Filtered Evidence, Filtered Arguments

1. Mistakes with Conservation of Expected Evidence
2. Explanation vs Rationalization
3. Timeless Modesty?
4. Co-Proofs
5. Thinking About Filtered Evidence Is (Very!) Hard
6. Subtle Forms of Confirmation Bias
7. Gears Level & Policy Level
8. Placing Yourself as an Instance of a Class
9. The Problematic Third Person Perspective
10. Confusions Concerning Pre-Rationality

Mistakes with Conservation of Expected Evidence

Epistemic Status: I've really spent some time wrestling with this one. I am highly confident in most of what I say. However, this differs from section to section. I'll put more specific epistemic statuses at the end of each section.

Some of this post is generated from mistakes I've seen people make (or, heard people complain about) in applying conservation-of-expected-evidence or related ideas. Other parts of this post are based on mistakes I made myself. I think that I used a wrong version of conservation-of-expected-evidence for some time, and propagated some wrong conclusions fairly deeply; so, this post is partly an attempt to work out the right conclusions for myself, and partly a warning to those who might make the same mistakes.

All of the mistakes I'll argue against have *some good insight behind them.* They may be something which is usually true, or something which points in the direction of a real phenomenon while making an error. I may come off as nitpicking.

1. "You can't predict that you'll update in a particular direction."

Starting with an easy one.

It can be tempting to simplify conservation of expected evidence to say you can't predict the direction which your beliefs will change. This is often approximately true, and it's exactly true in symmetric cases where your starting belief is 50-50 and the evidence is equally likely to point in either direction.

To see why it is wrong in general, consider an extreme case: a universal law, which you mostly already believe to be true. At any time, you could see a counterexample, which would make you jump to complete disbelief. That's a small probability of a very large update downwards. Conservation of expected evidence implies that you must move your belief upwards when you don't see such a counterexample. But, you consider that case to be quite likely. So, considering only which *direction* your beliefs will change, you can be fairly confident that your belief in the universal law will increase -- in fact, as confident as you are in the universal law itself.

The critical point here is direction vs magnitude. Conservation of expected evidence takes magnitude as well as direction into account. The small but very probable increase is balanced by the large but very improbable decrease.

The fact that we're talking about universal laws and counterexamples may fool you into thinking about logical uncertainty. You *can* think about logical uncertainty if you want, but this phenomenon is present in the fully classical Bayesian setting; there's no funny business with non-Bayesian updates here.

Epistemic status: confidence at the level of mathematical reasoning.

2. "Yes requires the possibility of no."

Scott's recent post, yes requires the possibility of no, is fine. I'm referring to a possible mistake which one could make in applying the principle illustrated there.

> “Those who dream do not know they dream, but when you are awake, you know you are awake.” -- Eliezer, Against Modest Epistemology

Sometimes, look around, and ask myself whether I'm in a dream. When this happens, I generally conclude very confidently that I'm awake.

I am not similarly capable of determining that I'm dreaming. My dreaming self doesn't have the self-awareness to question whether he is dreaming in this way.

(Actually, very occasionally, I do. I either end up forcing myself awake, or I become lucid in the dream. Let's ignore that possibility for the purpose of the thought experiment.)

I am not claiming that my dreaming self is never deluded into thinking he is awake. On the contrary, I have those repeatedly-waking-up-only-to-find-I'm-still-dreaming dreams occasionally. In those cases, I vividly believe myself to be awake. So, it's definitely possible for me to vividly believe I'm awake and be mistaken.

What I'm saying is that, when I'm asleep, I am not able to perform *the actually good test*, where I look around and really consciously consider whether or not I might be dreaming. Nonetheless, when I can perform that check, it seems quite reliable. If I want to know if I'm awake, I can just check.

A "yes-requires-the-possibility-of-no" mindset might conclude that my "actually good test" is no good at all, because it can't say "no". I believe the exact opposite: my test seems really quite effective, because I only successfully complete it while awake.

Sometimes, your thought processes really are quite suspect; yet, there's a sanity check you can run which tells you the truth. If you're deluding yourself, the *general category* of "things which you think are simple sanity checks you can run" is not trustworthy. If you're deluding yourself, you're not even going to think about the real sanity checks. But, that *does not in itself detract from the effectiveness of the sanity check.*

The general moral in terms of conservation of expected evidence is: "'*Yes*' *only requires the possibility of silence.*". In many cases, you can meaningfully say yes without being able to meaningfully say no. For example, the axioms of set theory could prove their own inconsistency. They could *not* prove themselves consistent (without also proving themselves inconsistent). This does not detract from the effectiveness of a proof of inconsistency! Again, although the example involves logic, there's nothing funny going on with logical uncertainty; the phenomenon under discussion is understandable in fully Bayesian terms.

Symbolically: as is always the case, you don't really want to update on the raw proposition, but rather, *the fact that you observed the proposition*, to account for selection bias. Conservation of expected evidence can be written

$P(H) = P(H|E)P(E) + P(H|\neg E)P(\neg E)$, but if we re-write it to explicitly show the

"observation of evidence", it becomes

$P(H) = P(H|obs(E))P(obs(E)) + P(H|\neg obs(E))P(\neg obs(E))$. It ***does not become***

$P(H) = P(H|obs(E))P(obs(E)) + P(H|obs(\neg E))P(obs(\neg E))$. In English: evidence is balanced between making the observation and not making the observation, ***not*** between the observation and the observation of the negation.

Epistemic status: confidence at the level of mathematical reasoning for the core claim of this section. However, some applications of the idea (such as to dreams, my central example) depend on trickier philosophical issues discussed in the next section. I'm only moderately confident I have the right view there.

