Best of LessWrong: November 2013

1. On learning difficult things
2. 2013 Less Wrong Census/Survey
3. Wait vs Interrupt Culture
4. No Universally Compelling Arguments in Math or Science
5. Yes, Virginia, You Can Be 99.99% (Or More!) Certain That 53 Is Prime
6. I notice that I am confused about Identity and Resurrection

Best of LessWrong: November 2013

1. On learning difficult things
2. 2013 Less Wrong Census/Survey
3. Wait vs Interrupt Culture
4. No Universally Compelling Arguments in Math or Science
5. Yes, Virginia, You Can Be 99.99% (Or More!) Certain That 53 Is Prime
6. I notice that I am confused about Identity and Resurrection

On learning difficult things

I have been autodidacting quite a bit lately. You may have seen my reviews of books on the MIRI course list. I've been going for about ten weeks now. This post contains my notes about the experience thus far.

Much of this may seem obvious, and would have seemed obvious if somebody had told me in advance. But nobody told me in advance. As such, this is a collection of things that were somewhat surprising at the time.

Part of the reason I'm posting this is because I don't know a lot of autodidacts, and I'm not sure how normal any of my experiences are. (Though on average, I'd guess they're about average.) As always, keep in mind that I am only one person and that your mileage may vary.

Pair up

When I began my quest for more knowledge, I figured that in this modern era, a well-written textbook and an account on math.stackexchange would be enough to get me through anything. And I was right... sort of.

But not really.

The problem is, most of the time that I get stuck, I get stuck on something incredibly stupid. I've either misread something somewhere or misremembered a concept from earlier in the book. Usually, someone looking over my shoulder could correct me in ten seconds with three words.

"Dude. Disjunction. *Dis*junction."

These are the things that eat my days.

In principle, places like stackexchange can get me unstuck, but they're an awkward tool for the job. First of all, my stupid mistakes are heavily contextualized. A full context dump is necessary before I can even ask my question, and this takes time. Furthermore, I feel dumb asking stupid questions on stackexchange-type sites. My questions are usually things that I can figure out with a close re-read (except, I'm not sure which part needs a re-read). I usually opt for a close re-read of everything rather than asking for help. This is even more time consuming.

The infuriating thing is that answering these questions usually doesn't require someone who already knows the answers: it just requires someone who didn't make exactly the same mistakes as me. I lose hours on little mistakes that could have been fixed within seconds if I was doing this with someone else.

That's why my number one piece of advice for other people attempting to learn on their own is *do it with a friend*. They don't need to be more knowledgeable than you to answer most of the questions that come up. They just need to make *different* misunderstandings, and you'll be able to correct each other as you go along.

The thing I miss most about college is tight feedback loops while learning. When autodidacting, the feedback loop can be long.

I still haven't managed to follow my own advice here. I'm writing this advice in part because it should motivate me to actually pair up. Unfortunately, there is nobody in my immediate circle who has the time or patience to read along with me, but there are a number of resources I have not yet explored (the LessWrong study hall, for example, or soliciting to actual mathematicians). It's on my list of things to do.

Read, reread, rereread

Reading *Model Theory* was one of the hardest things I've done. Not necessarily because the content was hard, but because it was the first time I actually learned something that was way outside my comfort zone.

The short version is that *Basic Category Theory* and *Naïve Set Theory* left me somewhat overconfident, and that I should have read a formal logic textbook before diving in. I had basic familiarity with logic, but no practice. Turns out practice is important.

Anyway, it's not like *Model Theory* was impossible just because I skipped my logic exercises. It was just *hard*. There are a number of little misconceptions you have when you're familiar with something but you've never applied it, and I found myself having to clean those out just to understand what *Model Theory* was trying to say to me.

In retrospect, this was an efficient way to strengthen my understanding of mathematical logic and learn *Model Theory* at the same time. (I've moved on to a logic textbook, and it's been a cakewalk.) That said, I wouldn't wish the experience on others.

In the process, I learned how to learn things that are way outside my comfort zone. In the past, all the stuff I've learned has been either easy, or an extension of things that I was already interested in and experienced with. Reading *Model Theory* was the first time in my life where I read a chapter of a textbook and it made *absolutely no sense*. In fact, it took about three passes per chapter before they made sense.

1. The first pass was barely sufficient to understand all the words and symbols. I constantly had to go research a topic. I followed proofs one step at a time, able to verify the validity of each step but not really understand what was going on. I came out the other end believing the results, but not knowing them.
2. Another pass was required to figure out what the book was actually trying to say to me. Once all the words made sense and I was comfortable with their usage, the second pass allowed me to see what the theorems and proofs were actually saying. This was nice, but it still wasn't sufficient: I understood the theorems, but they seemed like a random walk through theorem-space. I couldn't yet understand why anyone would say those particular things on purpose.
3. The third pass was necessary to understand the greater theory. I've never been particularly good at memorizing things, and it's not sufficient for me to believe and memorize a theorem. If it's going to stick, I have to understand why it's important. I have to understand why this theorem in particular is being stated, rather than another. I have to understand the problem that's being solved. A third pass was necessary to figure out the context in which the text made sense.

