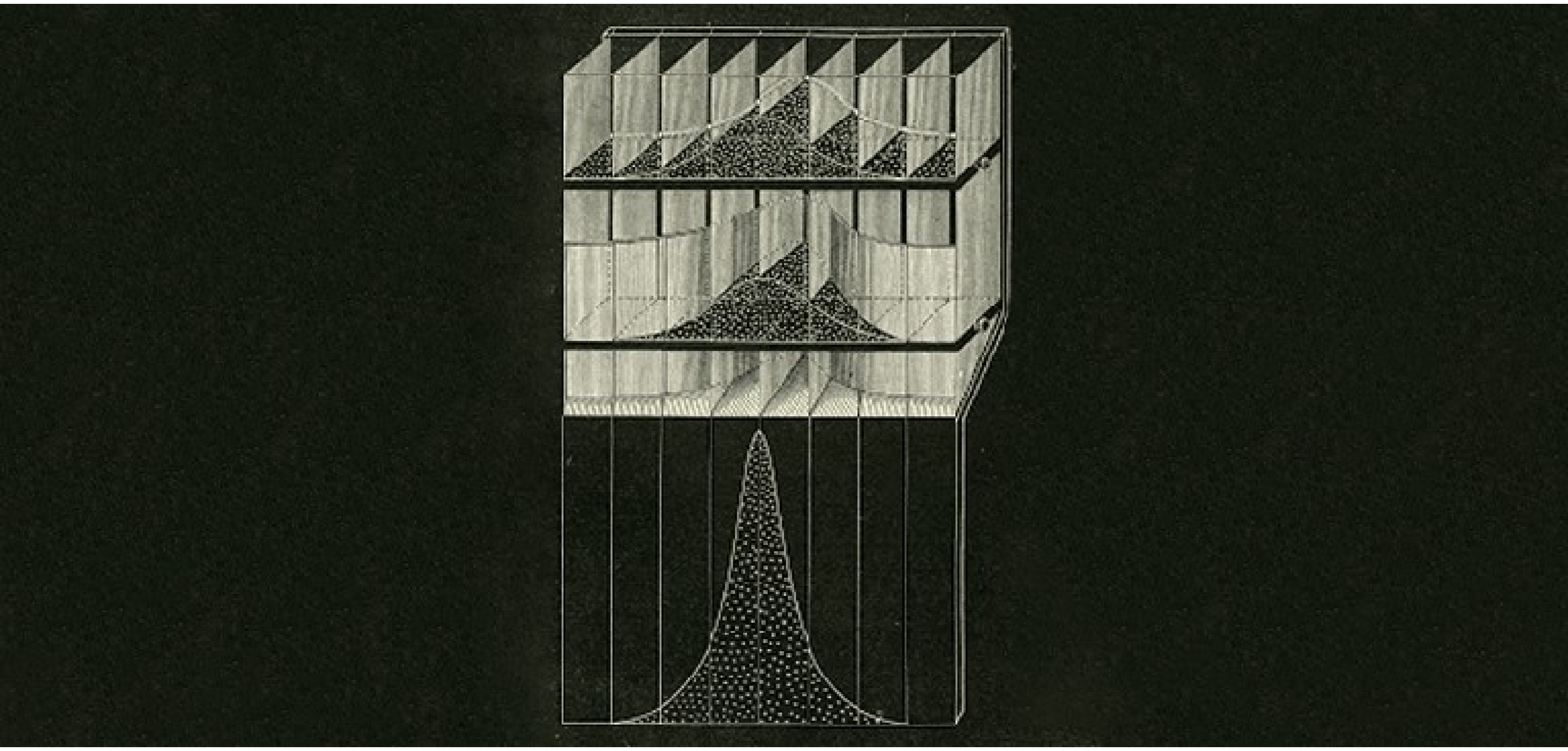

Probability and Predictions

The Pyramid And The Garden

I.

A recent breakthrough in pseudoscience: the location of the Great Pyramid of Giza encodes the speed of light to seven decimal places.

This is actually true. The speed of light in a vacuum is 299,792,458 meters per second. The coordinates of of the Great Pyramid are 29.9792458° N, 31.1342880° E (you can confirm with Google Maps that this gets you right on top of the Pyramid). The speed of light and the latitude number there have all the same digits. That's a pretty impressive coincidence.

You might think this is idiotic because the meter was invented by 1600s French people. If ancient aliens or Atlanteans built the pyramids, why would they encode their secret wisdom using a unit of measurement from 1600s France? But there's a way around this objection: the 1600s French people defined their meter as 1/10,000,000th the distance between the Equator and the North Pole. If the aliens also thought that was an interesting way to measure length, then they could have encoded their secret wisdom in it. So you wouldn't need aliens who could predict the thoughts of 1600s Frenchmen. Just aliens who *thought exactly like* 1600s Frenchmen.

(actually, a different group of 1600s Frenchmen proposed a different version of the meter, defined as the length of a pendulum with a half-period of one second. This turned out to be 99.7% of the 1/10,000,000th-the-way-to-the-North-Pole definition, so either one works unless you want super-exactness. I think a much more interesting conspiracy theory would be that aliens designed the Earth to encode secret wisdom about the periods of pendulums.)

But realistically, aliens who think suspiciously like French people probably weren't involved. So how do we explain the coincidence?

II.

The following is indebted to user mrfintoil's great explanation on metabunk.org.

First, it's *not* a coincidence to seven decimal places. Yes, that particular nine-digit sequence lands you atop the Great Pyramid. But that gives you way more precision than you need – cutting off the last three digits actually gets you closer rather than further from the center of the Pyramid. The only numbers that are doing any work are the 29.9792° N. So you really only get four decimal places worth of coincidence.

On the other hand, matching six digits is still pretty good. That's literally a one-in-a-million chance.

So here the explanation has to go to how hard the pseudoscientists worked to find a coincidence of this magnitude; in other words, how many degrees of freedom they had.

Here's an obvious example; as far as I can tell, the *longitude* of the Great Pyramid doesn't encode anything interesting at all. So it's not the equivalent of winning a one-in-a-million lottery with a single ticket. It's the equivalent of winning a one-in-a-million lottery with *two* tickets.

A second issue: if the latitude of the Great Pyramid had been 10.7925 N, that would be the speed of light in kilometers per hour, which would be an equally impressive match.

So just taking these two degrees of freedom, we have four lottery tickets:

1. The one where the latitude is the speed of light in meters/second
2. The one where the longitude is the speed of light in meters/second
3. The one where the latitude is the speed of light in kilometers/hour
4. The one where the longitude is the speed of light in kilometers/hour

In other words, the number of lottery tickets increases exponentially as we get more degrees of freedom.

Let me list out all the degrees of freedom I can think of and see where we end up. I am going to try my best to be as fair as possible to the ancient aliens. For example, I was considering saying that since there are three pyramids at Giza, we have to multiply by three, but to be honest the Great Pyramid is clearly greater than the other two, and it would be less elegant if Menkaure’s pyramid encoded some amazing cosmic constant, so I won’t raise that objection. I am going to try to be *really fricking fair*.

1. Latitude vs. longitude (2 options)

2. Speed of light in meters/second vs. kilometers/hour vs. cubits/second vs. cubits/hour. I’m avoiding using feet/miles, because that’s even more arbitrary than meters. But I think it would actually be even *more* convincing if the calculation actually used the real Egyptian unit, which I understand is the cubit. So let’s go with (4 options)

3. Great Pyramid vs. Sphinx. Like I said before, the other two pyramids at Giza are noticeably less impressive than the Great Pyramid. But the Sphinx is pretty impressive, and the ancient aliens folks talk about it just as much as the Pyramid, so I think that would be an equally good hit if it had been true. (2 options)

4. Use of a 90 degree latitude system vs. use of a 100 degree latitude system. I’m a little split on this one, because it wouldn’t look anywhere near as impressive if the pseudoscience sites had to explain that they found a really cool coincidence but it only worked if you converted normal latitude into a different hypothetical latitude system that had 100 degrees. But since we know the aliens/Atlanteans use base 10 anyway (they’re encoding their wisdom in the base 10 representation of the speed of light) it makes more sense for them to use a base 10 latitude system instead of replicating our own bizarre custom of using base 10 for everything else but having latitude go from 0 to 90. On the other hand, if these were Earth-based Atlanteans, they might have gotten the custom of dividing the circle into 360 parts for the same reason we did – there are about 360 days in a year. And if they were aliens, maybe we got our bizarre latitude convention from *them* – the idea of 360 degree circles is really old and lost in the mists of time. Overall I can see this one going either way, so I’m going to give it as (2 options)

5. Decimal point placement. The latitude 29.9792 N matches the speed of light exactly, but so would the latitudes 2.99792, 2.99792 S, and 29.9792 S. I checked these other sites at the same longitude as the Pyramid to see if there were any mysterious features. But they seem to be, respectively, a perfectly ordinary field in Uganda, a perfectly ordinary field in Tanzania, and a perfectly ordinary patch of ocean.

