Best of LessWrong: May 2017

1. Gears in understanding

Best of LessWrong: May 2017

1. Gears in understanding

Gears in understanding

Some (literal, physical) roadmaps are more useful than others. Sometimes this is because of how well the map corresponds to the territory, but sometimes it's because of features of the map that are irrespective of the territory. E.g., maybe the lines are fat and smudged such that you can't tell how far a road is from a river, or maybe it's unclear which road a name is trying to indicate.

In the same way, I want to point at a property of models that isn't about what they're modeling. It *interacts with* the clarity of what they're modeling, but only in the same way that smudged lines in a roadmap interact with the clarity of the roadmap.

This property is *how deterministically interconnected the variables of the model are*. There are a few tests I know of to see to what extent a model has this property, though I don't know if this list is exhaustive and would be a little surprised if it were:

1. Does the model pay rent? If it does, and if it were falsified, how much (and how precisely) could you infer other things from the falsification?
2. How incoherent is it to imagine that the model is accurate but that a given variable could be different?
3. If you knew the model were accurate but you were to forget the value of one variable, could you rederive it?

I think this is a really important idea that ties together a lot of different topics that appear here on Less Wrong. It also acts as a prerequisite frame for a bunch of ideas and tools that I'll want to talk about later.

I'll start by giving a bunch of examples. At the end I'll summarize and gesture toward where this is going as I see it.

Example: Gears in a box

Let's look at this collection of gears in an opaque box:

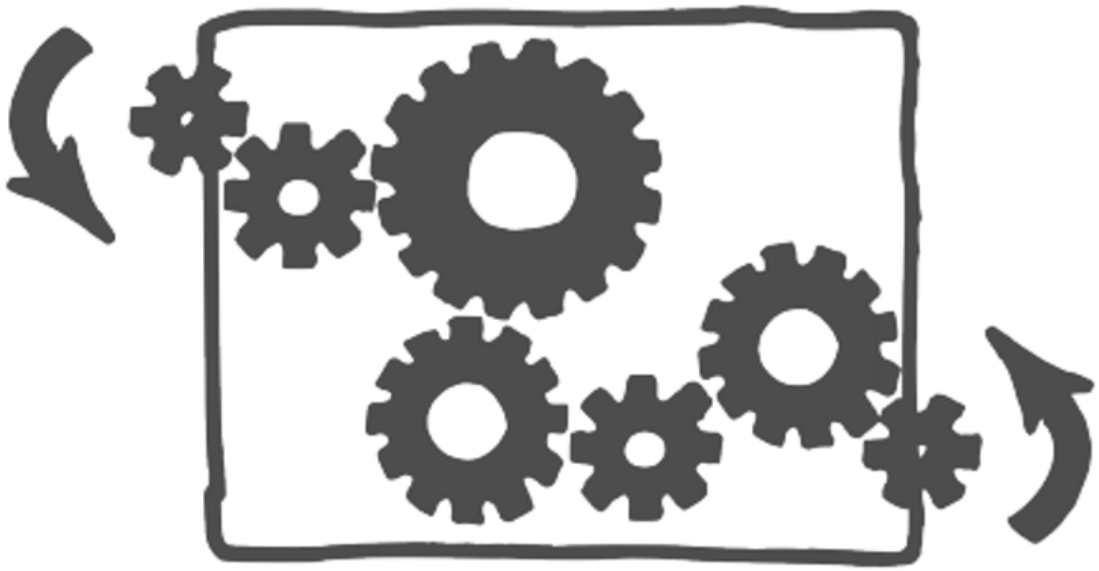
(Drawing courtesy of my colleague, Duncan Sabien.)

If we turn the lefthand gear counterclockwise, it's within our model of the gears on the inside that the righthand gear could turn either way. The model we're able to build for this system of gears does poorly on all three tests I named earlier:

- The model barely pays rent. If you speculate that the righthand gear turns one way and you discover it turns the other way, you can't really infer very much. All you can meaningly infer is that *if* the system of gears is pretty simple (e.g., nothing that makes the righthand gear alternate as the lefthand gear rotates counterclockwise), then the direction that the righthand gear turns determines whether the total number of gears is even or odd.
- The gear on the righthand side could just as well go either way. Your expectations aren't constrained.
- Right now you *don't* know which way the righthand gear turns, and you can't derive it.