After a third pass of any given chapter, the next chapter didn't seem quite so random. When the upcoming content started feeling like a natural progression instead of a random walk, I knew I was making progress.

I note this because this is the first time that I had to read a math text more than once to understand what was going on. I'm not talking about individual sentences or paragraphs, I'm talking about finishing a chapter, feeling like "wat", and then starting the whole chapter over. Twice.

I'm not sure if I'm being naïve (for never having needed to do this before) or slow (for having to do this for *Model Theory*), but I did not anticipate requiring three passes. Mostly, I didn't anticipate gaining as much as I did from a re-read; I would have guessed that something opaque on the first pass would remain opaque on a second pass.

This, I'm pretty sure, was naïvety.

So take note: if you stumble upon something that feels very hard, it might be more useful than anticipated to re-read it.

Cognitive exchange rates

When reading Model Theory, I was only able to convert 30-50% of my allotted "study time" into actual study.

This is somewhat surprising, as I had no such troubles with *Basic Category Theory* or *Naïve Set Theory*.

(I often have the *opposite* problem when writing code; this is probably due to the different reward structure.)

I was somewhat frustrated with my inability to study as much as I would have liked. My usual time-into-studying conversion rate is much higher (I'd guess 80%ish, though I haven't been measuring).

I'm not sure what factor made it harder for me to study model theory. I don't think it was the difficulty directly, as I often tend to work harder in the face of a challenge. I'd guess that it was either the slower rate of rewards (caused by a slower pace of learning) or actual cognitive exhaustion.

In the vein of cognitive exhaustion, there were a few times while reading *Model Theory* where I seem to have become cognitively exhausted before becoming physically exhausted. This was a first for me. I'm not referring to those times when you've done a lot of mental work and you shy away from doing anything difficult, that's happened to me plenty. Rather, in this case, I felt fully awake and ready to keep reading. And I did keep reading. It just… didn't work. I'd have trouble following simple proofs. I'd fail at parsing sentences that were quite clear after resting.

I'm still not sure what to make of this, and I don't have sufficient data to draw conclusions. However, it seems like there are mental states where my I feel awake and able to continue, but my mind is just not capable of doing the heavy lifting.

Again, the fact that I'm only just realizing this now is probably naïvety, but it's something to remember before getting frustrated with yourself.

Explain it to someone

As I've said before, one of the best ways to learn something is to do the problem sets. For *Model Theory*, though, there were times when I finished reading through a chapter and was not capable of doing the problems.

Re-reading helped, as mentioned above. Another thing that helped was explaining the concepts.

I explained model theory pretty extensively to a text file on my computer. I sketched the proofs in my own words and stated their significance. I explained the syntax being used. I tried to motivate each idea. (The notes are still lying around somewhere; I haven't posted them because they're pretty much a derivative work at this point.)

I found that this went a *long* way towards helping me track down places where I'd thought I learned something, but actually hadn't. If you're having trouble, go explain the concept to somebody (or to a text file). This can bridge the gap between "I read it" and "I can do the problems" quite well. For me, this technique often took problems from "unapproachable" to "easy" in one fell swoop.

Don't book yourself solid

I'm pretty good at avoiding stress. I have the (apparently rare) ability to drop all work-related concerns at the door when I leave. I don't even know *how* to get stressed by bad luck, especially if I made good choices given the information I had at the time. I get normally tense in stressful situations with time constraints, but I'm adept at avoiding the permastress that I've seen plague friends and family – unless I've booked myself solid.

I've had a packed schedule these past few weeks. I try to move the needle on at least two projects a day (more on weekends). Even if it's entirely reasonable to fit all these things into my schedule, I have not yet found a way to avoid the stress.

Even when I know that, if I push myself, I can read this much and write that much and code this feature all in one day, I haven't found a good way to push myself without pressure-stress.

I'm still hoping that I'll learn how to move quickly without stress as I learn my capabilities, but I'm not sure I've been adequately accounting for the cost of stress.

It's worth remembering that doing less than you're capable of *on purpose* might be a good strategy for maximizing long-term output.

There you go. Those are my notes gathered from trying to learn lots of things very quickly (and trying to learn one hard thing in particular). Comments are encouraged; I am by no means an expert.

2013 Less Wrong Census/Survey

It's that time of year again.

If you are reading this post, and have not been sent here by some sort of conspiracy trying to throw off the survey results, then you are the target population for the Less Wrong Census/Survey. Please take it. Doesn't matter if you don't post much. Doesn't matter if you're a lurker. Take the survey.

This year's census contains a "main survey" that should take about ten or fifteen minutes, as well as a bunch of "extra credit questions". You may do the extra credit questions if you want. You may skip all the extra credit questions if you want. They're pretty long and not all of them are very interesting. But it is very important that you not put off doing the survey or not do the survey at all because you're intimidated by the extra credit questions.