But a world where the pyramid was in Uganda and the ordinary field was in Egypt would be just as much of a hit as our current world. Therefore (4 options)

From these really simple things alone, we learn we've got 2 x 4 x 2 x 2 x 4 = 128 lottery tickets, reducing our 1/1 million chance of winning to something more like 1/10,000. Progress!

There are a few other degrees of freedom that I think are a little harder to judge, but still important:

6. What aspect of the Pyramid we're looking at. That is, it would have been equally interesting (maybe moreso!) if its height or width matched the speed of light exactly. So that's another (3 options). I guess if the ancient aliens were *really* good at what they were doing, they could have given the pyramid 299,792,458 *sides*, but I won't hold that against them. This should really make the multiplication more complicated because I can no longer use all the different ways of representing latitude vs. longitude, but I'll stick with the simple method for now.

7. Which *site* we're looking at. This one is hard, because I don't know if anywhere else has the ancient alien-related credibility of the Great Pyramid. The only equally mysterious site I can think of is Stonehenge, and *maybe* the Nazca Lines. I don't feel comfortable saying it would be *equally* impressive if Tiwanaku or Yonaguni had the right coordinates. I'll just say (2 options) for Pyramids and Stonehenge.

8. Which *constant* we're looking at. Sure, the Pyramid encoding the speed of light is pretty cool, but what about the Planck length? Avagadro's number? I'm split on whether I want to include mathematical constants like pi or e in here. I think if it encoded pi to some number of decimals places then I would just think that the Egyptians were more advanced at math than I thought but it wouldn't necessarily be earth-shattering. The Egyptians knowing e would be pretty shocking but still maybe not worth believing in ancient aliens over. There really aren't that many physical constants as cool as the speed of light, so I might just arbitrarily call this one (4 options).

So now we have a total of 128 x 3 x 2 x 4 = 3072 lottery tickets, for a 1/300 chance of winning the one-in-a-million lottery.

I would like to say "Ha ha, I sure proved those dumb conspiracy nuts wrong", except that a 1/300 chance is still a pretty impressive coincidence – what scientists call $p < 0.01$. And now I've used up all my excuses. I *think* what's going on here is that I'm still accepting the terms of the game – comparing only the exact categories used in the original calculation. Suppose that the latitude of the Great Pyramid was exactly 30.0000? That too would be impressive – it would prove that the pyramid builders knew the exact size and shape of the Earth and were able to build their Pyramid one third of the way between Equator and Pole. Suppose that the Great Pyramid was latitude 19.69724. That's the date humankind first landed on the moon in yyyy/mm/dd format – clearly the Pyramid was built by a time-traveling Nostradamus! Suppose that the Pyramid was built of stones of four different colors, with blue stones always paired opposite red stones, and yellow stones always paired opposite green stones. Then the ancient Egyptians were trying to tell us about the structure of DNA. What if the Pyramid, viewed from above, looked like a human brain?

Is it fair to take all of that into account? If so, does the remaining coincidence go away? I wish I were able to give these questions a more confident affirmative answer.

III.

I still believe that pseudoscience is helpful for understanding regular science. The loopholes that let people discover proofs of ESP or homeopathy are the same ones that let them discover proofs of power posing and ego depletion.

In the same way, numerology is helpful for understanding statistics. You can see the same factors at work, free from any lingering worry that maybe the theory you’re investigating is true after all.

Andrew Gelman writes about the garden of forking paths. The idea is: the scientific community accepts a discovery as meaningful if $p < 0.05$ - that is, if equally extreme data would only occur by coincidence 5% of the time or less. In other words, you need to win a lottery with a one-in-twenty chance if you want to get credit for discovering something absent any real effect to be discovered. But if a scientist forms their hypothesis after seeing their data, they might massage the precise wording of their hypothesis to better fit their data. If there are many different ways to frame the hypothesis, then they have many lottery tickets to choose from and a win is no longer so surprising. Gelman discusses a study claiming to find that women wear red or pink shirts during the most fertile part of their menstrual cycle, which sometimes involves red or pink coloration changes in primates. The study does detect the effect, $p < 0.05$. But there were a couple of different ways the researchers could have framed the problem. They could have looked at only red shirts. They could have looked at only pink shirts. They chose days 7-14 as most fertile. But they could also have chosen days 6-15 without really being *wrong*. They could have looked only at the unmarried women most likely to be trying to attract mates. A recent paper listed 34 different degrees of freedom that can be used in this kind of thing. Add up enough of them, and you have more than twenty tickets to the one-chance-in-twenty lottery and success is all but certain.

I used to call this the Elderly Hispanic Woman Effect, after drug studies where the drug has no effect in general, no effect on a subgroup of just men, no effect on a subgroup of just women, no effect on a subgroup of just blacks, no effect on a subgroup of just whites…but when you get to a subgroup of elderly Hispanic women, $p < 0.05$, apparently because it's synchronized with their unique biological needs. This is pretty obvious. The lesson of the Pyramid-lightspeed link is that sometimes it isn't. It just looks like some sudden and shocking coincidence. The other lesson of the Pyramid is that *I cannot consistently figure this kind of thing out*. I threw everything I had against the correlation, and I still ended up with $p = 0.003$. I don’t think this is because the Pyramid really *was* designed by aliens with a suspicious link to 1600s France. I think it’s because I’m not creative enough to fully dissect coincidences even when I’m looking for them.

This is always happening to me in real studies too. Something seems very suspicious. But their effect size is very high and their p-value is very significant. I can’t always figure out exactly what’s going on. But I should be reluctant to dismiss the possibility that I’m missing something and that there’s some reasonable explanation.

On Overconfidence

[Epistemic status: This is basic stuff to anyone who has read the Sequences, but since many readers here haven't I hope it is not too annoying to regurgitate it. Also, ironically, I'm not actually that sure of my thesis, which I guess means I'm extra-sure of my thesis]

I.

A couple of days ago, the Global Priorities Project came out with a calculator that allowed you to fill in your own numbers to estimate how concerned you should be with AI risk. One question asked how likely you thought it was that there would be dangerous superintelligences within a century, offering a drop down menu with probabilities ranging from 90% to 0.01%. And so people objected: there should be options to put in only one a million chance of AI risk! One in a billion! One in a…

For example, a commenter writes that: "the best (worst) part: the probability of AI risk is selected from a drop down list where the lowest probability available is 0.01%!! Are you kidding me??" and then goes on to say his estimate of the probability of human-level (not superintelligent!) AI this century is "very very low, maybe 1 in a million or less". Several people on Facebook and Tumblr say the same thing – 1/10,000 chance just doesn't represent how sure they are that there's no risk from AI, they want one in a million or more.

Last week, I mentioned that Dylan Matthews' suggestion that maybe there was only 10^-67 chance you could affect AI risk was stupendously overconfident. I mentioned that was thousands of lower than than the chance, *per second*, of getting simultaneously hit by a tornado, meteor, and al-Qaeda bomb, while *also* winning the lottery twice in a row. Unless you're comfortable with that level of improbability, you should stop using numbers like 10^-67.

But maybe it sounds like "one in a million" is much safer. That's only 10^-6, after all, way below the tornado-meteor-terrorist-double-lottery range…

So let's talk about overconfidence.

Nearly everyone is very very very overconfident. We know this from experiments where people answer true/false trivia questions, then are asked to state how confident they are in their answer. If people's confidence was well-calibrated, someone who said they were 99% confident (ie only 1% chance they're wrong) would get the question wrong only 1% of the time. In fact, people who say they are 99% confident get the question wrong about 20% of the time.

It gets worse. People who say there's only a 1 in 100,000 chance they're wrong? Wrong 15% of the time. One in a million? Wrong 5% of the time. They're not just overconfident, they are *fifty thousand times* as confident as they should be.

This is not just a methodological issue. Test confidence in some other clever way, and you get the same picture. For example, one experiment asked people how many numbers there were in the Boston phone book. They were instructed to set a range, such that the true number would be in their range 98% of the time (ie they would only be wrong 2% of the time). In fact, they were wrong 40% of the time. Twenty times too confident! What do you want to bet that if they'd been asked for a range so wide there was only a one in a million chance they'd be wrong, at least five percent of them would have bungled it?

Yet some people think they can predict the future course of AI with one in a million accuracy!

