*Epistemic status: I wrote this post quickly, and largely to solicit feedback on the claims I make in it. This is because (a) I'm not sure about these claims (or how I've explained them), and (b) the question of what I **should** believe on this topic seems important in general and for various other posts I'm writing. (So please comment if you have any thoughts on this!)*

*I've now read a bunch on topics **related** to the questions covered here, but I'm not an expert, and haven't seen or explicitly looked for a **direct** treatment of the questions covered here. It's very possible this has already been thoroughly and clearly covered elsewhere; if so, please comment the link!*

I lean towards the idea that we can always assign probabilities to propositions (or at least use something like an uninformative prior), even if sometimes we have incredibly little basis for making those probabilities. Sometimes people propose what seem to me to be *very weak* counterexamples to that claim, such as the following:

> there are situations with so many unique features that they can hardly be grouped with similar cases, such as the danger resulting from a new type of virus, or the consequences of military intervention in conflict areas. These represent cases of (Knightian) uncertainty where no data are available to estimate objective probabilities. While we may rely on our subjective estimates under such conditions, no objective basis exists by which to judge them (e.g., LeRoy & Singell, 1987). (source)

It seems obvious to me that a *wealth* of data *is* available for such cases. There have been *many* viruses and military interventions before. None of those situations will perfectly mirror the situations we're trying to predict, and that's definitely a very important point. We should therefore think very carefully about whether we're being too confident in our predictions (i.e., using too narrow a "confidence interval"[1] and thus not adequately preparing for especially "high" or "low" possibilities).

But we can clearly do better than nothing. To start small, you'd be comfortable with the claim that a new type of virus, if it hits this year, is more likely to kill somewhere between 0 and 1 billion people than somewhere between 1000 and 1001 billion people (i.e., far more than everyone alive), right? And in fact, we have *empirical evidence* that some people can reliably do better than chance (and better than "0 to 1 billion") in making predictions about geopolitical events like these, at least over timelines of a few years (from Tetlock's work).

[1] See this shortform post of mine for other ways of describing the idea that our probabilities might be relatively "untrustworthy".

AGI

What about something that seems more unique or unprecedented, and where we also may have to stretch our predictions further into the future, like artificial general intelligence (AGI) timelines? On that question, experts disagree wildly, and are seemingly quite swayed by things like how the question is asked (Katja Grace on 80k; search for "It's a bit complicated" in the transcript"). This makes me *highly* unconfident in any prediction I might make on the topic (and thus pushes me towards making decisions that are good given a wide range of possible timelines).

But I believe I know *more than nothing*. I believe I can reasonably assign *some probability distribution* (and then use something like the median or mean of that as if it were a point estimate, for certain purposes). If that seems like raw hubris, do you think it's worth actually behaving as if AGI is just as likely to be developed 1 minute from now as somewhere around 2 to 300 years from now? What about behaving as if it's likely to occur in some millennium 50 quintillion years from now, and not in this millennium? So you'd at least be fairly happy bounding your probability distribution somewhere in between those points 1 minute from now and 50 quintillion years from now, right?

One could say that all I've done there is argue that some probabilities we could assign would seem especially outrageous, not that we really can or should assign probabilities to this event. But if some probabilities *are* more reasonable than others (and it certainly *seems* they are, though I can't prove it), then we can do better by using those probabilities than by using something like an uninformative prior.[2] And as far as I'm aware, principles for decision making *without* probabilities essentially collapse to acting as if using an uninformative prior or predictably lead to seeming irrational and bad decisions (I'll be posting about this soon).

And in any case, we *do* have *relevant* data for the AGI question, even if we've never developed *AGI itself* - we have data on AI development more broadly, development related to computing/IT/robotics more broadly, previous transformative technologies (e.g., electricity), the current state of funding for AI, current governmental stances towards AI development, how funding and governmental stances have influenced tech in the past, etc.