It also contains a chance at winning a MONETARY REWARD at the bottom. You do not need to fill in all the extra credit questions to get the MONETARY REWARD, just make an honest stab at as much of the survey as you can.

Please make things easier for my computer and by extension me by reading all the instructions and by answering any text questions in the simplest and most obvious possible way. For example, if it asks you "What language do you speak?" please answer "English" instead of "I speak English" or "It's English" or "English since I live in Canada" or "English (US)" or anything else. This will help me sort responses quickly and easily. Likewise, if a question asks for a number, please answer with a number such as "4", rather than "four".

Last year there was some concern that the survey period was too short, or too uncertain. This year the survey will remain open until 23:59 PST December 31st 2013, so as long as you make time to take it sometime this year, you should be fine. Many people put it off last year and then forgot about it, so why not take it right now while you are reading this post?

Okay! Enough preliminaries! Time to take the...

***

2013 Less Wrong Census/Survey

***

Thanks to everyone who suggested questions and ideas for the 2013 Less Wrong Census/Survey. I regret I was unable to take all of your suggestions into account, because of some limitations in Google Docs, concern about survey length, and contradictions/duplications among suggestions. I think I got *most* of them in, and others can wait until next year.

By ancient tradition, if you take the survey you may comment saying you have done so here, and people will upvote you and you will get karma.

Wait vs Interrupt Culture

At the recent CFAR Workshop in NY, someone mentioned that they were uncomfortable with pauses in conversation, and that got me thinking about different conversational styles.

Growing up with friends who were disproportionately male and disproportionately nerdy, I learned that it was a normal thing to interrupt people. If someone said something you had to respond to, you'd just start responding. Didn't matter if it "interrupted" further words – if they thought you needed to hear those words before responding, they'd interrupt right back.

Occasionally some weird person would be offended when I interrupted, but I figured this was some bizarre fancypants rule from before people had places to go and people to see. Or just something for people with especially thin skins or delicate temperaments, looking for offense and aggression in every action.

Then I went to St. John's College – the talking school (among other things). In Seminar (and sometimes in Tutorials) there was a totally different conversational norm. People were always expected to wait until whoever was talking was done. People would apologize not just for interrupting someone who was already talking, but for accidentally saying something when someone else looked like they were about to speak. This seemed totally crazy. Some people would just blab on unchecked, and others didn't get a chance to talk at all. Some people would ignore the norm and talk over others, and nobody interrupted them back to shoot them down.

But then a few interesting things happened:

1) The tutors were able to moderate the discussions, gently. They wouldn't actually scold anyone for interrupting, but they would say something like, "That's interesting, but I think Jane was still talking," subtly pointing out a violation of the norm.

2) People started saying less at a time.

#1 is pretty obvious – with no enforcement of the social norm, a no-interruptions norm collapses pretty quickly. But #2 is actually really interesting. If talking at all is an implied claim that what you're saying is the most important thing that can be said, then polite people keep it short.

With 15-20 people in a seminar, this also meant that people rarely tried to force the conversation in a certain direction. When you're done talking, the conversation is out of your hands. This can be frustrating at first, but with time, you learn to trust not your fellow conversationalists individually, but the conversation itself, to go where it needs to. If you haven't said enough, then you trust that someone will ask you a question, and you'll say more.

When people are interrupting each other – when they're constantly tugging the conversation back and forth between their preferred directions – then the conversation itself is just a battle of wills. But when people just put in one thing at a time, and trust their fellows to only say things that relate to the thing that came right before – at least, until there's a very long pause – then you start to see genuine collaboration.

And when a lull in the conversation is treated as an opportunity to think about the last thing said, rather than an opportunity to jump in with the thing you were holding onto from 15 minutes ago because you couldn't just interrupt and say it – then you also open yourself up to being genuinely surprised, to seeing the conversation go somewhere that no one in the room would have predicted, to introduce ideas that no one brought with them when they sat down at the table.

By the time I graduated, I'd internalized this norm, and the rest of the world seemed rude to me for a few months. Not just because of the interrupting – but more because I'd say one thing, politely pause, and then people would assume I was done and start explaining why I was wrong – without asking any questions! Eventually, I realized that I'd been perfectly comfortable with these sorts of interactions before college. I just needed to code-switch! Some people are more comfortable with a culture of interrupting when you want to, and accepting interruptions. Others are more comfortable with a culture of waiting their turn, and courteously saying only one thing at a time, not trying to cram in a whole bunch of arguments for their thesis.

Now, I've praised the virtues of wait culture because I think it's undervalued, but there's plenty to say for interrupt culture as well. For one, it's more robust in "unwalled" circumstances. If there's no one around to enforce wait culture norms, then a few jerks can dominate the discussion, silencing everyone else. But someone who doesn't follow "interrupt" norms only silences themselves.