Imagine if every time you said you were sure of something to the level of 999,999/1 million, and you were right, the Probability Gods gave you a dollar. Every time you said this and you were wrong, you lost $1 million (if you don't have the cash on hand, the Probability Gods offer a generous payment plan at low interest). You might feel like getting some free cash for the parking meter by uttering statements like "The sun will rise in the east tomorrow" or "I won't get hit by a meteorite" without much risk. But would you feel comfortable predicting the course of AI over the next century? What if you noticed that most other people only managed to win $20 before they slipped up? Remember, if you say even one false statement under such a deal, all of your true statements you've said over years and years of perfect accuracy won't be worth the hole you've dug yourself.

Or – let me give you another intuition pump about how hard this is. Bayesian and frequentist statistics are pretty much the same thing [citation needed] – when I say "50% chance this coin will land heads", that's the same as saying "I expect it to land heads about one out of every two times." By the same token, "There's only a one in a million chance that I'm wrong about this" is the same as "I expect to be wrong on only one of a million statements like this that I make."

What do a million statements look like? Suppose I can fit twenty-five statements onto the page of an average-sized book. I start writing my predictions about scientific and technological progress in the next century. "I predict there will not be superintelligent AI." "I predict there will be no simple geoengineering fix for global warming." "I predict no one will prove P = NP." *War and Peace*, one of the longest books ever written, is about 1500 pages. After you write enough of these statements to fill a *War and Peace* sized book, you've made 37,500. You would need to write about 27 *War and Peace* sized books – enough to fill up a good-sized bookshelf – to have a million statements.

So, if you want to be confident to the level of one-in-a-million that there won't be superintelligent AI next century, you need to believe that you can fill up 27 *War and Peace* sized books with similar predictions about the next hundred years of technological progress – and be wrong – at most – once!

This is especially difficult because claims that a certain form of technological progress will not occur have a very poor track record of success, even when uttered by the most knowledgeable domain experts. Consider how Nobel-Prize winning atomic scientist Ernest Rutherford dismissed the possibility of nuclear power as "the merest moonshine" *less than a day* before Szilard figured out how to produce such power. In 1901, Wilbur Wright told his brother Orville that "man would not fly for fifty years" – two years later, they flew, leading Wilbur to say that "ever since, I have distrusted myself and avoided all predictions". Astronomer Joseph de Lalande told the French Academy that "it is impossible" to build a hot air balloon and "only a fool would expect such a thing to be realized"; the Montgolfier brothers flew less than a year later. This pattern has been so consistent throughout history that sci-fi titan Arthur C. Clarke (whose own predictions were often eerily accurate) made a heuristic out of it under the name Clarke's First Law: "When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong."

Also – one good heuristic is to look at what experts in a field think. According to Muller and Bostrom (2014), a sample of the top 100 most-cited authors in AI ascribed a > 70% probability to AI within a century, a 50% chance of superintelligence conditional on human-level, and a 10% chance of existential catastrophe conditional on human level AI. Multiply it out, and you get a couple percent chance of superintelligence-related existential catastrophe in the next century.

Note that my commenter wasn't disagreeing with the 4% chance. They were disagreeing with the possibility that *there would be human-level AI at all*, that is, the 70% chance! That means that he was saying, essentially, that he was confident he could write a million sentences – that is, twenty-seven *War and Peace*'s worth – all of which were trying to predict trends in a notoriously difficult field, all of which contradicted a well-known heuristic about what kind of predictions you should never try to make, all of which contradicted the consensus opinion of the relevant experts – and only have one of the million be wrong!

But if you feel superior to that because you don't believe there's only a one-in-a-million chance of human-level AI, you just believe there's a one-in-a-million chance of existential catastrophe, you are missing the point. Okay, you're not 300,000 times as confident as the experts, you're only 40,000 times as confident. Good job, here's a sticker.

Seriously, when people talk about being able to defy the experts a million times in a notoriously tricky area they don't know much about and only be wrong once – I don't know what to think. Some people criticize Eliezer Yudkowsky for being overconfident in his favored interpretation of quantum mechanics, but he doesn't even attach a number to that. For all I know, maybe he's only 99% sure he's right, or only 99.9%, or something. If you are absolutely outraged that he is claiming one-in-a-thousand certainty on something that doesn't much matter, shouldn't you be literally a thousand times more outraged when every day people are claiming one-in-a-million level certainty on something that matters very much? It is *almost impossible* for me to comprehend the mindsets of people who make a Federal Case out of the former, but are totally on board with the latter.

Everyone is overconfident. When people say one-in-a-million, they are wrong five percent of the time. And yet, people keep saying "There is only a one in a million chance I am wrong" on issues of making really complicated predictions about the future, where many top experts disagree with them, and where the road in front of them is littered with the bones of the people who made similar predictions before. HOW CAN YOU DO THAT?!

II.

I am of course eliding over an important issue. The experiments where people offering one-in-a-million chances were wrong 5% of the time were on true-false questions – those with only two possible answers. There are other situations where people can often say "one in a million" and be right. For example, I confidently predict that if you enter the lottery tomorrow, there's less than a one in a million chance you will win.

On the other hand, I feel like I can justify that. You want me to write twenty-seven *War and Peace* volumes about it? Okay, here goes. "Aaron Aaronson of Alabama will not win the lottery. Absalom Abramowtiz of Alaska will not win the lottery. Achitophel Acemoglu of Arkansas will not win the lottery." And so on through the names of a million lottery ticket holders.

I think this is what statisticians mean when they talk about “having a model”. Within the model where there are a hundred million ticket holders, and we know exactly one will be chosen, our predictions are on very firm ground, and our intuition pumps reflect that.

Another way to think of this is by analogy to dart throws. Suppose you have a target that is half red and half blue; you are aiming for red. You would have to be very very confident in your dart skills to say there is only a one in a million chance you will miss it. But if there is a target that is 999,999 millionths red, and 1 millionth blue, then you do not have to be at *all* good at darts to say confidently that there is only a one in a million chance you will miss the red area.

Suppose a Christian says “Jesus might be God. And he might not be God. 50-50 chance. So you would have to be incredibly overconfident to say you’re sure he isn’t.” The atheist might respond “The target is full of all of these zillions of hypotheses – Jesus is God, Allah is God, Ahura Mazda is God, Vishnu is God, a random guy we’ve never heard of is God. You are taking a tiny tiny submillimeter-sized fraction of a huge blue target, painting it red, and saying that because there are two regions of the target, a blue region and a red region, you have equal chance of hitting either.” Eliezer Yudkowsky calls this “privileging the hypothesis”.

There’s a tougher case. Suppose the Christian says “Okay, I’m not sure about Jesus. But either there is a Hell, or there isn’t. Fifty fifty. Right?”

I think the argument against this is that there are way more ways for there not to be Hell than there are for there to be Hell. If you take a bunch of atoms and shake them up, they usually end up as not-Hell, in much the same way as the creationists’ fabled tornado-going-through-a-junkyard usually ends up as not-a-Boeing-747. For there to be Hell you have to have some kind of mechanism for judging good vs. evil – which is a small part of the space of all mechanisms, let alone the space of all things – some mechanism for diverting the souls of the evil to a specific place, which same, some mechanism for punishing them – again same – et cetera. Most universes won’t have Hell unless you go through a *lot* of work to put one there. Therefore, Hell existing is only a very tiny part of the target. Making this argument correctly would require an in-depth explanation of formalizations of Occam’s Razor, which is outside the scope of this essay but which you can find on the LW Sequences.

But this kind of argumentation is really hard. Suppose I predict “Only one in 150 million chance Hillary Clinton will be elected President next year. After all, there are about 150 million Americans eligible for the Presidency. It could be any one of them. Therefore, Hillary covers only a tiny part of the target.” Obviously this is wrong, but it’s harder to explain how. I would say that your dart-aim is guided by an argument based on a concrete numerical model – something like “She is ahead in the polls by X right now, and candidates who are ahead in the polls by X usually win about 50% of the time, therefore, her real probability is more like 50%.”

Or suppose I predict “Only one in a million chance that Pythagoras’ Theorem will be proven wrong next year.” Can I get away with that? I can’t *quite* appeal to “it’s been proven”, because there might have been a mistake in (all the) proofs. But I could say: suppose there are five thousand great mathematical theorems that have undergone something like the level of scrutiny as Pythagoras’, and they’ve been known on average for two hundred years each. None of them have ever been disproven. That’s a numerical argument that the rate of theorem-disproving is less than one per million years, and I think it holds.

Another way to do this might be “there are three hundred proofs of Pythagoras’ theorem, so even accepting an absurdly high 10%-per-proof chance of being wrong, the chance is now only 10^-300.” Or “If there’s a 10% chance each mathematician reading a proof missing something, and one million mathematicians have read the proof of Pythagoras’ Theorem, then the probability that they all missed it is more like 10^-1,000,000.”