3. "But then what do you say to the Republican?"

I suspect that many readers are *less than fully on board* with the claims I made in the previous section. Perhaps you think I'm grossly overconfident about being awake. Perhaps you think I'm neglecting the outside view, or ignoring something to do with timeless decision theory.

A lot of my thinking in this post was generated by grappling with some points made in Inadequate Equilibria. To quote the relevant paragraph of against modest epistemology:

> Or as someone advocating what I took to be modesty recently said to me, after I explained why I thought it was sometimes okay to give yourself the discretion to disagree with mainstream expertise when the mainstream seems to be screwing up, in exactly the following words: “But then what do you say to the Republican?”

Let's put that in (pseudo-)conservation-of-expected-evidence terms: we know that just applying one's best reasoning will often leave one overconfident in one's idiosyncratic beliefs. Doesn't that mean "apply your best reasoning" is a bad test, which fails to conserve expected evidence? So, should we not adjust downward in general?

In the essay, Eliezer strongly advises allowing yourself to have an inside view even when there's an outside view which says inside views broadly similar to yours tend to be mistaken. But doesn't that go against what he said in Ethical Injunctions?

Ethical Injunctions argues that there are situations where you should not trust your reasoning, and fall back on a general rule. You do this because, in the vast majority of cases of that kind, your oh-so-clever reasoning is mistaken and the general rule saves you from the error.

In *Against Modest Epistemology*, Eliezer criticizes arguments which rely on putting arguments in very general categories and taking the outside view:

> At its epistemological core, modesty says that we should abstract up to a particular *very general* self-observation, condition on it, and then not condition on anything else because that would be inside-viewing. An observation like, “I’m familiar with the cognitive science literature discussing which debiasing techniques work well in practice, I’ve spent time on calibration and visualization

> exercises to address biases like base rate neglect, and my experience suggests that they've helped," is to be generalized up to, "I use an epistemology which I think is good." I am then to ask myself what average performance I would expect from an agent, conditioning only on the fact that the agent is using an epistemology that they think is good, and not conditioning on that agent using Bayesian epistemology or debiasing techniques or experimental protocol or mathematical reasoning or anything in particular.
>
> Only in this way can we force Republicans to agree with us... or something.

He instead advises that we should update on all the information we have, use our best arguments, reason about situations in full detail:

> If you're trying to estimate the accuracy of your epistemology, and you know what Bayes's Rule is, then—on naive, straightforward, traditional Bayesian epistemology—you ought to condition on both of these facts, and estimate *P*(accuracy|know_Bayes) instead of *P*(accuracy). Doing anything other than that opens the door to a host of paradoxes.

In *Ethical Injunctions,* he seems to warn against that very thing:

> But surely... if one is *aware of these reasons...* then one can simply redo the calculation, taking them into account. So we can rob banks if it seems like the right thing to do *after taking into account* the problem of corrupted hardware and black swan blowups. That's the rational course, right?
>
> There's a number of replies I could give to that.
>
> I'll start by saying that this is a prime example of the sort of thinking I have in mind, when I warn aspiring rationalists to beware of cleverness.

Now, maybe Eliezer has simply changed views on this over the years. Even so, that leaves *us* with the problem of how to reconcile these arguments.

I'd say the following: modest epistemology points out a *simple improvement* over the default strategy: "In any group of people who disagree, they can do better by moving their beliefs toward each other." "Lots of crazy people think they've discovered secrets of the universe, and the number of sane people who truly discover such secrets is quite small; so, we can improve the average by never believing we've discovered secrets of the universe." If we take a timeless decision theory perspective (or similar), this *is in fact an improvement*; however, it is *far from the optimal policy*, and has a form which blocks further progress.

Ethical Injunctions talks about rules with greater specificity, and less progress-blocking nature. Essentially, a proper ethical injunction is *actually the best policy you can come up with*, whereas the modesty argument stops short of that.

Doesn't the "actually best policy you can come up with" risk overly-clever policies which depend on broken parts of your cognition? Yes, but your meta-level arguments about which kinds of argument work should be independent sources of evidence from your object-level confusion. To give a toy example: let's say you really, really want 8+8 to be 12 due to some motivated cognition. You can still decide to check by applying basic arithmetic. You might *not* do this, because you *know* it isn't to the advantage of the motivated cognition. However, if you do check, it is actually quite difficult for the motivated cognition to warp basic arithmetic.

There's also the fact that choosing a modesty policy ***doesn't really help the republican***. I think that's the critical kink in the conservation-of-expected-evidence version of modest epistemology. If you, while awake, decide to doubt whether you're awake (no matter how compelling the evidence that you're awake seems to be), then *you're not really improving your overall correctness*.

So, all told, it seems like conservation of expected evidence has to be applied to the *details* of your reasoning. If your put your reasoning in a more generic category, it may appear that a much more modest conclusion is required by conservation of expected evidence. We can justify this in classical probability theory, though in this section it is even more tempting to consider exotic decision-theoretic and non-omnescience considerations than it was previously.

Epistemic status: the conclusion is mathematically true in classical Bayesian epistemology. I am subjectively >80% confident that the conclusion should hold in >90% of realistic cases, but it is unclear how to make this into a real empirical claim. I'm unsure enough of how ethical injunctions should work that I could see my views shifting significantly. I'll mention pre-rationality as one confusion I have which seems vaguely related.

4. "I can't credibly claim anything if there are incentives on my words."

Another rule which one might derive from Scott's *Yes Requires the Possibility of No* is: you can't really say anything if pressure is being put on you to say a particular thing.