Second, it's faster and easier to calibrate how much someone else feels the need to talk, when they're willing to interrupt you. It takes willpower to stop talking when you're not sure you were perfectly clear, and to trust others to pick up the slack. It's much easier to keep going until they stop you.

So if you're only used to one style, see if you can try out the other somewhere. Or at least pay attention and see whether you're talking to someone who follows the other norm. And don't assume that you know which norm is the "right" one; try it the "wrong" way and maybe you'll learn something.

Cross-posted at my personal blog.

No Universally Compelling Arguments in Math or Science

Last week, I started a thread on the widespread sentiment that people don't understand the metaethics sequence. One of the things that surprised me most in the thread was this exchange:

Commenter: "I happen to (mostly) agree that there aren't universally compelling arguments, but I still wish there were. The metaethics sequence failed to talk me out of valuing this."

Me: "But you realize that Eliezer is arguing that there aren't universally compelling arguments in any domain, including mathematics or science? So if that doesn't threaten the objectivity of mathematics or science, why should that threaten the objectivity of morality?"

Commenter: "Waah? Of course there are universally compelling arguments in math and science."

Now, I realize this is just one commenter. But the most-upvoted comment in the thread also perceived "no universally compelling arguments" as a major source of confusion, suggesting that it was perceived as conflicting with morality not being arbitrary. And today, someone mentioned having "no universally compelling arguments" cited at them as a decisive refutation of moral realism.

After the exchange quoted above, I went back and read the original No Universally Compelling Arguments post, and realized that while it had been obvious to *me* when I read it that Eliezer meant it to apply to everything, math and science included, it was rather short on concrete examples, perhaps in violation of Eliezer's own advice. The concrete examples can be found in the sequences, though... just not in that particular post.

First, I recommend reading The Design Space of Minds-In-General if you haven't already. TLDR; the space of minds in general ginormous and includes some downright weird minds. The space of human minds is a teeny tiny dot in the larger space (in case this isn't clear, the diagram in that post isn't remotely drawn to scale). Now with that out of the way...

There are minds in the space of minds-in-general that do not recognize *modus ponens.*

Modus ponens is the rule of inference that says that if you have a statement of the form "If A then B", and also have "A", then you can derive "B". It's a fundamental part of logic. But there are possible mind that reject it. A brilliant illustration of this point can be found in Lewis Carroll's dialog "What the Tortoise Said to Achilles" (for those not in the know, Carroll was a mathematician; *Alice in Wonderland* is secretly full of math jokes).

Eliezer covers the dialog in his post Created Already In Motion, but here's the short version: In Carroll's dialog, the tortoise asks Achilles to imagine someone rejecting a particular instance of *modus ponens* (drawn from Euclid's *Elements,* though that isn't important). The Tortoise suggests that such a person might be persuaded by adding

an additional premise, and Achilles goes along with it—foolishly, because this quickly leads to an infinite regress when the Tortoise suggests that someone might reject the new argument in spite of accepting the premises (which leads to another round of trying to patch the argument, and then..)

"What the Tortoise Said to Achilles" is one of the reasons I tend to think of the so-called "problem of induction" as a pseudo-problem. The "problem of induction" is often defined as the problem of how to justify induction, but it seems to make just as much senses to ask how to justify deduction. But speaking of induction...

There are minds in the space of minds-in-general that reason counter-inductively.

To quote Eliezer:

> There are possible minds in mind design space who have anti-Occamian and anti-Laplacian priors; they believe that simpler theories are less likely to be correct, and that the more often something happens, the less likely it is to happen again.
>
> And when you ask these strange beings why they keep using priors that never seem to work in real life... they reply, "Because it's never worked for us before!"

If this bothers you, well, I refer you back to Lewis' Carroll's dialog. There are also minds in the mind design space that ignore the standard laws of logic, and are furthermore totally unbothered by (what we would regard as) the absurdities produced by doing so. Oh, but if you thought that was bad, consider this...

There are minds in the space of minds-in-general that use a maximum entropy prior, and never learn anything.

Here's Eliezer again discussing a problem where you have to predict whether a ball drawn out of an urn will be red or white, based on the color of the balls that have been previously drawn out of the urn:

> Suppose that your prior information about the urn is that a monkey tosses balls into the urn, selecting red balls with 1/4 probability and white balls with 3/4 probability, each ball selected independently. The urn contains 10 balls, and we sample without replacement. (E. T. Jaynes called this the "binomial monkey prior".) Now suppose that on the first three rounds, you see three red balls. What is the probability of seeing a red ball on the fourth round?
>
> First, we calculate the prior probability that the monkey tossed 0 red balls and 10 white balls into the urn; then the prior probability that the monkey tossed 1 red ball and 9 white balls into the urn; and so on. Then we take our evidence (three red balls, sampled without replacement) and calculate the likelihood of seeing that evidence, conditioned on each of the possible urn contents. Then we update and normalize the posterior probability of the possible remaining urn contents. Then we average over the probability of drawing a red ball from each possible urn, weighted by that urn's posterior probability. And the answer is... *(scribbles frantically for quite some time)*... 1/4!
>
> Of course it's 1/4. We specified that each ball was independently tossed into the urn, with a known 1/4 probability of being red. Imagine that the monkey is tossing the balls to you, one by one; if it tosses you a red ball on one round, that doesn't change the probability that it tosses you a red ball on the next round. When we

> withdraw one ball from the urn, it doesn't tell us anything about the other balls in the urn.
>
> If you start out with a maximum-entropy prior, then you never learn anything, ever, no matter how much evidence you observe. You do not even learn anything wrong - you always remain as ignorant as you began.