But this can get tricky. Suppose I argued “There’s a good chance Pythagoras’ Theorem will be disproven, because of all Pythagoras’ beliefs – reincarnation, eating beans being super-evil, ability to magically inscribe things on the moon – most have since been disproven. Therefore, the chance of a randomly selected Pythagoras-innovation being wrong is > 50%.”

Or: “In 50 past presidential elections, none have been won by women. But Hillary Clinton is a woman. Therefore, the chance of her winning this election is less than 1/50.”

All of this stuff about adjusting for size of the target or for having good mathematical models is really hard and easy to do wrong. And then you have to add another question: are you sure, to a level of one-in-a-million, that you didn’t mess up your choice of model at all?

Let’s bring this back to AI. Suppose that, given the complexity of the problem, you predict with utter certainty that we will not be able to invent an AI this century. But if the modal genome trick pushed by people like Greg Cochran works out, within a few decades we might be able to genetically engineer humans far smarter than any who have ever lived. Given tens of thousands of such supergeniuses, might we be able to solve an otherwise impossible problem? I don’t know. But if there’s a 1% chance that we can perform such engineering, and a 1% chance that such supergeniuses can invent artificial intelligence within a century, then the probability of AI within the next century isn’t one in a million, it’s one in ten thousand.

Or: consider the theory that all the hard work of brain design has been done by the time you have a rat brain, and after that it’s mostly just a matter of scaling up. You can find my argument for the position in this post – search for “the hard part is evolving so much as a tiny rat brain”. Suppose there’s a 10% chance this theory is true, and a 10% chance that researchers can at least make rat-level AI this century. Then the chance of human-level AI is not one in a million, but one in a hundred.

Maybe you disagree with both of these claims. The question is: *did you even think about them before you gave your one in a million estimate*? How many other things are there that you never thought about? Now your estimate has, somewhat bizarrely, committed you to saying there’s a less than one in a million chance we will significantly enhance human intelligence over the next century, *and* a less than one in a million chance that the basic-scale-up model of intelligence is true. You may never have thought directly about these problems, but by saying “one in a million chance of AI in the next hundred years”, you are not only committing yourself to a position on them, but committing yourself to a position with one-in-a-million level certainty even though several domain experts who have studied these fields for their entire lives disagree with you!

A claim like “one in a million chance of X” not only implies that your model is strong enough to spit out those kinds of numbers, but that there’s only a one in a million chance you're using the wrong model, or missing something, or screwing up the calculations.

A few years ago, a group of investment bankers came up with a model for predicting the market, and used it to design a trading strategy which they said would meet certain parameters. In fact, they said that there was only a one in 10^135 chance it would fail to meet those parameters during a given year. A human just uttered the probability "1 in 10^135", so you can probably guess what happened. The very next year was the 2007 financial crisis, the model wasn't prepared to deal with the extraordinary fallout, the strategy didn't meet its parameters, and the investment bank got clobbered.

This is why I don't like it when people say we shouldn't talk about AI risk because it involves "Knightian uncertainty". In the real world, Knightian uncertainty collapses back down to plain old regular uncertainty. When you are an investment bank, the money you lose because of normal uncertainty and the money you lose because of Knightian uncertainty are denominated in the same dollars. Knightian uncertainty becomes just another reason not to be overconfident.

III.

I came back to AI risk there, but this isn't just about AI risk.

You might have read Scott Aaronson's recent post about Aumann Agreement Theorem, which says that rational agents should be able to agree with one another. This is a nice utopian idea in principle, but in practice, well, nobody seems to be very good at carrying it out.

I'd like to propose a more modest version of Aumann's agreement theorem, call it Aumann's Less-Than-Total-Disagreement Theorem, which says that two rational agents shouldn't both end up with 99.9…% confidence on opposite sides of the same problem.

The "proof" is pretty similar to the original. Suppose you are 99.9% confident about something, and learn your equally educated, intelligent, and clear-thinking friend is 99.9% confident of the opposite. Arguing with each other and comparing your evidence fails to make either of you budge, and neither of you can marshal the weight of a bunch of experts saying you're right and the other guy is wrong. Shouldn't the fact that your friend, using a cognitive engine about as powerful as your own, got so heavily different a conclusion make you worry that you're missing something?

But practically everyone is walking around holding 99.9…% probabilities on the opposite sides of important issues! I checked the Less Wrong Survey, which is as good a source as any for people's confidence levels on various tough questions. Of the 1400 respondents, about 80 were at least 99.9% certain that there were intelligent aliens elsewhere in our galaxy; about 170 others were at least 99.9% certain that they weren't. At least 80 people just said they were certain to one part in a thousand and then got the answer wrong! And some of the responses were things like "this box cannot fit as many zeroes as it would take to say how certain I am". Aside from stock traders who are about to go bankrupt, *who says that sort of thing??!*

And speaking of aliens, imagine if an alien learned about this particular human quirk. I can see them thinking *"Yikes, what kind of a civilization would you get with a species who routinely go around believing opposite things, always with 99.99…% probability?"*

Well, funny you should ask.

I write a lot about free speech, tolerance of dissenting ideas, open-mindedness, et cetera. You know which posts I'm talking about. There are a lot of reasons to support such a policy. But one of the big ones is – who the heck would burn heretics if they thought there was a 5% chance the heretic was right and they were wrong? Who would demand that dissenting opinions be banned, if they were only about 90% sure of their
own? Who would start shrieking about "human garbage" on Twitter when they fully expected that in some sizeable percent of cases, they would end up being wrong and the garbage right?

Noah Smith recently asked why it was useful to study history. I think at least one reason is to medicate your own overconfidence. I'm not just talking about things like "would Stalin have really killed all those people if he had considered that he was wrong about communism" – especially since I don't think Stalin worked that way. I'm talking about Neville Chamberlain predicting "peace in our time", or the centuries when Thomas Aquinas' philosophy was the preeminent Official Explanation Of Everything. I'm talking about Joseph "no one will ever build a working hot air balloon" Lalande. And yes, I'm talking about what Muggeridge writes about, millions of intelligent people thinking that Soviet Communism was great, and ending out disastrously wrong. Until you see how often people just like you have been wrong in the past, it's hard to understand how uncertain you should be that you are right in the present. If I had lived in 1920s Britain, I probably would have been a Communist. What does that imply about how much I should trust my beliefs today?

There's a saying that "the majority is always wrong". Taken literally it's absurd – the majority thinks the sky is blue, the majority don't believe in the Illuminati, et cetera. But what it *might* mean, is that in a world where everyone is overconfident, the majority will always be wrong about which direction to move the probability distribution in. That is, if an ideal reasoner would ascribe 80% probability to the popular theory and 20% to the unpopular theory, perhaps most real people say 99% popular, 1% unpopular. In that case, if the popular people are urging you to believe the popular theory more, and the unpopular people are urging you to believe the unpopular theory more, the unpopular people are giving you better advice. This would create a strange situation in which good reasoners are usually engaged in disagreeing with the majority, and also usually "arguing for the wrong side" (if you're not good at thinking probablistically, and almost no one is), but remain good reasoners and the ones with beliefs most likely to produce good outcomes. Unless you count "why are all of our good reasoners being burned as witches?" as a bad outcome.

I started off by saying this blog was about "the principle of charity", but I had trouble defining it and in retrospect I'm not that good at it anyway. What can be salvaged from such a concept? I would say "behave the way you would if you were less than insanely overconfident about most of your beliefs." This is the Way. The rest is just commentary.

Discussion Questions (followed by my own answers in ROT13)

1. What is your probability that there is a god? (Svir crepag)
2. What is your probability that psychic powers exist? (Bar va bar gubhfnaq)
3. What is your probability that anthropogenic global warming will increase temperatures by at least 1C by 2050? (Avargl crepag)
4. What is your probability that a pandemic kills at least one billion people in a 5 year

period by 2100? (Svsgrra creprag)
5. What is your probability that humans land on Mars by 2050? (Rvtugl creprag)
6. What is your probability that superintelligent AI (=AI better than almost every human at almost every cognitive task) exists by 2115? (Gjragl svir creprag)

If It’s Worth Doing, It’s Worth Doing With Made-Up Statistics

I do not believe that the utility weights I worked on last week – the ones that say living in North Korea is 37% as good as living in the First World – are objectively correct or correspond to any sort of natural category. So why do I find them so interesting?