Now, I agree that this is somewhat true, particularly in simple cases where pressure is being put on you to say one particular thing. However, I've suffered from learned helplessness around this. I sort of shut down when I can identify any incentives at all which could make my claims suspect, and hesitate to claim anything. This isn't a very useful strategy. Either "just say the truth", or "just say whatever you feel you're expected to say" are *both* likely better strategies.

One idea is to "call out" the pressure you feel. "I'm having trouble saying anything because I'm worried what you will think of me." This isn't always a good idea, but it can often work fairly well. Someone who *is* caving to incentives isn't very likely to say something like that, so it provides some evidence that you're being genuine. It can also open the door to other ways you and the person you're talking to can solve the incentive problem.

You can also "call out" something even if you're unable or unwilling to explain. You just say something like "there's some *thing* going on"... or "I'm somehow frustrated with this situation"... or whatever you can manage to say.

This "call out" idea also works (to some extent) on motivated cognition. Maybe you're worried about the social pressure on your beliefs because it might influence the accuracy of those beliefs. Rather than stressing about this and going into a spiral of self-analysis, you can just state to yourself that that's a thing which might be going on, and move forward. Making it explicit might open up helpful lines of thinking later.

Another thing I want to point out is that most people are willing to place at least a little faith in your honesty (and not irrationally so). Just because you have a story in

mind where they should assume you're lying doesn't mean that's the only possibility they are -- or should be -- considering. One problematic incentive doesn't fully determine the situation. (This one also applies internally: identifying one relevant bias or whatever doesn't mean you should block off that part of yourself.)

Epistemic status: low confidence. I imagine I would have said something very different if I were more an expert in this particular thing.

5. "Your true reason screens off any other evidence your argument might include."

In The Bottom Line, Eliezer describes a clever arguer who first writes the conclusion which they want to argue for at the bottom of a sheet of paper, and then comes up with as many arguments as they can to put above that. In the thought experiment, the clever arguer's conclusion is actually determined by who can pay the clever arguer more. Eliezer says:

> So the handwriting of the curious inquirer is entangled with the signs and portents and the contents of the boxes, whereas the handwriting of the clever arguer is evidence only of which owner paid the higher bid. There is a great difference in the indications of ink, though one who foolishly read aloud the ink-shapes might think the English words sounded similar.

Now, Eliezer is trying to make a point about *how you form your own beliefs* -- that the quality of the process which determines which claims *you* make is what matters, and the quality of any rationalizations you give doesn't change that.

However, reading that, I came away with the mistaken idea that *someone listening to a clever arguer should ignore all the clever arguments*. Or, generalizing further, *what you should do when listening to any argument is try to figure out what process wrote the bottom line*, ignoring any other evidence provided.

This isn't the worst possible algorithm. You really *should* heavily discount evidence provided by clever arguers, because it has been heavily cherry-picked. And almost everyone does a great deal of clever arguing. Even a hardboiled rationalist will tend to present evidence for the point they're trying to make rather than against (perhaps because that's a fairly good strategy for explaining things -- sampling evidence at random isn't a very efficient way of conversing!).

However, ignoring arguments and attending only to the original causes of belief has some absurd consequences. Chief among them is: it would imply that you should ignore mathematical proofs if the person who came up with the proof only searched for positive proofs and wouldn't have spend time trying to prove the opposite. (This ties in with the very first section -- failing to find a proof is like remaining silent.)

This is bonkers. Proof is proof. And again, this isn't some special non-Bayesian phenomenon due to logical uncertainty. A Bayesian can and should recognize decisive evidence, whether or not it came from a clever arguer.

Yet, I really held this position for a while. I treated mathematical proofs as an exceptional case, rather than as a phenomenon continuous with weaker forms of evidence. If a clever arguer presented anything *short* of a mathematical proof, I would remind myself of how convincing cherry-picked evidence can seem. And I'd notice how almost everyone mostly cherry-picked when explaining their views.

This strategy was throwing out data when it has been contaminated by selection bias, rather than making a model of the selection bias so that I could update on the data appropriately. It might be a good practice in scientific publications, but if you take it as a universal, you could find reasons to throw out just about *everything* (especially if you start worrying about anthropic selection effects).

The right thing to do is closer to this: figure out how convincing you expect evidence to look *given* the extent of selection bias. Then, update on the *difference* between what you see and what's expected. If a clever arguer makes a case which is much better than what you would have expected they could make, you can update up. If it is *worse* that you'd expect, even if the evidence would otherwise look favorable, you update *down*.

My view also made me uncomfortable presenting a case for my own beliefs, because I would think of myself as a clever-arguer any time I did something other than recount the actual historical causes of my belief (or honestly reconsider my belief on the spot). Grognor made a similar point in Unwilling or Unable to Explain:

> Let me back up. Speaking in good faith entails giving the real reasons you believe something rather than a persuasive impromptu rationalization. Most people routinely do the latter without even noticing. I'm sure I still do it without noticing. But when I do notice I'm about to make something up, instead I clam up and say, "I can't explain the reasons for this claim." I'm not willing to disingenuously reference a scientific paper that I'd never even heard of when I formed the belief it'd be justifying, for example. In this case silence is the only feasible alternative to speaking in bad faith.

While I think there's something to this mindset, I no longer think it makes sense to clam up when you can't figure out how you originally came around to the view which you now hold. If you think there are other good reasons, you can give them without violating good faith.

Actually, I really wish I could draw a sharper line here. I'm essentially claiming that a little cherry-picking is OK if you're just trying to convince someone of the view which you see as the truth, so long as you're not intentionally hiding anything. This is an uncomfortable conclusion.