You may think, while minds such as I've been describing are possible in theory, they're unlikely to evolve anywhere in the universe, and probably they wouldn't survive long if programmed as an AI. And you'd probably be right about that. On the other hand, it's not hard to imagine minds that are generally able to get along well in the world, but irredeemably crazy on particular questions. Sometimes, it's tempting to suspect some humans of being this way, and even if that isn't *literally* true of any humans, it's not hard to imagine as just a more extreme form of existing human tendencies. See e.g. Robin Hanson on near vs. far mode, and imagine a mind that will literally never leave far mode on certain questions, regardless of the circumstances.

It used to disturb me to think that there might be, say, young earth creationists in the world who couldn't be persuaded to give up their young earth creationism by any evidence or arguments, no matter how long they lived. Yet I've realized that, while there may or may not be actual human young earth creationists like that (it's an empirical question), there are certainly possible minds in the space of mind designs like that. And when I think about that fact, I'm forced to shrug my shoulders and say, "oh well" and leave it at that.

That means I can understand why people would be bothered by a lack of universally compelling arguments for their moral views... but you shouldn't be any *more* bothered by that than by the lack of universally compelling arguments against young earth creationism. And if you don't think the lack of universally compelling arguments is a reason to think there's no objective truth about the age of the earth, you shouldn't think it's a reason to think there's no objective truth about morality.

(Note: this may end up being just the first in a series of posts on the metaethics sequence. People are welcome to discuss what I should cover in subsequent posts in the comments.)

Added: Based on initial comments, I wonder if some people who describe themselves as being bothered the lack of universally compelling arguments would more accurately describe themselves as being bothered by the orthogonality thesis.

Yes, Virginia, You Can Be 99.99% (Or More!) Certain That 53 Is Prime

TLDR; though you can't be 100% certain of anything, a lot of the people who go around talking about how you can't be 100% certain of anything would be surprised at how often you can be 99.99% certain. Indeed, we're often justified in assigning odds ratios well in excess of a million to one to certain claims. Realizing this is important for avoiding certain rookie Bayesian's mistakes, as well as for thinking about existential risk.

---

53 is prime. I'm *very* confident of this. 99.99% confident, at the very least. How can I be so confident? Because of the following argument:

If a number is composite, it must have a prime factor no greater than its square root. Because 53 is less than 64, sqrt(53) is less than 8. So, to find out if 53 is prime or not, we only need to check if it can be divided by primes less than 8 (i.e. 2, 3, 5, and 7). 53's last digit is odd, so it's not divisible by 2. 53's last digit is neither 0 nor 5, so it's not divisible by 5. The nearest multiples of 3 are 51 (=17x3) and 54, so 53 is not divisible by 3. The nearest multiples of 7 are 49 (=7^2) and 56, so 53 is not divisible by 7. Therefore, 53 is prime.

(My confidence in this argument is helped by the fact that I was good at math in high school. Your confidence in your math abilities may vary.)

I mention this because in his post Infinite Certainty, Eliezer writes:

> Suppose you say that you're 99.99% confident that 2 + 2 = 4. Then you have just asserted that you could make 10,000 *independent* statements, in which you repose equal confidence, and be wrong, on average, around once. Maybe for 2 + 2 = 4 this extraordinary degree of confidence would be possible: "2 + 2 = 4" extremely simple, and mathematical as well as empirical, and widely believed socially (not with passionate affirmation but just quietly taken for granted). So maybe you really could get up to 99.99% confidence on this one.
>
> I don't think you could get up to 99.99% confidence for assertions like "53 is a prime number". Yes, it seems likely, but by the time you tried to set up protocols that would let you assert 10,000 *independent* statements of this sort—that is, not just a set of statements about prime numbers, but a new protocol each time—you would fail more than once. Peter de Blanc has an amusing anecdote on this point, which he is welcome to retell in the comments.

I think this argument that you can't be 99.99% certain that 53 is prime is fallacious. Stuart Armstrong explains why in the comments:

> *If you say 99.9999% confidence, you're implying that you could make one million equally fraught statements, one after the other, and be wrong, on average, about once.*
>
> Excellent post overall, but that part seems weakest - we suffer from an unavailability problem, in that we can't just think up random statements with those properties. When I said I agreed 99.9999% with "P(P is never equal to 1)" it

> doesn't mean that I feel I could produce such a list - just that I have a very high belief that such a list could exist.

