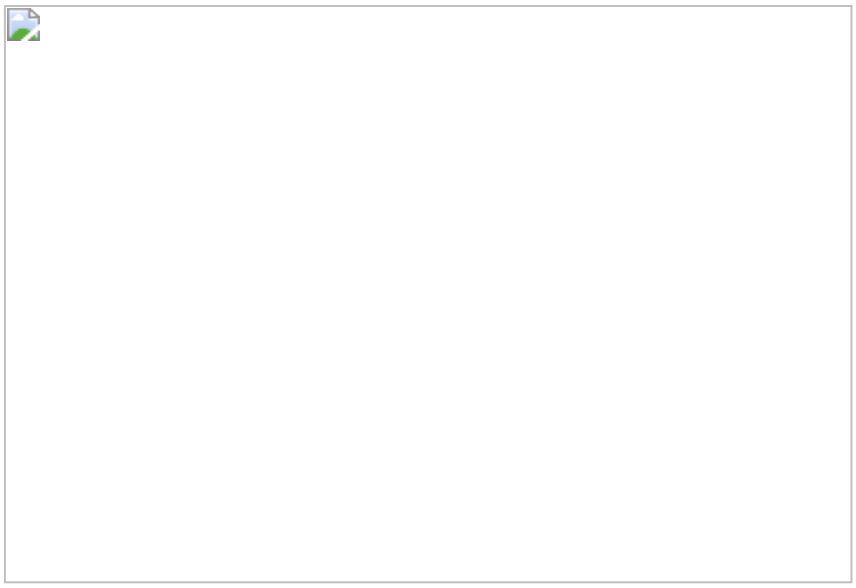
To observe something is to be affected by it - to have your brain and beliefs take on different states, depending on that thing's state. How can you know about something that doesn't affect your brain?

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

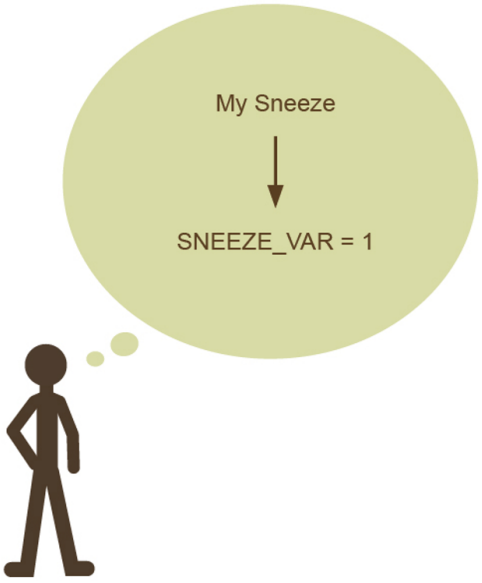
Let's say you're a Lord of the Matrix. You write a computer program which first computes the physical universe as we know it (or a discrete approximation), and then you add a couple of lower-tier effects as follows:

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

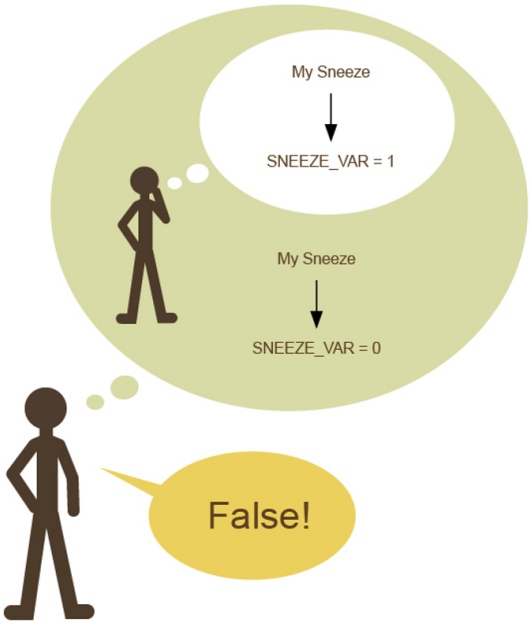
Second, every time I sneeze, the binary variable NO_SNEEZE will be set to the first of its two possible values.



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?"