Epistemic status: confident that the views I claim are mistaken are mistaken. Less confident about best-practice claims.

6. "If you can't provide me with a reason, I have to assume you're wrong."

If you take the conclusion of the previous section too far, you might reason as follows: if someone is trying to claim X, surely they're trying to give you some evidence toward X. If they claim X and then you challenge them for evidence, they'll try to tell you any evidence they have. So, if they come up with *nothing*, you have to update *down,* since you would have updated upwards otherwise. Right?

I think most people make this mistake due to simple conversation norms: when navigating a conversation, people have to figure out what everyone else is willing to assume, in order to make sensible statements with minimal friction. So, we look for obvious signs of whether a statement was accepted by everyone vs rejected. If someone was asked to provide a reason for a statement they made and failed to do so, that's a fairly good signal that the statement hasn't been accepted into the common background assumptions for the conversation. The fact that other people are likely to use this heuristic as well makes the signal even stronger. So, assertions which can't be backed up with reasons are likely to be rejected.

This is almost the opposite mistake from the previous section; the previous one was *justifications don't matter*, whereas this idea is *only justifications matter.*

I think something good happens when everyone in a conversation recognizes that *people can believe things for good reason without being able to articulate those reasons*. (This includes yourself!)

You can't just give everyone a pass to make unjustified claims and assert that they have strong inarticulable reasons. Or rather, you *can* give everyone a pass to do that, but you don't have to take them seriously when they do it. However, in environments of high intellectual trust, you *can* take it seriously. Indeed, applying the usual heuristic will likely cause you to update in the wrong direction.

Epistemic status: moderately confident.

Conclusion

I think all of this is fairly important -- if you're like me, you've likely made some mistakes along these lines. I also think there are many issues related to conservation of expected evidence which I still don't fully understand, such as explanation vs rationalization, ethical injunctions and pre-rationality. Tsuyoku Naritai!

Explanation vs Rationalization

Follow-up to: Toward a New Technical Explanation of Technical Explanation, The Bottom Line.

In The Bottom Line, Eliezer argues that arguments should only provide evidence to the extent that their conclusions were determined in a way which correlated them with reality. If you write down your conclusion at the bottom of the page, and then construct your argument, your argument does nothing to make the conclusion more entangled with reality.

This isn't precisely true. If you know that someone tried really hard to put together all the evidence for their side, and you still find the argument underwhelming you should probably update *against* what they're arguing. Similarly, if a motivated arguer finds a surprisingly compelling argument with much less effort than you expected, this should update you toward what they claim. So, you can still get evidence from the arguments of motivated reasoners, if you adjust for base rates of the argument quality you expected from them.

Still, motivated reasoning is bad for discourse, and aspiring rationalists seek to minimize it.

Yet, I think everyone has had the experience of trying to explain something and looking for arguments which will help the other person to get it. This is different than trying to convince / win an argument, right? I have been uneasy about this for a long time. Trying to find a good explanation is a lot like motivated cognition. Yet, trying to explain something to someone doesn't seem like it is wrong in the same way, does it?

A possible view which occurred to me is that you should only give the line of reasoning which originally convinced you. That way, you're sure you aren't selecting evidence; the evidence is selecting what you argue.

I think this captures some of the right attitude, but is certainly too strict. Teachers couldn't use this rule, since it is prudent to select good explanations rather than whichever explanation you heard first. I think the rule would also be bad for math research: looking for a proof is, mostly, a better use of your time than trying to articulate the mathematical intuitions which lead to a conjecture.

A second attempt to resolve the conflict: you must adopt different conversational modes for efficiently conveying information vs collaboratively exploring the truth. It's fine to make motivated arguments when you're trying to explain things well, but you should avoid them like the plague if you're trying to find out what's true in the first place.

I also think this isn't quite right, partly because I think good teaching is more like collaborative truth exploration, and partly because of the math research example I already mentioned.

I think this is what's going on: ***you're OK if you're looking for a gears-level explanation.*** Since gears-level explanations are more objective, it is harder to bend them with motivated cognition. They're also a handier form of knowledge to pass around from person to person, since they tend to be small and easily understood.

In the case of a mathematician who has a conjecture, a proof is a rigorous explanation which is quite unlikely to be wrong. You can think of looking for a proof as a way of *checking* the conjecture, sure; in that respect it might not seem like motivated cognition at all. However, that's if you doubt your conjecture and are looking for the proof as a test. I think there's *also* a case where you *don't* doubt your conjecture, and are looking for a proof to convince others. You might still change your mind if you can't find one, but the point is *you weren't wrong to search for a proof with the motive to convince* -- because of the rigorous nature of proofs, there is no selection-of-evidence problem.

If you are a physicist, and I ask what would happen if I do a certain thing with gyroscopes, you might give a quick answer without needing to think much. If I'm not convinced, you might proceed to try and convince me by explaining which physical principles are in play. You're doing something which looks like motivated cognition, but it isn't much of a problem because it isn't so easy to argue wrong conclusions from physical principles (if both of us are engaging with the arguments at a gears level). If I ask you to tell we what reasoning actually produced your quick answer rather than coming up with arguments, you might have nothing better to say than "intuition from long experience playing with gyroscopes and thinking about the physics".

If you are an expert of interior design, and tell me where I should put my couch, I might *believe* you, but still ask for an argument. Your initial statement may have been intuitive, but it isn't *wrong* for you to try and come up with more explicit reasons. *Maybe* you'll just come up with motivated arguments -- and you should watch out for that -- but maybe you'll articulate a model, not too far from your implicit reasoning, in which the couch just obviously does belong in that spot.