In other words, it's true that:

- If a well-calibrated person claims to be 99.99% certain of 10,000 independent statements, on average one of those statements should be false.

But it doesn't follow that:

- If a well-calibrated person claims to be 99.99% certain of one statement, they should be able to produce 9,999 other independent statements of equal certainty and be wrong on average once.

If it's not clear why this doesn't follow consider the anecdote Eliezer references in the quote above, which runs as follows: A gets B to agree that if 7 is not prime, B will give A $100. B then makes the same agreement for 11, 13, 17, 19, and 23. Then A asks about 27. B refuses. What about 29? Sure. 31? Yes. 33? No. 37? Yes. 39? No. 41? Yes. 43? Yes. 47? Yes. 49? No. 51? Yes. And suddenly B is $100 poorer.

Now, B claimed to be 100% sure about 7 being prime, which I don't agree with. But that's not what lost him his $100. What lost him his $100 is that, as the game went on, he got careless. If he'd taken the time to ask himself, "am I really as sure about 51 as I am about 7?" he'd probably have realized the answer was "no." He probably didn't check he primality of 51 as carefully as I checked the primality of 53 at the beginning of this post. (From the provided chat transcript, sleep deprivation may have also had something to do with it.)

If you tried to make 10,000 statements with 99.99% certainty, sooner or later you would get careless. Heck, before I started writing this post, I tried typing up a list of statements I was sure of, and it wasn't long before I'd typed 1 + 0 = 10 (I'd *meant* to type 1 + 9 = 10. Oops.) But the fact that, as the exercise went on, you'd start including statements that weren't really as certain as the first statement doesn't mean you couldn't be justified in being 99.99% certain of that first statement.

I almost feel like I should apologize for nitpicking this, because I agree with the main point of the "Infinite Certainty" post, that you should never assign a proposition probability 1. Assigning a proposition a probability of 1 implies that no evidence could ever convince you otherwise, and I agree that that's bad. But I think it's important to say that you're often justified in putting *a lot* of 9s after the decimal point in your probability assignments, for a few reasons.

One reason is arguments in the style of Eliezer's "10,000 independent statements" argument lead to inconsistencies. From another post of Eliezer's:

> I would be substantially more alarmed about a lottery device with a well-defined chance of 1 in 1,000,000 of destroying the world, than I am about the Large Hadron Collider being switched on.
>
> On the other hand, if you asked me whether I could make one million statements of authority equal to "The Large Hadron Collider will not destroy the world", and be wrong, on average, around once, then I would have to say no.
>
> What should I do about this inconsistency? I'm not sure, but I'm certainly not going to wave a magic wand to make it go away. That's like finding an

> inconsistency in a pair of maps you own, and quickly scribbling some alterations to make sure they're consistent.
>
> I would also, by the way, be substantially more worried about a lottery device with a 1 in 1,000,000,000 chance of destroying the world, than a device which destroyed the world if the Judeo-Christian God existed. But I would not suppose that I could make one billion statements, one after the other, fully independent and equally fraught as "There is no God", and be wrong on average around once.

Okay, so that's just Eliezer. But in a way, it's just a sophisticated version of a mistake a lot of novice students of probability make. Many people, when you tell them they can never be 100% certain of anything, respond switching to saying 99% or 99.9% whenever they previously would have said 100%.

In a sense they have the right idea—there are lots of situations where, while the appropriate probability is not 0, it's still negligible. But 1% or even 0.1% isn't negligible enough in many contexts. Generally, you should not be in the habit of doing things that have a 0.1% chance of killing you. Do so on a daily basis, and on average you will be dead in less than three years. Conversely, if you mistakenly assign a 0.1% chance that you will die each time you leave the house, you may never leave the house.

Furthermore, the ways this can trip people up aren't just hypothetical. Christian apologist William Lane Craig claims the skeptical slogan "extraordinary claims require extraordinary evidence" is contradicted by probability theory, because it actually wouldn't take all that much evidence to convince us that, for example, "the numbers chosen in last night's lottery were 4, 2, 9, 7, 8 and 3." The correct response to this argument is to say that the prior probability of a miracle occurring is orders of magnitude smaller than mere one in a million odds.

I suspect many novice students of probability will be uncomfortable with that response. They shouldn't be, though. After all, if you tried to convince the average Christian of Joseph Smith's story with the golden plates, they'd require *much* more evidence than they'd need to be convinced that last night's lottery numbers were 4, 2, 9, 7, 8 and 3. That suggests their prior for Mormonism is much less than one in a million.

This also matters a lot for thinking about futurism and existential risk. If someone is in the habit of using "99%" as shorthand for "basically 100%," they will have trouble grasping the thought "I am 99% certain this futuristic scenario will not happen, but the stakes are high enough that I need to take the 1% chance into account in my decision making." Actually, I suspect that problems in this vicinity explain much of the problems ordinary people (read: including average scientists) have thinking about existential risk.