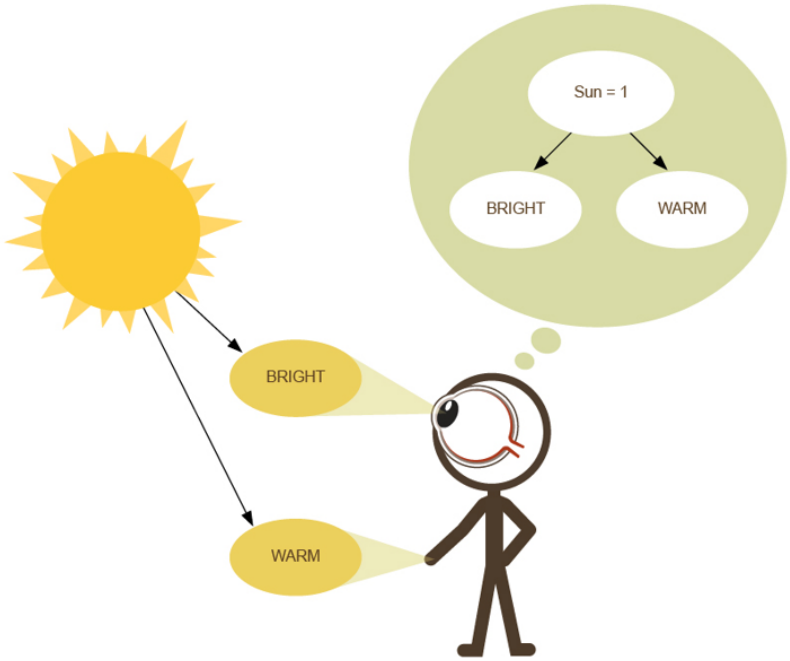
It will be all too easy for me to imagine that this belief is meaningful and could be true or false:



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



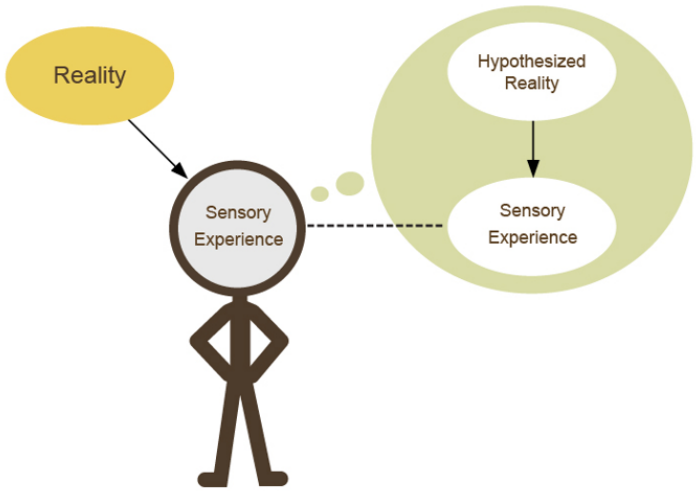
Standing here *outside* the Matrix, we might be tempted to compare your *beliefs* about "Sun = 1", to the real universe's state regarding the visibility of the sun in the sky (or rather, the Earth's rotational position).

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...

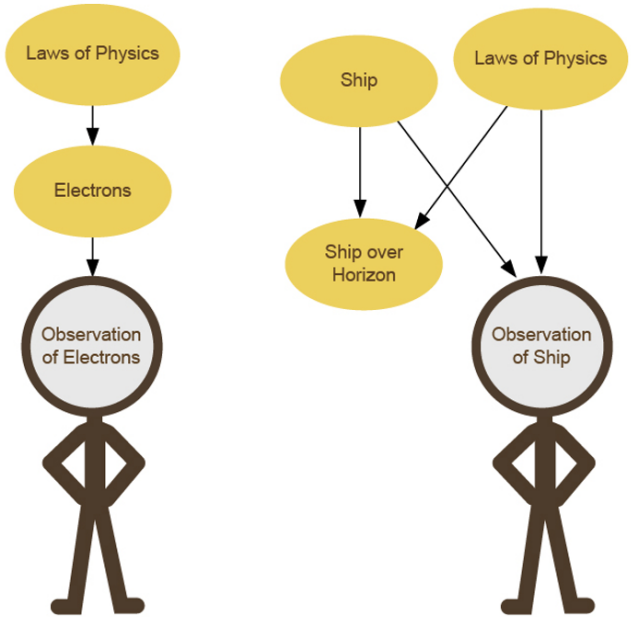


...then the correspondence between map and territory can at least *in principle* be point-wise evaluated by tracing causal links back from sensory experiences to reality, and tracing hypothetical causal links from sensory experiences back to hypothetical reality. We can't directly evaluate that truth-condition 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"