There's a lot of difference between math, physics, and interior design in terms of the amount of wiggle room gears-level arguments might have. There's almost no room for motivated arguments in formal proofs. There's lots of room in interior design. Physics is somewhere in between. I don't know how to cleanly distinguish in practice, so that we can have a nice social norm against motivated cognition while allowing explanations. (People seem to mostly manage on their own; I don't actually see so many people shutting down attempted explanations by labeling them motivated cognition.) Perhaps being aware of the distinction is enough.

The distinction is also helpful for explaining why you might want more information when you already believe someone. It's easy for me to speak from my gears level model and sound like I don't believe you yet, when really I'm just asking for an explanation. "Agents should maximize expected utility!" you say. "Convince me!" I say. "VNM Theorem!" you say. "What's the proof?" I say. You can't necessarily tell if I'm being skeptical or curious. We can convey more nuanced epistemics by saying things like "I trust you on things like this, but I don't have your models" or "OK, can you explain why?"

Probabilistic evidence provides nudges in one direction or another (sometimes strong, sometimes weak). These can be filtered by a clever arguer, collecting nudges in one direction and discarding the rest, to justify what they want you to believe. However, if this kind of probabilistic reasoning is like floating in a raft on the sea, a gears-level *explanation* is like finding firm land to stand on. Mathematics is bedrock; physics is firm soil; other subjects may be like shifting sand (it's all fake frameworks to greater/lesser extent) -- but it's more steady than water!

Timeless Modesty?

Looking back at my description of policy-level thinking, one might charge it of the same crime as outside view; namely, being overly likely to lead to modest epistemology:

> Policy-level thinking, on the other hand, helps you to not get lost in the details. It provides the rudder which can keep you moving in the right direction. It's better at cooperating with others, maintaining sanity before you figure out how it all adds up to normality, and optimizing your daily life.

I aim to clarify that "cooperating with others" and "maintaining sanity" does not mean modesty, here.

Both Eliezer's initial argument against the modesty argument and his recent detailed explication, and reductio ad absurdum, of that argument call to mind timeless decision theory. In his early post, Eliezer says:

> The central argument for Modesty proposes something like a Rawlsian veil of ignorance - how can you know which of you is the honest truthseeker, and which the stubborn self-deceiver?

Does reasoning from behind a veil of ignorance really support the modesty argument? Would a TDT agent use modest epistemology? I think not.

A TDT agent is supposed to think of itself as having logical control over anything implementing a relevantly similar decision procedure, for any particular decision being considered. We can think of such an agent reasoning about whether te re-write the value of a particular belief. From a position of ignorance, wouldn't it be better in expectation to average your belief with that of other people? Quoting the old post again:

> Should George - the customer - have started doubting his arithmetic, because five levels of Verizon customer support, some of whom cited multiple years of experience, told him he was wrong? Should he have adjusted his probability estimate in their direction? [...] Jensen's inequality proves even more straightforwardly that, if George and the five levels of tech support had averaged together their probability estimates, they would have improved their average log score.

It's *true*: committing to average beliefs whenever there's a disagreement makes you better off in expectation, if you're in a situation where you could just as well end up on either side of the error. But, there are two things stopping a TDT agent from reasoning in this way.

1. It often won't make sense to say you could just as easily be the one in the wrong.
2. Even if it does make sense, although averaging beliefs beats doing nothing, there are *much* better epistemological strategies available.

A TDT agent doesn't think "What if I'm not a TDT agent?" -- it only sees itself as having logical control over sufficiently similar decision procedres. Eliezer said:

> Those who dream do not know they dream, but when you are awake, you know you are awake.

The dreamer may lack the awareness needed to know whether it's a dream. The one who is awake can see the difference. The fact of seeing the difference is, in itself, enough to break the symmetry between the waking one and the dreamer. The TDT machinery isn't facing a relevantly similar decision in the two cases; it has no logical control over the dreamer.

So, the TDT argument for averaging could only make sense if the other person would consider averaging beliefs with you for the same reason. Even though it is an argument for *personal* gain (in a TDT sense), the potential gain just isn't there if the other person isn't executing the same decision procedure. If Alice and Bob disagree, and Alice moves her belief halfway between hers and Bob's out of modesty, but Bob doesn't move his, TDT can't be justifying Alice; Bob isn't running a sufficiently similar algorithm. If there's an argument for modest epistemology in that case, it has to be different from the TDT argument. Alice just considers Bob as someone else who has different beliefs, perhaps rational or perhaps not. But then, it seems like Alice should just be doing a regular Bayesian update on Bob's beliefs, with no extra averaging step.

As for the second point, even if you *are* both TDT agents, it *still* makes more sense to deal with each other in a more Bayesian way.

Imagine you could put a microchip in everyone's brain which could influence what they believed. You don't want to install specific facts -- that's crossing a line. What kind of program would you install, to improve everyone's beliefs as much as possible? You're essentially in the place of the TDT agent reasoning from behind the veil of ignorance. The chip really *will* end up in everyone. However, you can't just remove people's biases -- you're stuck with the wetware as-is. All you can do is add some programmed responses to tip things in the right direction. What's the best policy?

You could install modest epistemology. Objection #1 is gone; you know everyone is following the same policy. Do you nudge people to average their beliefs with each other whenever they disagree?

Ah, but remember that people's biases are intact. Wouldn't you be better off estimating some kind of epistemic trust in the other person, first, before averaging? People with very poor native calibration should just copy the beliefs of those with good calibration, and those with good calibration should mostly not average with others. This will increase the expected score much more.