[2] I think that my "1 minute" example doesn't demonstrate the superiority of certain probability distributions to an uninformative prior. This is because we could argue that the issue there is that "1 minute from now" is far more *precise* than "2 to 300 years from now", and an uninformative prior would favour the less precise prediction, just as we'd like it too. But I think my other example *does* indicate, if our intuitions on that are trustworthy, that some probability distributions can be superior to an uninformative prior. This is because, in that example, predictions mentioned spanned the same amount of time (a millennium), just starting at different points (~now vs ~50 quintillion years from now).

Supernatural-type claims

But that leads me to what *does* seem like it *could* be a strong type of counterexample to the idea that we can always assign probabilities: claims of a "supernatural", "metaphysical", or "unobservable" nature. These are very fuzzy and debatable terms, but defining them isn't my main purpose here, so instead I'll just jump into some examples:

1. What are the odds that "an all-powerful god" exists?
2. What are the odds that "ghosts" exist?
3. What are the odds that "magic" exists?
4. What are the odds that "non-naturalistic moral realism" is correct (or that "non-natural objective moral facts" exist)?[3]

My intuitions would suggest I should assign a very low probability to each of these propositions.[4] But what basis would I have for that? More specifically, what basis would I have for any *particular* probability (or probability distribution) I assign? And what would it even mean?

This is Chris Smith's statement of this apparent issue, which was essentially what prompted this post:

> Kyle is an atheist. When asked what odds he places on the possibility that an all-powerful god exists, he says "2%."
>
> [...] I don't know what to make of [Kyle's] probability estimate.
>
> [Kyle] wouldn't be able to draw on past experiences with different realities (i.e., Kyle didn't previously experience a bunch of realities and learn that some of them had all-powerful gods while others didn't). If you push someone like Kyle to explain why they chose 2% rather than 4% or 0.5%, you almost certainly won't get a clear explanation.
>
> If you gave the same "What probability do you place on the existence of an all-powerful god?" question to a number of self-proclaimed atheists, you'd probably get a wide range of answers.
>
> I bet you'd find that some people would give answers like 10%, others 1%, and others 0.001%. While these probabilities can all be described as "low," they differ by orders of magnitude. If probabilities like these are used alongside probabilistic decision models,

[3] These terms can be defined in many different ways. Footnote 15 of this is probably a good quick source. This page is also relevant, but I've only skimmed it myself.

[4] Though in the case of non-naturalistic moral realism, I might still act as though it's correct, to a substantial extent, based on a sort of expected value reasoning or Pascal's wager. But I'm not sure if that makes sense, and it's not directly relevant for the purposes of this post. (I hope to write a separate post about that idea later.)

> they could have extremely different implications. Going forward, I'm going to call probability estimates like these "hazy probabilities."

I can sympathise with Smith's concerns, though I think ultimately we *can* make sense of Kyle's probability estimate, and that Kyle *can* have at least some grounding for it. I'll now try to explain why I think that, partly to solicit feedback on whether this thinking (and my explanation of it) makes sense.

In the non-supernatural cases mentioned earlier, it seemed clear to me that we had relevant data and theories. We have data on previous viruses and military interventions (albeit likely from different contexts and circumstances), and some relevant theoretical understandings (e.g., from biology and epidemiology, in the virus case). We lack data on a previous completed instance of AGI development, but we have data on cases we could argue are *somewhat* analogous (e.g., industrial revolution, development and roll-out of electricity, development of the atomic bomb, development of the internet), and we have theoretical understandings that can guide us in our reference class forecasting.

But do we have *any* relevant data or theories for the supernatural-type cases?

Assuming theoretical observability

Let's first make the assumption (which I'll reverse later) that these propositions, if true, would *at some point* have at least *some*, *theoretically observable* consequences. That is, we'll first assume that we're *not* dealing with an utterly unverifiable, unfalsifiable hypothesis, the truth of which would have no impact on the world anyway (see also Carl Sagan's dragon).[5] This seems to be the assumption Smith is making, as he writes "Kyle didn't previously experience a bunch of realities and learn that some of them had all-powerful gods while others didn't", implying that *it would be theoretically possible* to learn whether a given reality had an all-powerful god.