Raising the waterline

Among the goals of Less Wrong is to "raise the sanity waterline" of humanity. We've also talked about "raising the rationality waterline": the phrase is somewhat popular around these parts, which suggests that the metaphor is catchy. But is that all there is to it, a catchy metaphor? Or can the phrase be more usefully cashed out?

While reading Nate Silver's *The Signal and the Noise*, I came across a discussion of "raising the waterline" which fleshes out the metaphor with a more substantial model. This model preserves some of the salient aspects of the metaphor as discussed on LW, for instance the perception that the current waterline (as regards sanity and rationality) is "ridiculously low". More interestingly, it fleshes out some of the specific ways that a "waterline" belief should constrain our future sensory experiences, maybe even to the point of *quantifying* what should result from low (or rising) waterlines.

This is intended as a short series:

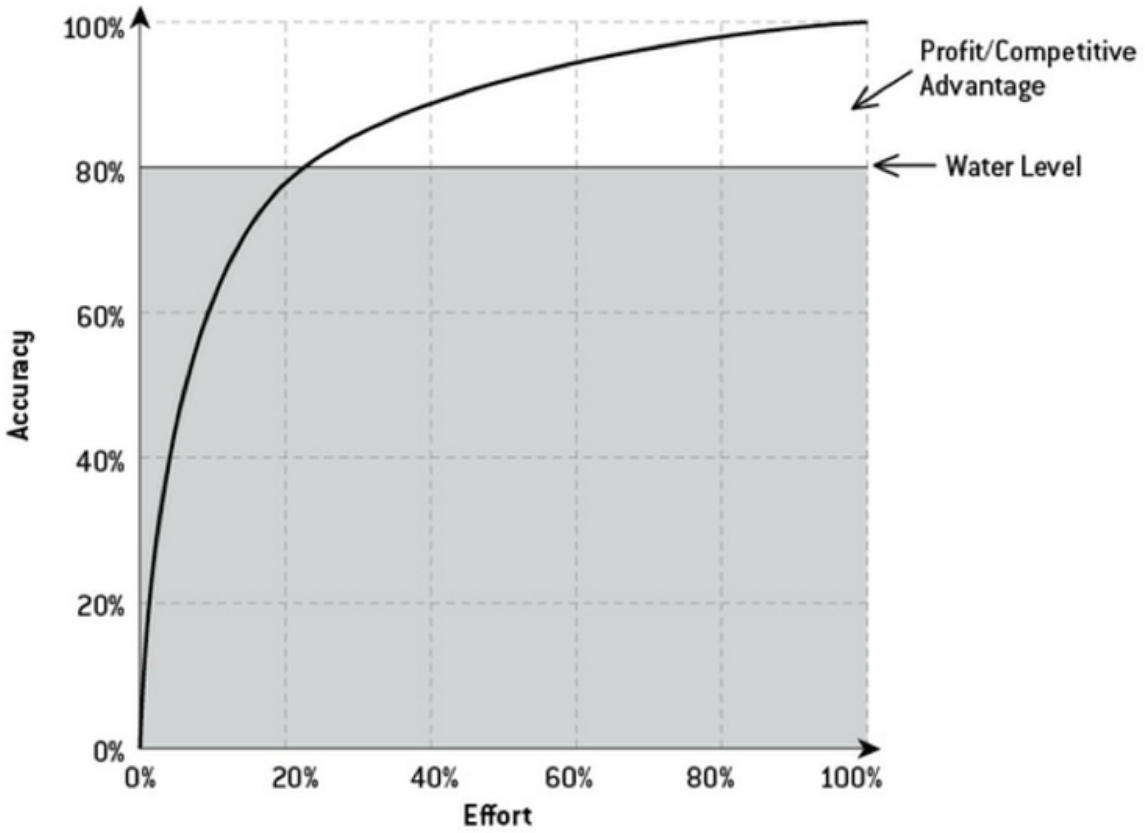
- "**Raising the waterline**", this introductory post, will summarize Nate Silver's "waterline" model, within its original context of playing Poker, which Silver frames as a game of prediction under uncertainty. Poker therefore serves as a "toy model" for a much more general class of problems.
- "**Raising the forecasting waterline**" will extend the discussion to the kind of forecasts studied by Philip Tetlock's Good Judgement Project, a prediction game somewhat similar to PredictionBook and related to prediction markets; I will leverage the waterline model to extract useful insights from my participation in GJP.
- "**Raising the discussion waterline**", a shamelessly speculative coda, will relate the previous two posts to the question of "how do Internet discussions reliably lead to correct inferences from true beliefs, or fail to do so"; I will argue that the waterline model brings some hope that a few basic tactics could nevertheless provide large wins, and raise the more general question of what other low waterlines we could aim to exploit.