But in order to do that, you need to set up good calibration estimation. Before you know it, you're setting up a whole system of Bayesian mechanics. Given the limitations of the situation, it's not clear whether the optimal policy would result in agreement (be it Aumann-style Bayesian agreement or modest belief-averaging style).

So, the modest epistemologist isn't being creative enough in their policy. They recognize a legitimate improvement over the do-nothing strategy, but not a very good policy on the whole.

Co-Proofs

At the recommendation of Jacobian, I've been reading *Too Like the Lightening*. It is a thoughtful book which has several points of interest to rationalists (imho), but there is one concept which I think is nice enough to pluck out and discuss in itself, rather than being satisfied to suggest that people read the book. I also want to suggest a different name than the one from the book.

If you think discussion of a logical concept which is mentioned in a book is a spoiler, maybe stop here.

At one point, there is a discussion in which one character is explaining how much some other characters must already know. The term "anti-proof" is used to refer to failure to falsify a hypothesis. Having a short term for this concept seems like a really good idea. We have the phrase "absence of evidence is evidence of absence", but we don't have a word for the positive case, where absence of counter-evidence speaks in favor of a hypothesis.

Unfortunately, "anti-proof" sounds more like the former than the latter, even though it is being used for the latter in the book. A more appropriate term would be "co-proof", since it is the absence of a proof of the negation.

For example, an alibi would refute someone's involvement in a crime. The absence of an alibi, then, is a co-proof of their involvement: it does not prove involvement by any means, but it *must* constitute some supporting evidence, by conservation of expected evidence.

By "proof of H" I mean an observation which would make the probability of H very close to 1. (How close is "very close" depends on standards of proof in a context, with mathematics demanding the highest standards.) By "refutation" I mean a proof of the negation. So, a co-proof is an observation whose negation would have taken the probability of H to very near zero:

E is a co-proof of H := $P(H|\neg E) \approx 0$

Why are co-proofs of interest? Popperian epistemology is the claim that scientific hypotheses can be supported only by co-proofs; we attempt to refute things, and if something has survived enough refutation attempts, it is considered to be a strong hypothesis. Bayesians are not Popperians, but Popper was still mostly right about this; so, having a short name for it seems useful.

Thinking About Filtered Evidence Is (Very!) Hard

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

The content of this post would not exist if not for conversations with Zack Davis, and owes something to conversations with Sam Eisenstat.

There's been some talk about filtered evidence recently. I want to make a mathematical observation which causes some trouble for the Bayesian treatment of filtered evidence. [OK, when I *started* writing this post, it was "recently". It's been on the back burner for a while.]

This is also a continuation of the line of research about trolling mathematicians, and hence, relevant to logical uncertainty.

I'm going to be making a mathematical argument, but, I'm going to keep things rather informal. I think this increases the clarity of the argument for *most* readers. I'll make some comments on proper formalization at the end.

Alright, here's my argument.

According to the Bayesian treatment of filtered evidence, you need to update on *the fact that the fact was presented to you*, rather than the raw fact. This involves reasoning about the algorithm which decided which facts to show you. The point I want to make is that this can be incredibly computationally difficult, *even if the algorithm is so simple that you can predict what it will say next.* IE, I don't need to rely on anything like "humans are too complex for humans to really treat as well-specified evidence-filtering algorithms".

For my result, we imagine that a Bayesian reasoner (the "listener") is listening to a series of statements made by another agent (the "speaker").

First, I need to establish some terminology:

Assumption 1. *A listener will be said to have a* **rich hypothesis space** *if the listener assigns some probability to the speaker enumerating any computably enumerable set of statements.*

The intuition behind this assumption is supposed to be: due to computational limitations, the listener may need to restrict to some set H of easily computed hypotheses; for example, the hypotheses might be poly-time or even log-poly. This prevents hypotheses such as "the speaker is giving us the bits of a halting oracle in order", as well as "the speaker has a little more processing power than the listener". However, the hypothesis space is not *so* restricted as to limit the world to being a finite-state machine. The listener can imagine the speaker proving complicated theorems, so long as it is done *sufficiently slowly for the listener to keep up*. In such a model, the listener might imagine the speaker staying quiet for quite a long time

(observing the null string over and over, or some simple sentence such as 1=1) while a long computation completes; and only then making a complicated claim.

This is also not to say that I *assume* my listener considers *only* hypotheses in which it can 100% keep up with the speaker's reasoning. The listener can also have probabilistic hypotheses which recognize its inability to perfectly anticipate the speaker. I'm only pointing out that my result *does not rely on* a speaker which the listener can't keep up with.

What it *does* rely on is that there are not too many restrictions on what the speaker *eventually* says.

Assumption 2. *A listener* **believes a speaker to be honest** *if the listener distinguishes between "X" and "the speaker claims X at time t" (aka "*claims_t*-X"), and also has beliefs such that* $P(X|\ \text{claims}_t\text{-}X)=1$ *when* $P(\text{claims}_t\text{-}X) > 0$*.*

This assumption is, basically, saying that the agent trusts its observations; the speaker can filter evidence, but the speaker cannot falsify evidence.

Maybe this assumption seems quite strong. I'll talk about relaxing it after I sketch the central result.

Assumption 3. *A listener is said to have* **minimally consistent beliefs** *if each proposition* X *has a negation* X*, and* $P(X)+P(X^*)\leq 1$.