Suppose that Joe peeks inside the box and tells you "Oh, the righthand gear will rotate clockwise." You imagine that Joe is more likely to say this if the righthand gear turns clockwise than if it doesn't, so this seems like relevant evidence that the righthand gear turns clockwise. This gets stronger the more people like Joe who look in the box and report the same thing.

Now let's peek inside the box:



…and now we have to wonder what's up with Joe.

The second test stands out for me especially strongly. There is *no way* that the obvious model about what's going on here could be right *and* Joe is right. And it doesn't matter *how many people* agree with Joe in terms of the logic of this statement: Either *all of them are wrong*, or your model is wrong. This logic is immune to social pressure. It means that there's a chance that you can accumulate evidence about how well your map matches the territory here, and if that converges on your map being basically correct, then you are on firm epistemic footing to disregard the opinion of *lots* of other people. Gathering evidence about the map/territory correspondence has higher leverage for seeing the truth than does gathering evidence about what others think.

The first test shows something interesting too. Suppose the gear on the right really *does* move clockwise when you move the left gear counterclockwise. What does that imply? Well,

it means your initial model (if it's what I imagine it is) is wrong — but there's a limited space of possibilities about *ways in which* it can be wrong. For instance, maybe the second gear from the left is on a vertical track and moves upward instead of rotating. By comparison, something like "Gears work in mysterious ways" just won't cut it.

If we combine the two, we end up staring at Joe and noticing that we can be a lot more precise than just "Joe is wrong". We know that either Joe's model of the gears is wrong (e.g., he thinks some gear is on a vertical track), Joe's model of the gears is *vague* and isn't constrained the way ours is (e.g., he was just counting gears and made a mistake), or Joe is lying. The first two give us testable predictions: If his model is wrong, then it's wrong *in some specific way*; and if it's vague, then there should be some place where it does poorly on the three tests of model interconnectedness. If we start zooming in on these two possibilities while talking to Joe and it turns out that neither of those are true, then it becomes a lot more obvious that Joe is just bullshitting (or we failed to think of a fourth option).

Because of this example, in CFAR we talk about how "Gears-like" or how "made of Gears" a model is. (I capitalize "Gears" to emphasize that it's an analogy.) When we notice an interconnection, we talk about "finding Gears". I'll use this language going forward.

Example: Arithmetic

If you add 25+18 using the standard addition algorithm, you have to carry a 1, usually by marking that above the 2 in 25.

Fun fact: it's possible to get that right without having any clue what the 1 represents or why you write it there.

This is actually a pretty major issue in math education. There's an in-practice tension between (a) memorizing and drilling algorithms that let you compute answers quickly, and (b) "really understanding" why those algorithms work.

Unfortunately, there's a kind of philosophical debate that often happens in education when people talk about what "understand" means, and I find it pretty annoying. It goes something like this:

- Person A: "The student said they carry the 1 because that's what their teacher told them to do. So they don't really understand the addition algorithm."
- Person B: "What do you mean by reallyunderstand'?Wˆ'swrongwiththejustificationofA person who knows this subject really well says this works, and I believe them'?"
- A: "But that reason isn't about the *mathematics*. Their justification isn't mathematical. It's social."
- B: "Mathematical justification is social. The style of proof that topologists use wouldn't be accepted by analysts. What constitutes a 'proof' or a 'justification' in math is socially agreed upon."
- A: "Oh, come on. We can't just agree that e=3 and make that true. Sure, maybe the way we *talk* about math is socially constructed, but we're talking about something real."
- B: "I'm not sure that's true. But even if it were, how could you know whether you're talking about that `something real' as opposed to one of the social constructs we're using to share perspectives about it?"

Et cetera.