A few weeks ago I got to go to a free CFAR tutorial (you can hear about these kinds of things by signing up for their newsletter). During this particular tutorial, Julia tried to explain Bayes’ Theorem to some, er, rationality virgins. I record a heavily-edited-to-avoid-recognizable-details memory of the conversation below:

Julia: So let’s try an example. Suppose there’s a five percent chance per month your computer breaks down. In that case…
Student: Whoa. Hold on here. That’s not the chance my computer will break down.
Julia: No? Well, what do you think the chance is?
Student: Who knows? It might happen, or it might not.
Julia: Right, but can you turn that into a number?
Student: No. I have no idea whether my computer will break. I’d be making the number up.
Julia: Well, in a sense, yes. But you’d be communicating some information. A 1% chance your computer will break down is very different from a 99% chance.
Student: I don’t know the future. Why do you want to me to pretend I do?
Julia: *(who is heroically nice and patient)* Okay, let’s back up. Suppose you buy a sandwich. Is the sandwich probably poisoned, or probably not poisoned?
Student: Exactly which sandwich are we talking about here?

In the context of a lesson on probability, this is a problem I think most people would be able to avoid. But the student’s attitude, the one that rejects hokey quantification of things we don’t actually know how to quantify, is a pretty common one. And it informs a lot of the objections to utilitarianism – the problem of quantifying exactly how bad North Korea shares some of the pitfalls of quantifying exactly how likely your computer is to break (for example, “we are kind of making this number up” is a pitfall).

The explanation that Julia and I tried to give the other student was that imperfect information still beats zero information. Even if the number “five percent” was made up (suppose that this is a new kind of computer being used in a new way that cannot be easily compared to longevity data for previous computers) it encodes our knowledge that computers are unlikely to break in any given month. Even if we are wrong by a very large amount (let’s say we’re off by a factor of four and the real number is 20%), if the insight we encoded into the number is sane we’re still doing better than giving no information at all (maybe model this as a random number generator which chooses anything from 0 – 100?)

This is part of why I respect utilitarianism. Sure, the actual badness of North Korea may not be exactly 37%. But it’s probably not twice as good as living in the First World. Or even 90% as good. But it’s probably not two hundred times worse than death either. There is definitely nonzero information transfer going on here.

But the typical opponents of utilitarianism have a much stronger point than the guy at the CFAR class. They're not arguing that utilitarianism fails to outperform zero information, they're arguing that it fails to outperform our natural intuitive ways of looking at things, the one where you just think "North Korea? Sounds awful. The people there deserve our sympathy."

Remember the Bayes mammogram problem? The correct answer is 7.8%; most doctors (and others) intuitively feel like the answer should be about 80%. So doctors – who are specifically trained in having good intuitive judgment about diseases – are wrong by an order of magnitude. And it "only" being *one* order of magnitude is not to the doctors' credit: by changing the numbers in the problem we can make doctors' answers as wrong as we want.

So the doctors probably would be better off explicitly doing the Bayesian calculation. But suppose some doctor's internet is down (you have NO IDEA how much doctors secretly rely on the Internet) and she can't remember the prevalence of breast cancer. If the doctor thinks her guess will be off by less than an order of magnitude, then making up a number and plugging it into Bayes will be more accurate than just using a gut feeling about how likely the test is to work. Even making up numbers based on basic knowledge like "Most women do not have breast cancer at any given time" might be enough to make Bayes Theorem outperform intuitive decision-making in many cases.

And a *lot* of intuitive decisions are off by way more than the make-up-numbers ability is likely to be off by. Remember that scope insensitivity experiment where people were willing to spend about the same amount of money to save 2,000 birds as 200,000 birds? And the experiment where people are willing to work harder to save one impoverished child than fifty impoverished children? And the one where judges give criminals several times more severe punishments on average just before they eat lunch than just after they eat lunch?

And it's not just neutral biases. We've all seen people who approve wars under Republican presidents but are *horrified* by the injustice and atrocity of wars under Democratic presidents, even if it's just the same war that carried over to a different administration. If we forced them to stick a number on the amount of suffering caused by war before they knew what the question was going to be, that's a bit harder.

Thus is it written: "It's easy to lie with statistics, but it's easier to lie without them."

Some things work okay on System 1 reasoning. Other things work badly. Really really badly. Factor of a hundred badly, if you count the bird experiment.

It's hard to make a mistake in calculating the utility of living in North Korea that's off by a factor of *a hundred*. It's hard to come up with values that make a war suddenly become okay/abominable when the President changes parties.

Even if your data is completely made up, the way the 5% chance of breaking your computer was made up, the fact that you can apply normal non-made-up arithmetic to these made-up numbers will mean that you will very often *still* be less wrong than if you had used your considered and thoughtful and phronetic opinion.

On the other hand, it's pretty easy to accidentally Pascal's Mug yourself into giving everything you own to a crazy cult, which System 1 is good at avoiding. So it's nice to have data from both systems.

In cases where we really don’t know what we’re doing, like utilitarianism, one can still make System 1 decisions, but making them with the System 2 data in front of you can change your mind. Like “Yes, do whatever you want here, just be aware that X causes two thousand people to die and Y causes twenty people an amount of pain which, in experiments, was rated about as bad as a stubbed toe”.

And cases where we don’t really know what we’re doing have a wonderful habit of developing into cases where we *do* know what we’re doing. Like in medicine, people started out with “doctors’ clinical judgment obviously trumps everything, but just in case some doctors forgot to order clinical judgment, let’s make some toy algorithms”. And then people got better and better at crunching numbers and now there are cases where doctors should never use their clinical judgment under any circumstances. I can’t find the article right now, but there are even cases where doctors armed with clinical algorithms consistently do worse than clinical algorithms without doctors. So it looks like at some point the diagnostic algorithm people figured out what they were doing.

I generally support applying made-up models to pretty much any problem possible, just to notice where our intuitions are going wrong and to get a second opinion from a process that has no common sense but is also lacks systematic bias (or else has unpredictable, different systematic bias).

This is why I’m disappointed that no one has ever tried expanding the QALY concept to things outside health care before. It’s not that I think it will work. It’s that I think it will fail to work in a different way than our naive opinions fail to work, and we might learn something from it.

EDIT: Edited to include some examples from the comments. I also really like ciphergoth’s quote: “Sometimes pulling numbers out of your arse and using them to make a decision is better than pulling a decision out of your arse.”

Techniques for probability estimates

Utility maximization often requires determining a probability of a particular statement being true. But humans are not utility maximizers and often refuse to give precise numerical probabilities. Nevertheless, their actions reflect a "hidden" probability. For example, even someone who refused to give a precise probability for Barack Obama's re-election would probably jump at the chance to take a bet in which ey lost $5 if Obama wasn't re-elected but won $5 million if he was; such decisions demand that the decider covertly be working off of at least a vague probability.

When untrained people try to translate vague feelings like "It seems Obama will probably be re-elected" into a precise numerical probability, they commonly fall into certain traps and pitfalls that make their probability estimates inaccurate. Calling a probability estimate "inaccurate" causes philosophical problems, but these problems can be resolved by remembering that probability is "subjectively objective" - that although a mind "hosts" a probability estimate, that mind does not arbitrarily determine the estimate, but rather calculates it according to mathematical laws from available evidence. These calculations require too much computational power to use outside the simplest hypothetical examples, but they provide a standard by which to judge real probability estimates. They also suggest tests by which one can judge probabilities as well-calibrated or poorly-calibrated: for example, a person who constantly assigns 90% confidence to eir guesses but only guesses the right answer half the time is poorly calibrated. So calling a probability estimate "accurate" or "inaccurate" has a real philosophical grounding.

There exist several techniques that help people translate vague feelings of probability into more accurate numerical estimates. Most of them translate probabilities from forms without immediate consequences (which the brain supposedly processes for signaling purposes) to forms with immediate consequences (which the brain supposedly processes while focusing on those consequences).

Prepare for Revelation

What would you expect if you believed the answer to your question were about to be revealed to you?

In Belief in Belief, a man acts as if there is a dragon in his garage, but every time his neighbor comes up with an idea to test it, he has a reason why the test wouldn't work. If he imagined Omega (the superintelligence who is always right) offered to reveal the answer to him, he might realize he was expecting Omega to reveal the answer "No, there's no dragon". At the very least, he might realize he was worried that Omega would reveal this, and so re-think exactly how certain he was about the dragon issue.

This is a simple technique and has relatively few pitfalls.

Bet on it

At what odds would you be willing to bet on a proposition?

Suppose someone offers you a bet at even odds that Obama will be re-elected. Would you take it? What about two-to-one odds? Ten-to-one? In theory, the knowledge that money is at stake should make you consider the problem in "near mode" and maximize your chances of winning.