The idea behind minimally consistent beliefs is that the listener need not be logically omniscient, but does avoid outright contradictions. This is important, since assuming logical omniscience would throw out computability from the start, making any computational-difficulty result rather boring; but totally throwing out logic would make my result impossible. Minimal consistency keeps an extremely small amount of logic, but, it is enough to prove my result.

Theorem(/Conjecture). *It is not possible for a Bayesian reasoner, observing a sequence of remarks made by a speaker, to simultaneously:*

- *Have a rich hypothesis space.*
- *Believe the speaker to be honest.*
- *Have minimally consistent beliefs.*
- *Have computable beliefs.*

Proof sketch. Suppose assumptions 1-3. Thanks to the rich hypothesis space assumption, the listener will assign some probability to the speaker enumerating theorems of PA (Peano Arithmetic). Since this hypothesis makes distinct predictions, it is possible for the confidence to rise above 50% after finitely many observations. At that point, since the listener expects each theorem of PA to *eventually* be listed, with probability > 50%, and the listener believes the speaker, the listener *must* assign > 50% probability to each theorem of PA! But this implies that the listener's beliefs are not computable, since if we had access to them we could *separate theorems of PA from contradictions* by checking whether a sentence's probability is > 50%. □

So goes my argument.

What does the argument basically establish?

The argument is supposed to be surprising, because *minimally consistent beliefs* are compatible with computable beliefs; and *rich hypothesis space* is compatible with beliefs which are computable on observations alone; yet, when combined with a belief that the speaker is honest, we get an incomputability result.

My take-away from this result is that we cannot *simultaneously* use our unrestricted ability to predict sensory observations accurately *and* have completely coherent beliefs about the world which produces those sensory observations, at least if our "bridge" between the sensory observations and the world includes something like language (whereby sensory observations contain complex "claims" about the world).

This is because using the *full force* of our ability to predict sensory experiences includes some hypotheses which *eventually* make surprising claims about the world, by incrementally computing increasingly complicated information (like a theorem prover which slowly but inevitably produces all theorems of PA). In other words, a rich sensory model contains *implicit information about the world* which we cannot immediately compute the consequences of (in terms of probabilities about the hidden variables out there in the world). This "implicit" information can be *necessarily* implicit, in the same way that PA is *necessarily* incomplete.

To give a non-logical example: suppose that your moment-to-moment anticipations of your relationship with a friend are pretty accurate. It might be that if you roll those anticipations forward, you inevitably become closer and closer until the friendship becomes a romance. *However,* you can't necessarily predict that right now; even though the anticipation of each next moment is relatively easy, you face a halting-problem-like difficulty if you try to anticipate what the *eventual* behavior of your relationship is. Because our ability to look ahead is bounded, each new consequence can be predictable without the overall outcome being predictable.

Thus, in order for an agent to use the full force of its computational power on predicting sensory observations, it must have *partial* hypotheses -- similar to the way logical induction contains traders which focus only on special classes of sentences, or Vanessa's incomplete Bayesianism contains incomplete hypotheses which do not try to predict everything.

So, this is an argument against strict Bayesianism. In particular, it is an argument against strict Bayesianism *as a model of updating on filtered evidence!* I'll say more about this, but first, let's talk about possible holes in my argument.

Here are some concerns you might have with the argument.

One might possibly object that the perfect honesty requirement is unrealistic, and therefore conclude that the result does not apply to realistic agents.

- I would point out that the assumption is not so important, so long as the listener *can conceive of the possibility of perfect honesty,* and assigns it nonzero probability. In that case, we can consider P(X|honesty) rather than P(X). Establishing that some conditional beliefs are not computable seems similarly damning.
- Furthermore, because the "speaker" is serving the role of our *observations*, the perfect honesty assumption is just a version of P(X|observe-X)=1. IE, *observing*

X gives us X. This is true in typical filtered-evidence setups; IE, filtered evidence can be misleading, but it can't be false.
- However, one might further object that *agents need not be able to conceive of* "*perfect honesty*", because this assumption has an unrealistically aphysical, "perfectly logical" character. One might say that *all* observations are imperfect; none are perfect evidence of what is observed. In doing so, we can get around my result. This has some similarity to the assertion that zero is not a valid probability. I don't find this response particularly appealing, but I also don't have a strong argument against it.

Along similar lines, one might object that the result depends on an example ("the speaker is enumerating theorems") which comes from logic, as opposed to any realistic physical world-model. The example does have a "logical" character -- we're not explicitly reasoning about evidence-filtering algorithms interfacing with an empirical world and selectively telling us some things about it. However, I want to point out that I've assumed extremely little "logic" -- the only thing I use is that you don't expect a sentence and its negation to both be true. Observations corresponding to theorems of PA are just an example used to prove the result. The fact that P(X) can be very hard to compute even when we restrict to easily computed $P(claims_t\text{-}X)$ is very general; even if we do restrict attention to finite-state-machine hypotheses, we are in P-vs-NP territory.

What does this result say about logical uncertainty?

Sam's untrollable prior beat the trollable-mathematician problem by the usual Bayesian trick of explicitly modeling the sequence of observations -- updating on I-observed-X-at-this-time rather than only X. (See also the illustrated explanation.)

However, it did so at a high cost: Sam's prior is *dumb.* It isn't able to perform rich Occam-style induction to divine the hidden rules of the universe. It doesn't *believe in* hidden rules; it believes "if there's a law of nature constraining everything to fit into a pattern, *I will eventually observe that law directly*." It shifts its probabilities when it makes observations, but, in some sense, it doesn't shift them *very much*; and indeed, that property seems key to the computability of that prior.