That assumption still leaves open the possibility that, *even if these propositions were true*, it'd be extremely unlikely we'd observe any evidence of them at all. This clearly makes it harder to assign probabilities to these propositions that are likely to track reality well. But is it *impossible* to assign *any* probabilities, or to make sense of probabilities that we assign?

It seems to me (though I'm unsure) that we could assign probabilities using something like the following process:

1. Try to think of all (or some sample of) the propositions that we know have ever been made that are similar to the proposition in question. This could mean something like one or more of the following:

[5] I acknowledge that this may mean that these claims aren't "actually supernatural", but they still seem like more-challenging-than-usual cases for the idea that we can always assign meaningful probabilities.

    a. All claims of a religious nature.
    b. All claims that many people would consider “supernatural”.
    c. All claims where no one really had a particular idea of what consequences we should expect to observe if they were true rather than false. (E.g., ghosts, given that they’re often interpreted as being meant to be invisible and incorporeal.)
    d. All claims that are believed to roughly the same level by humanity as a whole or by some subpopulation (e.g., scientists).
2. Try to figure out how many of these propositions later turned out to be true.
    a. This may require debating what counts as still being the same proposition, if the proposition was originally hardly specified. For example, does the ability to keep objects afloat using magnets count as levitation?
3. Do something along the lines of reference class forecasting using this “data”.
    a. This’ll likely require deciding whether certain data points count as a relevant claim turning out to not be true or just not yet turning out to be true. This may look like inside-view-style thinking about roughly how likely we think it’d be that we’d have observed evidence for that claim by now if it *is* true.
    b. We might do something like giving some data points more or less “weight” depending on things like how similar they seem to the matter at hand or how confident we are in our assessment of whether that data point “turned out to be true” or not. (I haven’t thought through in detail precisely how you’d do this; you might instead construct multiple separate reference classes, and then combine these like in model combination, giving different weights to the different classes.)
4. If this reference class forecasting suggests odds of 0%, this seems too confident; it seems that we should never use probabilities of 0 or 1. It seems that one option for handling this would be Laplace’s solution to the rule of succession.
    a. For example, if we found that 18 out of 18 relevant claims for which we “have data” “turned out to be false”, our reference class forecast might suggest there’s a *100%* chance (because 18/18=1) that the claim under consideration will turn out to be false too. To avoid this absolute certainty, we add 1 to the numerator and 2 to the denominator (so we do 19/20=0.95), and find that there’s a 95% chance the claim under consideration will turn out to be false too.
    b. There may be alternative solutions too, such as letting the inside view considerations introduced in the next step move one away from absolute certainty.
5. Construct an inside-view relevant to how likely the claim is to be true. This may involve considerations like:
    a. Knowledge from other fields like physics, and thinking about how consistent this claim is with that knowledge (and perhaps also about how well consistency with knowledge from other fields has predicted truth in the past).
    b. The extent to which the claim violates Occam’s razor, and how bad it is for a claim to do so (perhaps based on how well sticking to Occam’s razor has seemed to predict the accuracy of claims in the past).

    c. Explanations for why the claim would be made and believed as widely as it is even if it *isn't* true. E.g., explanations from the evolutionary psychology of religion, or explanations based on how memetics.
6. Combine the reference class forecast and the inside view somehow. (Perhaps qualitatively, or perhaps via explicit model combination.)

I don't expect that many people *actually, explicitly* use the above process (I personally haven't). But I think it'd be *possible* to do so. And if we want to know "what to make of" probability estimates for these sorts of claims, we could perhaps think of what we actually do, which is more implicit/intuitive, as approximating that explicit process. (But that's a somewhat separate and debatable claim; my core claims are consistent with the idea that *in practice* people are coming to their probability assignments quite randomly.)

Another, probably more realistic way people could arrive at probability estimates for these sorts of claims is through:

1. Do some *very vague*, *very implicit* version of the above.
    a. E.g., just "thinking about" how often things "like this" have seemed true in the past (without actually counting up various cases), and "thinking about" how likely the claim seems to you, when you bear in mind things like physics and Occam's razor.
2. Then introspect on how likely this claim "feels" to you, and try to arrive at a number to represent that.
    a. One method to do so is Hubbard's "equivalent bet test" (described here).