The problem with this method is that it only works when utility is linear with respect to money and you're not risk-averse. In the simplest case I should be indifferent to a $100,000 bet at 50% odds that a fair coin would come up tails, but in fact I would refuse it; winning $100,000 would be moderately good, but losing $100,000 would put me deeply in debt and completely screw up my life. When these sorts of consideration become paramount, imagining wagers will tend to give inaccurate results.

Convert to a Frequency

How many situations would it take before you expected an event to occur?

Suppose you need to give a probability that the sun will rise tomorrow. "999,999 in a million" doesn't immediately sound wrong; the sun seems likely to rise, and a million is a very high number. But if tomorrow is an average day, then your probability will be linked to the number of days it will take before you expect that the sun will fail to rise on at least one. A million days is three thousand years; the Earth has existed for far more than three thousand years without the sun failing to rise. Therefore, 999,999 in a million is too low a probability for this occurrence. If you think the sort of astronomical event that might prevent the sun from rising happens only once every three billion years, then you might consider a probability more like 999,999,999,999 in a trillion.

In addition to converting to a frequency across time, you can also convert to a frequency across places or people. What's the probability that you will be murdered tomorrow? The best guess would be to check the murder rate for your area. What's the probability there will be a major fire in your city this year? Check how many cities per year have major fires.

This method fails if your case is not typical: for example, if your city is on the losing side of a war against an enemy known to use fire-bombing, the probability of a fire there has nothing to do with the average probability across cities. And if you think the reason the sun might not rise is a supervillain building a high-tech sun-destroying machine, then consistent sunrises over the past three thousand years of low technology will provide little consolation.

A special case of the above failure is converting to frequency across time when considering an event that is known to take place at a certain distance from the present. For example, if today is April 10th, then the probability that we hold a Christmas celebration tomorrow is much lower than the 1/365 you get by checking on what percentage of days we celebrate Christmas. In the same way, although we know that the sun will fail to rise in a few billion years when it burns out its nuclear fuel, this shouldn't affect its chance of rising tomorrow.

Find a Reference Class

How often have similar statements been true?

What is the probability that the latest crisis in Korea escalates to a full-blown war? If there have been twenty crisis-level standoffs in the Korean peninsula in the past 60 years, and only one of them has resulted in a major war, then (war|crisis) = .05, so long as this crisis is equivalent to the twenty crises you're using as your reference class.

But finding the reference class is itself a hard problem. What is the probability Bigfoot exists? If one makes a reference class by saying that the yeti doesn't exist, the Loch Ness monster doesn't exist, and so on, then the Bigfoot partisan might accuse you of assuming the conclusion - after all, the likelihood of these creatures existing is probably similar to and correlated with Bigfoot. The partisan might suggest asking how many creatures previously believed not to exist later turned out to exist - a list which includes real animals like the orangutan and platypus - but then one will have to debate whether to include creatures like dragons, orcs, and Pokemon on the list.

This works best when the reference class is more obvious, as in the Korea example.

Make Multiple Statements

How many statements could you make of about the same uncertainty as a given statement without being wrong once?

Suppose you believe France is larger than Italy. With what confidence should you believe it? If you made ten similar statements (Germany is larger than Austria, Britain is larger than Ireland, Spain is larger than Portugal, et cetera) how many times do you think you would be wrong? A hundred similar statements? If you think you'd be wrong only one time out of a hundred, you can give the statement 99% confidence.

This is the most controversial probability assessment technique; it tends to give lower levels of confidence than the others; for example, Eliezer wants to say there's a less than one in a million chance the LHC would destroy the world, but doubts he could make a million similar statements and only be wrong once. Komponisto thinks this is a failure of imagination: we imagine ourselves gradually growing tired and making mistakes, whereas this method only works if the accuracy of the millionth statement is exactly the same as the first.

In any case, the technique is only as good as the ability to judge which statements are equally difficult to a given statement. If I start saying things like "Russia is larger than Vatican City! Canada is larger than a speck of dust!" then I may get all the statements right, but it won't mean much for my Italy-France example - and if I get bogged down in difficult questions like "Burundi is larger than Equatorial Guinea" then I might end up underconfident. In cases where there is an obvious comparison ("Bob didn't cheat on his test", "Sue didn't cheat on her test", "Alice didn't cheat on her test") this problem disappears somewhat.

Imagine Hypothetical Evidence

How would your probabilities adjust given new evidence?

Suppose one day all the religious people and all the atheists get tired of arguing and decide to settle the matter by experiment once and for all. The plan is to roll an n-sided numbered die and have the faithful of all religions pray for the die to land on "1". The experiment will be done once, with great pomp and ceremony, and never repeated, lest the losers try for a better result. All the resources of the world's skeptics and security forces will be deployed to prevent any tampering with the die, and we assume their success is guaranteed.

If the experimenters used a twenty-sided die, and the die comes up 1, would this convince you that God probably did it, or would you dismiss the result as a coincidence? What about a hundred-sided die? Million-sided? If a successful result on a hundred-sided die wouldn't convince you, your probability of God's existence must be less than one in a hundred; if a million-sided die would convince you, it must be more than one in a million.

This technique has also been denounced as inaccurate, on the grounds that our coincidence detectors are overactive and therefore in no state to be calibrating anything else. It would feel very hard to dismiss a successful result on a thousand-sided die, no matter how low the probability of God is. It might also be difficult to visualize a hypothetical where the experiment can't possibly be rigged, and it may be unfair to force subjects to imagine a hypothetical that would practically never happen (like the million-sided die landing on one in a world where God doesn't exist).

These techniques should be experimentally testable; any disagreement over which do or do not work (at least for a specific individual) can be resolved by going through a list of difficult questions, declaring confidence levels, and scoring the results with log odds. Steven's blog has some good sets of test questions (which I deliberately do *not* link here so as to not contaminate a possible pool of test subjects); if many people are interested in participating and there's a general consensus that an experiment would be useful, we can try to design one.

Confidence levels inside and outside an argument

Related to: Infinite Certainty

Suppose the people at FiveThirtyEight have created a model to predict the results of an important election. After crunching poll data, area demographics, and all the usual things one crunches in such a situation, their model returns a greater than 999,999,999 in a billion chance that the incumbent wins the election. Suppose further that the results of this model are your only data and you know nothing else about the election. What is your confidence level that the incumbent wins the election?

Mine would be significantly less than 999,999,999 in a billion.

When an argument gives a probability of 999,999,999 in a billion for an event, then probably the majority of the probability of the event is no longer in "But that still leaves a one in a billion chance, right?". The majority of the probability is in "That argument is flawed". Even if you have no particular reason to believe the argument is flawed, the background chance of an argument being flawed is still greater than one in a billion.

More than one in a billion times a political scientist writes a model, ey will get completely confused and write something with no relation to reality. More than one in a billion times a programmer writes a program to crunch political statistics, there will be a bug that completely invalidates the results. More than one in a billion times a staffer at a website publishes the results of a political calculation online, ey will accidentally switch which candidate goes with which chance of winning.

So one must distinguish between levels of confidence internal and external to a specific model or argument. Here the model's internal level of confidence is 999,999,999/billion. But my external level of confidence should be lower, even if the model is my only evidence, by an amount proportional to my trust in the model.

Is That Really True?

One might be tempted to respond "But there's an equal chance that the false model is too high, versus that it is too low." Maybe there was a bug in the computer program, but it prevented it from giving the incumbent's real chances of 999,999,999,999 out of a *trillion*.

The prior probability of a candidate winning an election is 50%[1]. We need information to push us away from this probability in either direction. To push significantly away from this probability, we need strong information. Any weakness in the information weakens its ability to push away from the prior. If there's a flaw in FiveThirtyEight's model, that takes us away from their probability of 999,999,999 in of a billion, and back closer to the prior probability of 50%

We can confirm this with a quick sanity check. Suppose we know nothing about the election (ie we still think it's 50-50) until an insane person reports a hallucination that an angel has declared the incumbent to have a 999,999,999/billion chance. We would not be tempted to accept this figure on the grounds that it is equally likely to be too high as too low.

A second objection covers situations such as a lottery. I would like to say the chance that Bob wins a lottery with one billion players is 1/1 billion. Do I have to adjust this upward to cover the possibility that my model for how lotteries work is somehow flawed? No. Even if I am misunderstanding the lottery, I have not departed from my prior. Here, new information really does have an equal chance of going against Bob as of going in his favor. For example, the lottery may be fixed (meaning my original model of how to determine lottery winners is fatally flawed), but there is no greater reason to believe it is fixed in favor of Bob than anyone else.[2]

Spotted in the Wild

The recent Pascal's Mugging thread spawned a discussion of the Large Hadron Collider destroying the universe, which also got continued on an older LHC thread from a few years ago. Everyone involved agreed the chances of the LHC destroying the world were less than one in a million, but several people gave extraordinarily low chances based on cosmic ray collisions. The argument was that since cosmic rays have been performing particle collisions similar to the LHC's zillions of times per year, the chance that the LHC will destroy the world is either literally zero, or else a number related to the probability that there's some chance of a cosmic ray destroying the world so miniscule that it hasn't gotten actualized in zillions of cosmic ray collisions. Of the commenters mentioning this argument, one gave a probability of 1/3*10^22, another suggested 1/10^25, both of which may be good numbers for the internal confidence of this argument.