I agree with what Eliezer has said about being ware of picking numbers out of thin air and trying to do math with them. (Or if you are going to pick numbers out of thin air, at least be ready to abandon your numbers at the drop of a hat.) Such advice goes double for dealing with very small probabilities, which humans seem to be *especially* bad at thinking about.

But it's worth trying to internalize a sense that there are several very different categories of improbable claims, along the lines of:

- Things that have a probability of something like 1%. These are things you really don't want to bet your life on if you can help it.
- Things that have a probability of something like one in a million. Includes many common ways to die that don't involve doing anything most people would regard as especially risky. For example, these stats suggest the odds of a 100 mile car trip killing you are somewhere on the order of one in a million.
- Things whose probability is truly negligible outside alternate universes where your evidence is radically different than what it actually is. For example, the risk of the Earth being overrun by demons.

Furthermore, it's worth trying to learn to think coherently about which claims belong in which category. That includes not being afraid to assign claims to the third category when necessary.

Added: I also recommend the links in this comment by komponisto.

I notice that I am confused about Identity and Resurrection

I've spent quite a bit of time trying to work out how to explain the roots of my confusion. I think, in the great LW tradition, I'll start with a story.

[Editor's note: The original story was in 16th century Mandarin, and used peculiar and esoteric terms for concepts that are just now being re-discovered. Where possible, I have translated these terms into their modern mathematical and philosophical equivalents. Such terms are denoted with curly braces, {like so}.]

Once upon a time there was a man by the name of Shen Chun-lieh, and he had a beautiful young daughter named Ah-Chen. She died.

Shen Chun-lieh was heartbroken, moreso he thought than any man who had lost a daughter, and so he struggled and scraped and misered until he had amassed a great fortune, and brought that fortune before me - for he had heard it told that I was could resurrect the dead.

I frowned when he told me his story, for many things are true after a fashion, but wisdom is in understanding the nature of that truth - and he did not bear the face of a wise man.

"Tell me about your daughter, Ah-Chen.", I commanded.

And so he told me.

I frowned, for my suspicions were confirmed.

"You wish for me to give you this back?", I asked.

He nodded and dried his tears. "More than anything in the world."

"Then come back tomorrow, and I will have for you a beautiful daughter who will do all the things you described."

His face showed a sudden flash of understanding. Perhaps, I thought, this one might see after all.

"But", he said, "will it be Ah-Chen?"

I smiled sagely. "What do you mean by that, Shen Chun-lieh?"

"I mean, you said that you would give me 'a' daughter. I wish for MY daughter."

I bowed to his small wisdom. "Indeed I did. If you wish for YOUR daughter, then you must be much, much more precise with me."

He frowned, and I saw in his face that he did not have the words.

"You are wise in the way of the Tao", he said, "surely you can find the words in my heart, so that even such as me could say them?"

I nodded. "I can. But it will take a great amount of time, and much courage from you. Shall we proceed?"

He nodded.

I am wise enough in the way of the Tao. The Tao whispers things that have been discovered and forgotten, and things that have yet to be discovered, and things that may never be discovered. And while Shen Chun-lieh was neither wise nor particularly courageous, his overwhelming desire to see his daughter again propelled him with an intensity seldom seen in my students. And so it was, many years later, that I judged him finally ready to discuss his daughter with me, in earnest.

"Shen", I said, "it is time to talk about your Ah-Chen."

His eyes brightened and he nodded eagerly. "Yes, Teacher."

"Do you understand why I said on that first day, that you must be much, much more precise with me?"

"Yes, Teacher. I had come to you believing that the soul was a thing that could be conjured back to the living, rather than a {computational process}."

"Even now, you are not quite correct. The soul is not a {computational process}, but a {specification of a search space} which describes any number of similar {computational processes}. For example, Shen Chun-lieh, would you still be Shen Chun-lieh if I were to cut off your left arm?"

"Of course, Teacher. My left arm does not define who I am."

"Indeed. And are you still the same Shen Chun-lieh who came to me all those years ago, begging me to give him back his daughter Ah-Chen?"

"I am, Teacher, although I understand much more now than I did then."

"That you do. But tell me - would you be the same Shen Chun-lieh if you had not come to me? If you had continued to save and to save your money, and craft more desperate and eager schemes for amassing more money, until finally you forgot the purpose of your misering altogether, and abandoned your Ah-Chen to the pursuit of gold and jade for its own sake?"

"Teacher, my love for Ah-Chen is all-consuming; such a fate could never befall me."

"Do not be so sure, my student. Remember the tale of the butterfly's wings, and the storm that sank an armada. Ever-shifting is the Tao, and so ever-shifting is our place in it."

Shen Chun-lieh understood, and in a brief moment he glimpsed his life as it could have been, as an old Miser Shen hoarding gold and jade in a great walled city. He shuddered and prostrated himself.

"Teacher, you are correct. And even such a wretch as Miser Shen, that wretch would still be me. But I thank the Buddha and the Eight Immortal Sages that I was spared that fate."