Many people may find that method quite suspicious. But there's evidence that, at least in some domains, it's possible to become fairly "well calibrated", and thus do better than chance at assigning probability estimates, following "calibration training" (see here and here). Ideally, the person using that method would have engaged in such calibration training before. If they have, they might add a third step, or add as part of step 2, an adjustment to account for them tending to over- or underestimate probabilities (or perhaps probabilities of roughly this kind).

I'm not aware of any evidence of whether people can become well-calibrated for these "supernatural-type claims". And I believe there's somewhat limited evidence on how well calibration training generalises across domains. So I think there are major reasons for skepticism, which I'd translate into large confidence intervals on my probability distributions.

But I'm also not aware of any *extremely compelling* arguments or evidence indicating that people *wouldn't* be able to become well-calibrated for these sorts of claims, or that calibration training *wouldn't* generalise to domains like this. So for now, I *think* I'd say that we can make sense of probability estimates for claims like these, and that we should have at least a *very weak* expectation that methods like the above will result in better probability estimates than if we acted as though we knew *nothing at all*.

Assuming no impact on the natural world

I think the much trickier case is if we assume that the truth of these claims would *never* affect the (natural/physical/whatever) world at all, and would thus *never* be observable. I think the standard rationalist response to this possibility is dismissiveness, and the argument that, under those conditions, whether or not these claims are true is an utterly meaningless and unimportant question. The claims are empty, and not worth arguing about.

I find this response *very compelling*, and it's the one I've typically gone with. I think that, if we can show that probabilities can be meaningfully assigned to all claims that could ever theoretically affect the natural world at all, that's probably good enough.

But what if, for the sake of the argument, we entertain the possibility that some claims may never affect the natural world, and yet still be important? Me not dismissing that possibility outright and immediately may annoy some readers, and I can sympathise with that. But it seems to me at least interesting to think about. And here's one case where that possibly actually *does* seem to me like it could be important:

What if non-naturalistic moral realism is "correct", and what that means is that "moral facts" will *never* affect the natural world, and will thus *never* be observable, even in principle - *but* our actions are still somehow relevant to these moral facts. E.g., what if it could be the case that it's "good" for us to do one thing rather than another, in a sense that we "really should" care about, but "goodness" itself leaves no trace at all in the natural world. (This could be something like epiphenomenalism, but here I'm going quite a bit beyond what I really know.)

In this case, I think reference forecasting is useless, because we'd never have any data on the truth or falsehood of any claims of the right type.

But at first glance, it still *seems to me* like we may be able to make some headway using inside views or something like arriving at a "feeling" about the likelihood and then quantifying this using the equivalent bet test. I'm very unsure about that, because usually those methods should rely on at least *some*, *somewhat relevant* data. But it seems like perhaps we can still usefully use considerations like how often Occam's razor has worked well in the past.

And this also reminds me of Scott Alexander's post on building intuitions on non-empirical arguments in science (additional post on that here). It also seems reminiscent of some of Eliezer Yudkowsky's writing on the many-worlds interpretation of quantum mechanics, though I read those posts a little while ago and didn't have this idea in mind at the time.[6]

[6] To be clear, I'm not necessarily claiming that Alexander or Yudkowsky would approve of using this sort of logic for topics like non-naturalistic moral realism or the existence of a god, rather than just dismissing those questions outright as meaningless or utterly inconsequential. I'm just drawing what *seems to me*, *from memory*, some potential connections.

Closing remarks

This quick post has become longer than planned, so I'll stop there. The basic summary is that I tentatively claim we *can* always assign meaningful probabilities, even to supernatural-type (or even actually supernatural) claims. I'm *not* claiming we should be confident in these probabilities, and in fact, I expect many people should massively reduce their confidence in their probability estimates. I'm also *not* claiming that the probabilities people *actually* assign are reliably better than chance - that's an empirical question, and again there'd likely be issues of overconfidence.

As I said at the start, a major aim of this post is to get feedback on my thinking. So please let me know what you think in the comments.