(I would *love* to see debates like this happen in a milieu of mutual truth-seeking. Unfortunately, that's not what academia rewards, so it probably isn't going to happen there.)

I think Person A is trying to gesture at the claim that the student's model of the addition algorithm isn't made of Gears (and implicitly that it'd be better if it were). I think this clarifies

both what A is saying and why it matters. In terms of the tests:

- The addition algorithm totally pays rent. E.g., if you count out 25 tokens and another 18 tokens and you then count the total number of tokens you get, that number should correspond to what the algorithm outputs. If it turned out that the student does the algorithm but the answer *doesn't* match the token count, then the student can only conclude that the addition algorithm isn't useful for the tokens. There isn't a lot else they can deduce. (By way of contrast, if I noticed this, then I'd conclude that either I'd made a mistake in running the algorithm or I'd made a mistake in counting, and I'd be very confident that at least one of those two things is true.)
- The student could probably readily imagine a world in which you *aren't* supposed to carry the 1 but the algorithm still works. This means their model isn't very constrained, at least as we're imagining it. (Whereas attempting to think about carrying being *wrong* to do for getting the right answer makes my head explode.)
- If the student forgot what their teacher said about what to do when a column adds up to more than nine, we imagine they wouldn't spontaneously notice the need to carry the 1. (If I forgot about carrying, though, I'd get confused about what to do with this extra ten and would come up with something mathematically equivalent to "carrying the 1".)

I find this to be a useful tabooing of the word "understand" in this context.

Example: My mother

My mother really likes learning about history.

Right now, this is probably an unattached random fact in your mind. Maybe a month down the road I could ask you "How does my mother feel about learning history?" and you could try to remember the answer, but you could just as well believe the world works another way.

But for me, that's not true at all. If I forgot her feelings about learning about history, I could make a pretty educated guess based on my overall sense of her. I wouldn't be Earth-shatteringly *shocked* to learn that she doesn't like reading about history, but I'd be really confused, and it'd throw into question my sense of why she likes working with herbs and why she likes hanging out with her family. It would make me think that I hadn't quite understood what kind of person my mother is.

As you might have noticed, this is an application of tests 1 and 3. In particular, my model of my mom isn't *totally made of* Gears in the sense that I could tell you what she's feeling right now or whether she defaults to thinking in terms of partitive division or quotative division. But the tests illustrate that my model of my mother is *more* Gears-like than *your* model of her.

Part of the point I'm making with this example is that "Gears-ness" isn't a binary property of models. It's more like a spectrum, from "random smattering of unconnected facts" to "clear axiomatic system with well-defined logical deductions". (Or at least that's how I'm currently imagining the spectrum!)

Also, I speculate that this is part of what we mean when we talk about "getting to know" someone: it involves increasing the Gears-ness of our model of them. It's not about just getting some isolated facts about where they work and how many kids they have and what they like doing for hobbies. It's about fleshing out an *ability to be surprised* if you were to learn some new fact about them that didn't fit your model of them.

(There's also an empirical question in getting to know someone of how well your Gears-ish model *actually matches* that person, but that's about the map/territory correspondence. I want to be careful to keep talking about properties of maps here.)

This lightly Gears-ish model of people is what I think lets you deduce what Mr. Rogers probably would have thought about, say, people mistreating cats on Halloween even though I don't know if he ever talked about it. As per test #2, you'd probably be pretty shocked and confused if you were given compelling evidence that he had *joined in*, and I imagine it'd take a *lot* of evidence. And then you'd have to update a *lot* about how you view Mr. Rogers (as per test #1). I think a lot of people had this kind of "Who even *is* this person?" experience when lots of criminal charges came out against Bill Cosby.

Example: Gyroscopes

Most people feel visceral surprise when they watch how gyroscopes behave. Even if they *logically know* the suspended gyroscope will rotate instead of falling, they usually *feel* like it's bizarre somehow. Even people who get gyroscopes' behavior into their intuitions probably had to train it for a while first and found them surprising and counterintuitive.