The Model

The waterline model is introduced in Chaper 10, "The Poker Bubble", to explain how for a period of time in the 2000's Silver found it fairly easy to make a living from playing online poker, but this source of revenue later dried up altogether.

> I was one of those people. I lived the poker dream for a while, and then it died. I learned that poker sits at the muddy confluence of the signal and the noise. My years in the game taught me a great deal about the role that chance plays in our lives and the delusions it can produce when we seek to understand the world and predict its course.

One graph neatly summarizes all features of the model, and strikes me as a good candidate for illlustrating the "one picture worth a thousand words" dictum:



The horizontal dimension is "effort" or "experience". One possible unit of measurement there could be "hours" - with the caveat that they should be hours of *deliberate* practice. (I'm not entirely sure why this is expressed as a percentage - I'll come back to taking the graph with a grain of salt.)

The vertical dimension is labeled "accuracy", but we could more simply call it "gain". One possible unit of measurement could be "money earned over some period of time playing the game".

Chance, skill and practice

I want to briefly come back to the distinction between experience and practice, as this is a very important but often missed point.

Poker, like all games of chance, driven by random reinforcement, is a highly addictive activity. (In the days following having my interest piqued by Silver's description and wanting to give it a try, I found myself losing more hours to the game than I care to think about - and I was only playing computer opponents for virtual chips, poker's methadone compared to the crack-like properties of online play for real money [1].) It is entirely possible to spend a lot of time in actual play without ever having much to show for it in terms of improvement.

Thus, a less easily measurable but more appropriate construct for the horizontal axis would be something like "number of basic insights absorbed, in the appropriate order". One way to operationalize this would be to devise a number of tests, for

instance, and apply these tests externally to the behaviour of a player: do they fold under circumstance X, raise under circumstance Y, are they able to quantify this or that aspect of the game?

This distinction is well-known in other domains, such as software engineering, where "10 years of experience is not the same as one year of experience repeated 10 times" turns out to be a useful mantra in hiring situations.

Poker's Pareto Principle

The most interesting features of the model are the curve itself, relating effort and gain; and the "waterline".

The curve isn't linear, but follows the same shape as a Pareto distribution: it obeys the "80/20"principle most often associated with Pareto's original observation in the domain of economics - twenty percent of the population holds eighty percent of the wealth. Here, the idea is that twenty percent of the effort is enough to get you at a level of performance better than eighty percent of the population - not bad!

> In poker, for instance, simply learning to fold your worst hands, bet your best ones, and make some effort to consider what your opponent holds will substantially mitigate your losses. If you are willing to do this, then perhaps 80 percent of the time you will be making the same decision as one of the best poker players like Dwan—even if you have spent only 20 percent as much time studying the game.

Silver dubs this the "Pareto Principle of Prediction" - it applies well beyond poker, to a large class of activities based on skills that require deliberate practice.