But the connection between this argument and the general LHC argument flows through statements like "collisions produced by cosmic rays will be exactly like those produced by the LHC", "our understanding of the properties of cosmic rays is largely correct", and "I'm not high on drugs right now, staring at a package of M&Ms and mistaking it for a really intelligent argument that bears on the LHC question", all of which are probably more likely than 1/10^20. So instead of saying "the probability of an LHC apocalypse is now 1/10^20", say "I have an argument that has an internal probability of an LHC apocalypse as 1/10^20, which lowers my probability a bit depending on how much I trust that argument".

In fact, the argument has a potential flaw: according to Giddings and Mangano, the physicists officially tasked with investigating LHC risks, black holes from cosmic rays might have enough momentum to fly through Earth without harming it, and black holes from the LHC might not[3]. This was predictable: this was a simple argument in a complex area trying to prove a negative, and it would have been presumptous to believe with greater than 99% probability that it was flawless. If you can only give 99% probability to the argument being sound, then it can only reduce your probability in the conclusion by a factor of a hundred, not a factor of 10^20.

But it's hard for me to be properly outraged about this, since the LHC did not destroy the world. A better example might be the following, taken from an online discussion of creationism[4] and apparently based off of something by Fred Hoyle:

> In order for a single cell to live, all of the parts of the cell must be assembled before life starts. This involves 60,000 proteins that are assembled in roughly 100 different combinations. The probability that these complex groupings of proteins could have happened just by chance is extremely small. It is about 1 chance in 10 to the 4,478,296 power. The probability of a living cell being assembled just by chance is so small, that you may as well consider it to be impossible. This means that the probability that the living cell is created by an intelligent creator, that designed it, is extremely large. The probability that God created the living cell is 10 to the 4,478,296 power to 1.

Note that someone just gave a confidence level of 10^4478296 to one and was wrong. This is the sort of thing that should *never ever happen*. This is possibly the *most wrong anyone has ever been*.

It is hard to say in words exactly how wrong this is. Saying "This person would be willing to bet the entire world GDP for a thousand years if evolution were true against a one in one million chance of receiving a single penny if creationism were true" doesn't even begin to cover it: a mere 1/10^25 would suffice there. Saying "This person believes he could make one statement about an issue as difficult as the origin of cellular life per Planck interval, every Planck interval from the Big Bang to the present day, and not be wrong even once" only brings us to 1/10^61 or so. If the chance of getting Ganser's Syndrome, the extraordinarily rare psychiatric condition that manifests in a compulsion to say false statements, is one in a hundred million, and the world's top hundred thousand biologists all agree that evolution is true, then this person should preferentially believe it is more likely that all hundred thousand have simultaneously come down with Ganser's Syndrome than that they are doing good biology[5]

This creationist's flaw wasn't mathematical; the math probably does return that number. The flaw was confusing the internal probability (that complex life would form completely at random in a way that can be represented with this particular algorithm) with the external probability (that life could form without God). He should have added a term representing the chance that his knockdown argument just didn't apply.

Finally, consider the question of whether you can assign 100% certainty to a mathematical theorem for which a proof exists. Eliezer has already examined this issue and come out against it (citing as an example this story of Peter de Blanc's). In fact, this is just the specific case of differentiating internal versus external probability when internal probability is equal to 100%. Now your probability that the theorem is false is entirely based on the probability that you've made some mistake.

The many mathematical proofs that were later overturned provide practical justification for this mindset.

This is not a fully general argument against giving very high levels of confidence: very complex situations and situations with many exclusive possible outcomes (like the lottery example) may still make it to the 1/10^20 level, albeit probably not the 1/10^4478296. But in other sorts of cases, giving a very high level of confidence requires a check that you're not confusing the probability inside one argument with the probability of the question as a whole.

Footnotes

1. Although technically we know we're talking about an incumbent, who typically has a much higher chance, around 90% in Congress.

2. A particularly devious objection might be "What if the lottery commissioner, in a fit of political correctness, decides that "everyone is a winner" and splits the jackpot a billion ways? If this would satisfy your criteria for "winning the lottery", then this mere possibility should indeed move your probability upward. In fact, since there is probably greater than a one in one billion chance of this happening, the majority of your probability for Bob winning the lottery should concentrate here!

3. Giddings and Mangano then go on to re-prove the original "won't cause an apocalypse" argument using a more complicated method involving white dwarf stars.

4. While searching creationist websites for the half-remembered argument I was looking for, I found what may be my new favorite quote: "Mathematicians generally agree that, statistically, any odds beyond 1 in 10 to the 50th have a zero probability of ever happening."

5. I'm a little worried that five years from now I'll see this quoted on some creationist website as an actual argument.

The Logician And The God-Emperor

Once upon a time a logician accomplished a great deed, and the God-Emperor offered him a choice of rewards. “You may,” said the God-Emperor “have the hand of my eldest daughter, who is the heir to the throne, yet plain to look upon. Or you may take my youngest daughter, who is beautiful beyond words, but without inheritance.”

The next day, the God-Emperor caught the logician in bed with *both* his daughters. Enraged, he hurled threats and abuse at the scholar, who responded with a grin: “Guess someone never learned the difference between ‘or’ and ‘xor’.”

The God-Emperor ordered the logician brought to the throne room in chains, and told him “You have offended me and betrayed my generosity, so you will be subjected to trial by ordeal. I have placed in front of you seven chests. Six of the chests contain skulls. One of the chests contains the key to your chains. I have asked the most devious minds in my kingdom to prepare a logic puzzle giving hints as to which chest is which. You may open a single chest. If you do not find the chest with the key on your first try, you will be slathered in barbecue sauce and thrown to the wolves.”

The logician approached the chests, and upon each was written a clue in complicated logical notation. He examined all seven, and then stood a while, deep in thought. Finally, he opened the third chest. Inside was a golden key.

“Very impressive!” said the God-Emperor. Then he yelled “Guards! Slather this man in barbecue sauce and throw him to the wolves!”

“But…but!” babbled the terrified logician “…but you said…!”

The God-Emperor grinned. “Guess someone never learned the difference between ‘if’ and ‘iff’.”

Reverse Psychology

[Content warning: suicide]

I.

It all started when I made that phone call.

I was really bad. All the tenure-track positions I'd applied to had politely declined, and I saw my future in academia gradually slipping away from me. Then the night before, my boyfriend had said he thought maybe we should start seeing other people. I didn't even know if we were broken up or not, and at that point I couldn't bring myself to care. I sat on my bed, thinking about things for a while, and finally I called the suicide hotline.

"Hello?" a woman's voice answered on the other side. Somehow, just hearing someone else made me feel about five times better.

"Hello," I said, a little more confidently. "I've been thinking of committing suicide. I need help."

"Okay," she said. "Is there a gun in your house?"

"No."

"All right. The first thing you need to do is get one. Overdosing on pills is common, but it almost never works. You can get a firearm at almost any large sporting goods store, but if there aren't any near you, we can start talking about maybe jumping from a high…"

"What the HELL?" I interrupted, suddenly way more angry than depressed. "You're supposed to @#!$ing tell me not to do it!"

"This is the suicide hotline," the woman said, now sounding confused. Then, "Are you sure you weren't thinking of the suicide *prevention* hotline?"

"Give me a break! I took a psychology class in undergrad, I know what a suicide hotline is!"

"I'm sorry you seem to be upset. But this is the suicide hotline. It's like how there's the Walk For Breast Cancer, but also the Walk Against Breast Cancer."

"There's the what? But…I was *in* the Walk For Breast Cancer! I thought…"

"It sounds like you have some issues," said the woman, politely.

"Ugh," I said. "Yeah."

"Do you feel like you need professional help?"

"Yeah."

"I do have a free clinic with an opening available tomorrow at three PM, would you like me to slot you in for an appointment?"

So you're probably wondering why in the world I would take an appointment arranged by the suicide hotline that wasn't a suicide prevention hotline. The answer is – were you even listening? A free clinic? With an appointment available the next day? Normally I was lucky if I found a place with an opening in less than two months and a co-pay that wasn't completely ruinious. You *bet* I was taking that appointment before someone else snatched it up.