I smiled benevolently and helped him to his feet. "Then suppose that you had died and not your daughter, and one day a young woman named Ah-Chen had burst into my door, flinging gold and jade upon my table, and described the caring and wonderful father that she wished returned to her? What could she say about Shen Chun-lieh that would allow me to find his soul amongst the infinite chaos of the Nine Hells?"

"I..." He looked utterly lost.

"Tell me, Shen Chun-lieh, what is the meaning of the parable of the {Ship of Theseus}?"

"That personal identity cannot be contained within the body, for the flow of the Tao slowly strips away and the flow of the Tao slowly restores, such that no single piece of my body is the same from one year to the next; and within the Tao, even the distinction of 'sameness' is meaningless."

"And what is the relevance of the parable of the {Shroedinger's Cat} to this discussion?"

"Umm... that... let me think. I suppose, that personal identity cannot be contained within the history of choices that have been made, because for every choice that has been made, if it was truly a 'choice' at all, it was also made the other way in some other tributary of the Great Tao."

"And the parable of the tiny {Paramecium}?"

"That neither is the copy; there are *two* originals."

"So, Shen. Can you yet articulate the dilemma that you present to me?"

"No, Teacher. I fear that yet again, you must point it out to your humble student."

"You ask for Ah-Chen, my student. But *which one*? Of all the Ah-Chens that could be brought before you, which would satisfy you? Because there is no hard line, between {configurations} that you would recognize as your daughter and {configurations} which you would not. So why did my original offer, to construct you a daughter that would do all the things you described Ah-Chen as doing, not appeal to you?"

Shen looked horrified. "Because she would not BE Ah-Chen! Even if you made her respond perfectly, it would not be HER! I do not simply miss my six-year-old girl; I miss what she could have become! I regret that she never got to see the world, never got to grow up, never got to..."

"In what sense did she never do these things? She died, yes; but even a dead Ah-Chen is still an Ah-Chen. She has since experienced being worms beneath the earth, and flowers, and then bees and birds and foxes and deer and even peasants and noblemen. All these are Ah-Chen, so why is it so important that she appear before you as YOU remember her?"

"Because I miss her, and because she has no conscious awareness of those things."

"Ah, but then which conscious awareness do you wish her to have? There is no copy; all possible tributaries of the Great Tao contain an original. And each of those originals experience in their own way. You wish me to pluck out a {configuration} and present it

to you, and declare "This one! This one is Ah-Chen!". But which one? Or do you leave that choice to me?"

"No, Teacher. I know better than to leave that choice to you. But... you have shown me many great wonders, in alchemy and in other works of the Tao. If her brain had been preserved, perhaps frozen as you showed me the frozen koi, I could present that to you and you could reconstruct her {configuration} from that?"

I smiled sadly. "To certain degrees of precision, yes, I could. But the question still remains - you have only narrowed down the possible {configurations}. And what makes you say that the boundary of {configurations} that are achievable from a frozen brain are correct? If I smash that brain with a hammer, melt it, and paint a portrait of Ah-Chen with it, is that not a {configuration} that is achievable from that brain?"

Shen looked disgusted. "You... how can you be so wise and yet not understand such simple things? We are talking about people! Not paintings!"

I continued to smile sadly. "Because these things are not so simple. 'People' are not things, as you said before. 'People' are {sets of configurations}; they are {specifications of search spaces}. And those boundaries are so indistinct that anything that claims to capture them is in error."

Now it was Shen's turn to look animated. "Just because the boundary cannot be drawn perfectly, does not make the boundary meaningless!"

I nodded. "You have indeed learned much. But you still have not described the purpose of your boundary-drawing. Do you wish for Ah-Chen's resurrection for yourself, so that you may feel less lonely and grieved, or do you wish it for Ah-Chen's sake, so that she may see the world anew? For these two purposes will give us very different boundaries for what is an acceptable Ah-Chen."

Shen grimaced, as war raged within his heart. "You are so wise in the Tao; stop these games and do what I mean!"

And so it was that Miser Shen came to live in the walled city of Ch'in, and hoarded gold and jade, and lost all memory and desire for his daughter Ah-Chen, until it was that the Tao swept him up into another tale.

So, there we are. My confusion is in two parts:

1. When I imagine resurrecting loved ones, what makes me believe that even a perfectly preserved brain state is any more 'resurrection' than an overly sophisticated wind-up toy that happens to behave in ways that fulfill my desire for that loved one's company? In a certain sense, avoiding true 'resurrection' should be PREFERABLE - since it is possible that a "wind-up toy" could be constructed that provides a superstimulus version of that loved one's company, while an actual 'resurrection' will only be as good as the real thing.

2. When I imagine being resurrected "myself", how different from this 'me' can it be and still count? How is this fundamentally different from "I will for the future to contain a being like myself", which is really just "I will for the future to contain a being like I

imagine myself to be" - in which case, we're back to the superstimulus option (which is perhaps a little weird in this case, since I'm not there to receive the stimulus).

I'd really like to discuss this.