Raising the waterline

The curve divides into three parts: in the first part, progress is very rapid, disproportionate to the amount of effort you put in. In the middle, you are "grinding" - progress is steadier, requiring the accumulation and honing of a number of distinct techniques. Finally, as you near the top of the performance curve, ever-smaller gains in performance require ever-greater refinement of your existing skills and acquisition of subtle nuances of technique. This is the ten-thousand hour domain, that of "mastery".

The "waterline" represents the typical level of performance that you can expect to see in the player population. Silver represents it as a horizontal line, so we need to think of it as a "gain" level - the typical (say, median) poker player is earning (or losing) the amount of money per time period implied by the particular position of the waterline.

The key idea in the waterline is that in many cases it's not how well you do in an absolute sense that matters - it's how well you do *relative to the competition*. This is especially true in a zero-sum game! If the waterline is low, then you can make very handsome gains at the cost of a limited investment in acquiring the basic skills. But if the waterline is high, there is no alternative, before you can beat the competition, to grinding your way through the lower insights, until you finally level up enough:

> When a field is highly competitive, it is only through this painstaking effort around the margin that you can make any money. There is a “water level” established by the competition and your profit will be like the tip of an iceberg: a small sliver of competitive advantage floating just above the surface, but concealing a vast bulwark of effort that went in to support it.

(This, Silver argues, is what happened to online poker after 2006 and the Unlawful Internet Gambling Enforcement Act. The professional or semi-professional players, who derived an actual income stream from the game, continued playing, but the weaker amateurs quit, in the face of a tougher environment. Bad for poker, maybe; good for rationalists, at least insofar as it led Silver to start work on his election forecasting site FiveThirtyEight.com which has debunked a fair amount of madness about political polling, and ultimately to his book, which may introduce a broader population to largely Less Wrong compatible ideas.)

The (somewhat) bad news

Another important caveat here - this is, as far as I can tell, purely a conceptual model: I'm not aware that there's much hard empirical data that supports the curve having the shape pictured above. Silver does have statistics to show that in the case of poker, the population of mediocre players has a key role in "feeding" the better players, and based on before-and-after numbers, makes a good case that bad players drying up is bad for the better players. However, he doesn't cite anything that would suggest the precise relationship between effort and gain has been measured and shown to fit the curve.

The good news

This model and its more fleshed-out description of what a "ridiculously low" waterline entails are great news: it gives us some testable predictions. Suppose you find yourself in a competitive situation; conditional on the Pareto principle being applicable there, if you notice that applying a short set of uncomplicated techniques reliably results in outperforming your peers, that constitutes some evidence of a low waterline.

For instance, if it is indeed the case that the waterline is very low in the skill of "thinking probabilistically", then those of us who have heeded the lessons of Less Wrong should be able to perform very well in a competitive forecasting environment, by applying only a few basic tools that unfortunately (for them) the general population doesn't yet possess.

Failure to observe this would also have *interesting* consequences, raising the strength of alternate hypotheses: for instance a) we're not as good at applying even those basic techniques as we thought we were, or b) this particular competitive domain is not after all governed by a Pareto distribution, or c) the waterline is not as low as we thought it was.

It turns out I was able to get some experience in just the kind of setting where this type of test could be performed: the Good Judgment Project. That's my next post.

[1] This post is one way I hope to redeem those hours, turning them into a more productive effort after the fact.

P.S.: as this has been asked elsewhere, would I recommend Nate Silver's book? I certainly enjoyed it a lot, even though not all of it was new to me, and I often found myself wishing for clearer separation between the undoubtedly fascinating stories he tells - about financial panics, stock market crashes, baseball, poker, presidential elections, and so on - and the insights he draws from them, such as the waterline model. But enjoying it isn't quite the same as being ready to endorse it to others; I'm not quite sure yet how much value it would have, depending on the audience. Veteran Less Wrongers might not respond to it the same way as the general public, for instance. I found it valuable enough that I might invest some time in writing a short chapter-by-chapter summary and overall review, to answer that question for myself.