Dr. Trauer's office looked gratifyingly normal. There was a houseplant, a diagram of the cranial nerves, some Abilify® merchandise, and on the wall one of those Magic Eye stereographic images that resolved into a 3D picture of the human brain. Dr. Trauer himself looked like your average doctor – a little past middle age, a little overweight, a short greying beard. He motioned me to sit down and took the paperwork I'd been filling out.

"Hmmmm," he said, reading it over. "29 years old, postdoc in biochem, recent relationship trouble…mmmm…you did the right thing."

"In coming here?"

"No, in considering suicide. After getting rejected from a tenure-track position, your life is pretty much over."

"WHAT?"

"I mean, here you are, hundreds of thousands of dollars in debt, with only one area of expertise, and now you've been rejected from it. I can totally see why you might think it's worth ending it all."

"But…there are lots of other things I can do! I can get a job in industry! I can work in something else! Even if I can't find a job right away, I have parents who can help support me."

"Industry!" Dr. Trauer was having none of it. "A bunch of bloodsuckers. Do you realize how bad work in the private sector is these days? They'll abuse you and then spit you out, and once you've been out of university too long nobody else will want you."

"Lots of people want biochemists! If I work for a company for a few years, I'll have more experience and maybe that will make me more attractive to employers! What…what kind of a psychiatrist *are* you, anyway?"

"Cindy didn't tell you?"

"Cindy?"

"The woman on the phone."

"She didn't really tell me anything!"

"Well," said Dr. Trauer. "To answer your question, we're dark side psychiatrists. This is the state's only dark side psychiatry clinic."

"Dark side psychiatry? *Really?*"

"We're a…well, some people say sect, but I like to think of it as more of a guild…dedicated to improving negative mental health. Think of it this way. When you're a hijacked murder-monkey hurtling toward your inevitable death, sanity is a completely ridiculous thing to have. And when the universe is fifteen billion light-years across and almost entirely freezing void, the idea that people should have ‘coping skills’ boggles the imagination. An emotionally healthy person is a person who isn’t paying attention, and our job is to cure them.”

“There’s more than one of you?”

“Oh, yes. There’s a thriving dark side psychiatric community. There are dark side psychopharmacologists – you’d be amazed what a few doses of datura can do to a person. There are dark side psychotherapists who analyze and break down people’s positive cognitions. There are dark side child psychiatrists who catch people when they’re young, before sanity has had a chance to take root and worsen. And there are dark side geriatric psychiatrists, who go from nursing home to nursing home, making sure that the elderly are not warehoused and neglected at exactly the time it is most important to ensure that stroke or dementia does not protect them from acute awareness of the nearness of death.”

“That’s awful!” I said.

“Is it? Look where sanity’s gotten you. You want to kill yourself, but you don’t have the courage. Work with me for ten sessions, and I promise you we can help you *get* that courage.”

“You’re a @#!$ing quack,” I said. “And if you think killing yourself is so great, how come you haven’t done it yourself yet?”

“Who says I haven’t?” asked Dr. Trauer.

His hand went to his face, and he plucked out his right eye, revealing an empty void surrounded by the bleached whiteness of bone. I screamed and ran out of the clinic and didn’t stop running until I was in my house and had locked the door beside me.

II.

“…and that’s pretty much the whole story, doctor,” she told me. “And then I looked to see if there were any *real* psychiatrists in the area and someone referred me to you.”

“Well,” I said, my face unreadable. “I can certainly see why you’re complaining of, how did you put it, ‘depression and acute stress disorder’.”

“Not so acute anymore. It took me two months to get an appointment at your clinic.”

“Oh,” I said. Then, “Sorry, we’re sort of backed up.” Then, “Okay. We’ve got a lot we have to work on here. Let me tell you how we’re going to do it. We’re going to use a form of therapy that challenges your negative cognitions. We’re going to take the things that are bothering you, examine the evidence for them, and see if there are alternative explanations.”

“What do you mean?” she asked.

“Well,” I said. “It seems to be this Dr. Trauer incident that’s traumatized you a lot. I can see why you would be stressed out. The way you tell it, it sounds absolutely terrifying.”

“You don’t believe me,” she said, not accusatory, just stating a fact.

“I think it would be helpful to examine alternate explanations,” I said. “I’m willing to assume it happened exactly as you tell it. I can see why you would think Dr. Trauer wanted you to commit suicide. But are there any alternative explanations for the same event?”

“I don’t see how there can be,” she said. “He outright said that he thought I should kill myself.”

“Right. But from what you know of psychiatrists and therapy – and you did say you took some classes in undergrad – are there any other reasons he might have said something like that?”

She thought for a second. “Wait,” she told me. “There’s a technique in therapy called paradoxical intention. Where you take a patient’s irrational thought, and then defend and amplify it. And then when the patient hears it from someone else, she realizes how silly it sounds and starts arguing against it, and then it’s really hard to keep believing it after you’ve shot it down yourself.”

I nodded. “That’s definitely a therapeutic method, and sometimes a very effective one. Do you have any evidence that this is what Dr. Trauer was doing?”

“Yes! As soon as he said I should commit suicide, I started arguing against him. He told me that if I couldn’t get a tenure track position there would be no other jobs available, and I told him there would be! Then he told me that the jobs would be terrible and I’d never be able to make a happy life for myself with them, and I argued that I would! That must have been what he was going for!”

She suddenly looked really excited. Then, just as suddenly, the worry returned to her face.

“But then what happened with his eye? I swear I saw him take it right out of the socket.”

I nodded. “Can you think of any alternate explanations for that?”

Thinking about it that way, it only took her like five seconds. She slapped her head like she’d been an idiot. “A glass eye. He probably had some kind of injury, had to put in a glass eye, and could take it out any time he wanted. He must have thought it would be a funny gag and didn’t realize how traumatized I’d be. Or he wanted to scare me into realizing how much I wanted to live. Or something.”

I nodded. “That does sound like a reasonable explanation.”

“But…don’t people with glass eyes usually have like scar tissue and normal skin behind them? This guy, I swear it was just the bone and this empty socket, like you were seeing straight to his skull.”

“You’re asking the right questions,” I said. “Now think a little more.”

“Hmmmm,” she said. “I guess I was really, really stressed out at the time. And I only saw it for, like, a fraction of a second. Maybe my brain was playing tricks on me.”

“That can definitely happen,” I agreed.

She looked a lot better now. “I owe you a lot of thanks,” she said. “I’ve only been here for, like, fifteen minutes, and already I think a lot of my stress has gone away. All of this really makes sense. That paradoxical intention thing is actually kind of brilliant. And I can’t deny that it worked – I haven’t been suicidal since I talked to the guy. In fact…okay, this is going to sound really strange, but…maybe I should go back to Dr. Trauer.”

I wrinkled my forehead.

“It’s not that I don’t like you,” she said. “But he had this amazing free clinic, and what he did for me that day…now that I realize what was going on, that was actually pretty incredible.”

“Hold on a second,” I said.

I left the room, marched up to the front desk, took the directory of medical providers in the area off the shelf, marched back to the room. I started flipping through the pages. It was in alphabetical order…Tang…Thompson…Tophet…there we go. Trauer. My gaze lingered there maybe just a second too long, and she asked if I was okay.

“Um, yeah,” I said. “It’s just that he doesn’t – he doesn’t take your insurance. That’s the problem.”

“It’s okay,” she told me. “He said it was a free clinic. So that shouldn’t a problem.”

“Well, uh…the thing is…when you see out-of-network providers, your insurance actually charges, charges an extra fee. Even if the visit itself is free.”

She looked skeptical. “I’ve never heard of that.”

“It’s new. With Obamacare.”

“Really? How high a fee is it?”

“It’s…um…ten thousand dollars. Yeah, I know, right? Thanks, Obama.”

“Wow,” she said. “I definitely can’t afford that. I guess I’ll keep coming here. Not that there’s anything wrong with that. You’ve been very nice. It’s just that…with Dr. Trauer…well…sorry, I’ll stop talking now. Thanks a lot, doctor.” She stood up and shook my hand before heading for the door. “Seriously, I can’t believe how much you’ve helped me.”

No, I thought, as she departed *you can’t*. I told her she was asking the right questions, and she was, but not all of them.

For example, *why would a man with only one working eye have a stereographic Magic Eye image in his office?*

I picked up my provider directory again, stared a second time at the entry for Dr. Trauer. There was a neat line through it in red pen, and above, in my secretary’s careful handwriting, “DECEASED”.

Before returning the directory to the front desk, I took my own pen and added “DO NOT REFER” in big letters underneath.