Somehow, for most people, it seems coherent to imagine a world in which physics works exactly the same way *except that* when you suspend one end of a gyroscope, it falls like a non-spinning object would and just keeps spinning.

If this is true of you, that means your model of the physics around gyroscopes does poorly on test #2 of how Gears-like it is.

The *reason* gyroscopes do what they do is actually something you can derive from Newton's Laws of Motion. Like the gears example, you can't actually have a coherent model of rotation that allows (a) Newton's Laws and (b) a gyroscope that *doesn't* rotate instead of falling when suspended on one end in a gravitational field. So if both (a) and (b) seem plausible to you, then your model of rotation isn't coherent. It's missing Gears.

This is one of the beautiful (to me) things about physics: *everything* is made of Gears. Physics is (I think) the system of Gears you get when you stare at any physical object's behavior and ask "What makes you do that?" in a Gears-seeking kind of way. It's a different level of abstraction than the "Gears of people" thing, but we kind of expect that *eventually*, at least in theory, a sufficient extension of physics will connect the Gears of mechanics to the Gears of what makes a romantic relationship last while feeling good to be in.

I want to rush to clarify that I'm *not* saying that *the world* is made of Gears. That's a type error. I'm suggesting that *the property of Gears-ness in models* is tracking a true thing about the world, which is why making models more Gears-like can be so powerful.

Gears-ness is not the same as goodness

I want to emphasize that, while I think that more Gears are better all else being equal, there are other properties of models that I think are worthwhile.

The obvious one is accuracy. I've been intentionally sidestepping that property throughout most of this post. This is where the rationalist virtue of empiricism becomes critical, and I've basically ignored (but hopefully never defied!) empiricism here.

Another is *generativity*. Does the model *inspire a way of experiencing* in ways that are useful (whatever "useful" means)? For instance, many beliefs in God or the divine or similar are too abstract to pay rent, but some people still find them helpful for reframing how they emotionally experience beauty, meaning, and other people. I know of a few ex-atheists who say that having *become* Christian causes them to be nicer people and has made their relationships better. I think there's reason for epistemic fear here to the extent that those religious frameworks sneak in claims about how the world actually works — but if you're epistemically careful, it seems possibly worthwhile to explore how to tap the power of faith without taking epistemic damage.

I also think that even if you're trying to lean on the Gears-like power of a model, lacking Gears doesn't mean that the activity is worthless. In fact, I think this is *all we can do* most of the time, because most of our models don't connect all the way down to physics. E.g., I'm thinking of getting my mother a particular book as a gift because I think she'll really like it, but I can *also* come up with a within-my-model-of-her story about why she might *not* really care about it. I don't think the fact that my model of her is weakly constrained means that (a) I shouldn't use the model or that (b) it's not worthwhile to explore the "why" behind both my being right and my being wrong. (I think of it as a bit of pre-computation: whichever way the world goes, my model becomes a little more "crisp", which is to say, more Gears-like. It just so happens that I know *in what way* beforehand.)

I mention this because sometimes in rationalist contexts, I've felt a pressure to *not talk about* models that are missing Gears. I don't like that. I think that Gears-ness is a *really super important* thing to track, and I think there's something epistemically dangerous about *failing to notice* a lack of Gears. Clearly noting, at least in your own mind, where there are and aren't Gears seems really good to me. But I think there are *other* capacities that are *also* important when we're trying to get epistemology right.

Gears seem valuable to me *for a reason*. I'd like us to keep that reason in mind rather than getting too fixated on Gears-ness.

Going forward

I think this frame of Gears-ness of models is super powerful for cutting through confusion. It helps our understanding of the world become immune to social foolishness and demands a kind of rigor to our thinking that I see as unifying lots of ideas in the Sequences.

I'll want to build on this frame as I highlight other ideas. In particular, I haven't spoken to *how we know Gears are worth looking for*. So while I view this as a powerful weapon to use in our war against sanity drought, I think it's *also* important to examine the smithy in which it was forged. I suspect that won't be my *very next* post, but it's one I have in mind upcoming.