Raising the forecasting waterline (part 1)

Previously: Raising the waterline, **see also:** 1001 PredictionBook Nights (LW copy), Techniques for probability estimates

Low waterlines imply that it's relatively easy for a novice to outperform the competition. (In poker, as discussed in Nate Silver's book, the "fish" are those who can't master basic techniques such as folding when they have a poor hand, or calculating even roughly the expected value of a pot.) Does this apply to the domain of making predictions? It's early days, but it looks as if a smallish set of tools - a conscious status quo bias, respecting probability axioms when considering alternatives, considering references classes, leaving yourself a line of retreat, detaching from sunk costs, and a few more - can at least place you in a good position.

A bit of backstory

Like perhaps many LessWrongers, my first encounter with the notion of calibrated confidence was "A Technical Explanation of Technical Explanation". My first serious stab at publicly expressing my own beliefs as quantified probabilities was the Amanda Knox case - an eye-opener, waking me up to how everyday opinions could correspond to degrees of certainty, and how these had *consequences*. By the following year, I was trying to improve my calibration for work-related purposes, and playing with various Web sites, like PredictionBook or Guessum (now defunct).

Then the Good Judgment Project was announced on Less Wrong. Like several of us, I applied, unexpectedly got in, and started taking forecasting more seriously. (I tend to apply myself somewhat better to learning when there is a competitive element - not an attitude I'm particularly proud of, but being aware of that is useful.)

The GJP is both a contest and an experimental study, in fact a group of related studies: several distinct groups of researchers (1,2,3,4) are being funded by IARPA to each run their own experimental program. Within each, small or large number of participants have been recruited, allocated to different experimental conditions, and encouraged to compete with each other (or even, as far as I know, for some experimental conditions, collaborate with each other). The goal is to make predictions about "world events" - and if possible to get them more right, collectively, than we would individually.[1]

Tool 1: Favor the status quo

The first hint I got that my approach to forecasting needed more explicit thinking tools was a blog post by Paul Hewitt I came across late in the first season. My scores in that period (summer 2011 to spring 2012) had been decent but not fantastic; I ended up 5th on my team, which itself placed quite modestly in the contest.

Hewitt pointed out that in general, you could do better than most other forecasters by favoring the status quo outcome.[2] This may not quite be on the same order of effectiveness as the poker advice to "err on the side of folding mediocre hands more

often", but it makes a lot of sense, at least for the Good Judgment Project (and possibly for many of the questions we might worry about). Many of the GJP questions refer to possibilities that loom large in the media at a given time, that are highly available - in the sense of the availability heuristic. This results in a tendency to favor forecasts of change from status quo.

For instance, one of the Season 1 questions was "Will Marine LePen cease to be a candidate for President of France before 10 April 2012?" (also on PredictionBook). Just because the question is being asked doesn't mean that you should assign "yes" and "no" equal probabilities of 50%, or even close to 50%, any more than you should assign 50% to the proposition "I will win the lottery".

Rather, you might start from a relatively low prior probability that anyone who undertakes something as significant as a bid for national presidency would throw in the towel before the contest even starts. Then, try to find evidence that positively favors a change. In this particular case, there was such evidence - the National Front, of which she was the candidate, consistently reports difficulties rounding up the endorsements required to register a candidate legally. However, only once in the past (1981) had this resulted in their candidate being barred (admittedly a very small sample). It would have been a mistake to weigh that evidence excessively. (I got a good score on that question, compared to the team, but definitely owing to a "home ground advantage" as a French citizen rather than my superior forecasting skills.)

Tool 2: Flip the question around

The next technique I try to apply consistently is respecting the axioms of probability. If the probability of event **A** is 70%, then the probability of **not-A** is 30%.

This may strike everyone as obvious... it's not. In Season 2, several of my team-mates are on record as assigning a 75% probability to the proposition "The number of registered Syrian conflict refugees reported by the UNHCR will exceed 250,000 at any point before 1 April 2013".

That number was reached today, six months in advance of the deadline. This was clear as early as August. The trend in the past few months has been an increase of 1000 to 2000 a day, and the UNHCR have recently provided estimates that this number will eventually reach 700,000. The kicker is that this number is *only* the count of people who are fully processed by the UNHCR administration and officially in their database; there are tens of thousands more in the camps who only have "appointments to be registered".

I've been finding it hard to understand why my team-mates haven't been updating to, maybe not 100%, but at least 99%; and how one wouldn't see these as the only answers worth considering. At any point in the past few weeks, to state your probability as 85% or 91% (as some have quite recently) was to say, "There is *still* a one in ten chance that the Syrian conflict will suddenly stop and all these people will go home, maybe next week?."

This is kind of like saying "There is a one in ten chance Santa Claus will be the one distributing the presents this year." It feels like a huge "clack".

I can only speculate as to what's going on there. Queried for a probability, people are translating something like "Sure, **A** is happening" into a biggish number, and reporting

that. They are totally failing to flip the question around and explicitly consider what it would take for **not-A** to happen. (Perhaps, too, people have been so strongly cautioned by cautions, from Tetlock and others, against being overconfident that they reflexively shy away from the extreme numbers.)

Just because you're expressing beliefs as percentages doesn't mean that you are automatically applying the axioms of probability. Just because you use "75%" as a shorthand for "I'm pretty sure" doesn't mean you are thinking probabilistically; you must train the skill of seeing that for some events, its complement "25%" also counts as "I'm pretty sure". The axioms are more important than the use of numbers - in fact for this sort of forecast "91%" strikes me as needlessly precise; increments of 5% are more than enough, away from the extremes.

Tool 3: Reference class forecasting

The order in which I'm discussing these "basics of forecasting" reflects not so much their importance, as the order in which I tend to run through them when encountering a new question. (This might not be the optimal order, or even very good - but that should matter little if the waterline is indeed low.)

Using reference classes was actually part of the "training package" of the GJP. From the linked post comes the warning that "deciding what's the proper reference class is not straightforward". And in fact, this tool only applies in some cases, not systematically. One of our recently closed questions was "Will any government force gain control of the Somali town of Kismayo before 1 November 2012?". Clearly, you could spend quite a while trying to figure out an appropriate reference class here. (In fact, this question also stands as a counter-example to the "Favor status quo" tool, and flipping the question around might not have been too useful either. All these tools require some discrimination.)

On the other hand, it came in rather handy in assessing the short-term question we got late september: "What change will occur in the FAO Food Price index during September 2012?" - with barely two weeks to go before the FAO was to post the updated index in early October. More generally, it's a useful tool when you're asked to make predictions regarding a numerical indicator, for which you can observe past data.

The FAO price data can be retrieved as a spreadsheet (.xsl download). Our forecast question divided the outcomes into four: **A**) an increase of 3% or more, **B**) an increase of less than 3%, **C**) a decrease of less than 3%, **D**) a decrease of more than 3%, **E**) "no change" - meaning a change too small to alter the value rounded to the nearest integer.

It's not clear from the chart that there is any consistent seasonal variation. A change of 3% would have been about 6.4 points; since 8/2011 there had been four month-on-month changes of that magnitude, 3 decreases and 1 increase. Based on that reference class, the probability of a small change (B+C+E) came out to about 2/3. The probability for "no change" (E) to 1/12 - the August price was the same as the July price. The probability for an increase (A+B), roughly the same as for a decrease (C+D). My first-cut forecast allocated the probability mass as follows: 15/30/30/15/10.

However, I figured I did need to apply a correction, based on reports of a drought in the US that could lead to some food shortages. I took 10% probability mass from the

"decrease" outcomes and allocated it to the "increase" outcomes. My final forecast was 20/35/25/10/10. I didn't mess around with it any more than that. As it turned out, the actual outcome was **B**! My score was bettered by only 3 forecasters, out of a total of 9.

Next up: lines of retreat, ditching sunk costs, loss functions

This post has grown long enough, and I still have 3+ tools I want to cover. Stay tuned for Part 2!

---

[1] The GJP is being run by Phil Tetlock, known for his "hedgehog and fox" analysis of forecasting. At that time I wasn't aware of the competing groups - one of them, DAGGRE, is run by Robin Hanson (of OB fame) among others, which might have made it an appealing alternate choice if I'd know about it.

[2] Unfortunately, the experimental condition Paul belonged to used a prediction market where forecasters played virtual money by "betting" on predictions; this makes it hard to translate the numbers he provides into probabilities. The general point is still interesting.