So, a natural question arises: is this an *essential* property of an untrollable prior? Or can we construct a "rich" prior which entertains hypotheses about the deep structure of the universe, learning about them in an Occam-like way, which is nonetheless still untrollable?

The present result is a first attempt at an answer: given my (admittedly a bit odd) notion of rich hypothesis space, it is indeed impossible to craft a computable prior over logic with some minimal good properties (like believing what's proved to it). I don't directly address a trollability-type property, unfortunately; but I do think I get close to the heart of the difficulty: a "deep" ability to adapt in order to predict data better stands in contradiction with computability of the latent probability-of-a-sentence.

So, how should we think about filtered evidence?

Orthodox Bayesian (OB): We can always resolve the problem by distinguishing between X and "I observe X", and conditioning on **all** the evidence available. Look

how nicely it works out in the Monty Hall problem and other simple examples we can write down.

Skeptical Critique (SC): You're ignoring the argument. You can't handle cases where running your model forward is easier than answering questions about what happens eventually; in those cases, many of your beliefs will either be uncomputable or incoherent.

OB: That's not a problem for me. Bayesian ideals of rationality apply to the logically omniscient case. What they give you is an idealized notion of rationality, which defines the **best** an agent **could** do.

SC: Really? Surely your Bayesian perspective is supposed to have some solid implications for finite beings who are not logically omniscient. I see you giving out all this advice to machine learning programmers, statisticians, doctors, and so on.

OB: Sure. We might not be able to achieve perfect Bayesian rationality, but whenever we see something less Bayesian than it could be, we can correct it. That's how we get closer to the Bayesian ideal!

SC: That sounds like cargo-cult Bayesianism to me. If you spot an inconsistency, **it matters how you correct it**; you don't want to go around correcting for the planning fallacy by trying to do everything faster, right? Similarly, if your rule-of-thumb for the frequency of primes is a little off, you don't want to add composite numbers to your list of primes to fudge the numbers.

OB: No one would make those mistakes.

SC: That's because there are, in fact, rationality principles which apply. You **don't** just cargo-cult Bayesianism by correcting inconsistencies any old way. A boundedly rational agent has rationality constraints which apply, guiding it to better approximate "ideal" rationality. And those rationality constraints don't actually need to refer to the "ideal" rationality. **The rationality constraints are about the update, not in the ideal which the update limits to.**

OB: Maybe we can imagine some sort of finite Bayesian reasoner, who treats logical uncertainty as a black box, and follows the evidence toward unbounded-Bayes-optimality in a bounded-Bayes-optimal way...

SC: Maybe, but I don't know of a good picture which looks like that. The picture we **do** have is given by logical induction: we learn to avoid Dutch books by noticing lots of Dutch books against ourselves, and gradually becoming less exploitable.

OB: That sounds a lot like the picture I gave.

SC: Sure, but it's more precise. And more importantly, it's **not** a Bayesian update -- there is a kind of family resemblance in the math, but it isn't learning through a Bayesian update in a strict sense.

OB: Ok, so what does all this have to do with filtered evidence? I still don't see why the way I handle that is wrong.

SC: Well, isn't the standard Bayesian answer a little suspicious? The numbers conditioning on X don't come out to what you want, so you introduce something new

to condition on, observe-X, which can have different conditional probabilities. Can't you get whatever answer you want, that way?

OB: I don't think so? The numbers are dictated by the scenario. The Monty Hall problem has a right answer, which determines how you should play the game if you want to win. You can't fudge it without changing the game.

SC: Fair enough. But I still feel funny about something. Isn't there an infinite regress? We jump to updating on observe-X when X is filtered. What if observe-X is filtered? Do we jump to observe-observe-X? What if we can construct a "meta Monty-Hall problem" where it isn't sufficient to condition on observe-X?

OB: If you observe, you observe that you observe. And if you observe that you observe, then you must observe. So there's no difference.

SC: If you're logically perfect, sure. But a boundedly rational agent need not realize immediately that it observed X. And certainly it need not realize and update on the entire sequence "X", "I observed X", "I observed that I observed X", and so on.

OB: Ok...

SC: To give a simple example: call a sensory impression "subliminal" when it is weak enough that only X is registered. A stronger impression also registers "observe-X", making the sensory impression more "consciously available". Then, we cannot properly track the effects of filtered evidence for subliminal impressions. Subliminal impressions would always register as if they were unfiltered evidence.

OB: ...no.

SC: What's wrong?

OB: An agent should come with a basic notion of sensory observation. If you're a human, that could be activation in the nerves running to sensory cortex. If you're a machine, it might be RBG pixel values coming from a camera. That's the only thing you ever have to condition on; all your evidence has that form. Observing a rabbit means *getting pixel values corresponding to a rabbit.* We don't start by conditioning on "rabbit" and then patch things by adding "observe-rabbit" as an additional fact. We condition on *the complicated observation corresponding to the rabbit*, which happens to, *by inference,* tell us that there is a rabbit.

SC: That's... a bit frustrating.

OB: How so?

SC: The core Bayesian doctrine is the Kolmogorov axioms, together with the rule that we update beliefs via Bayesian conditioning. A common extension of Bayesian doctrine *grafts on a distinction between observations and hypotheses,* naming some special events as observable, and others as non-observable hypotheses. I want you to *notice when you're using the extension rather than the core.*

OB: How is that even an extension? It just sounds like a special case, which happens to apply to just about any organism.

SC: But you're restricting the rule "update beliefs by Bayesian conditioning" -- you're saying that it only works for *observations*, not for other kinds of events.

OB: Sure, but you could never update on those other kinds of events anyway.

SC: Really, though? Can't you? Some information you update on comes from sensory observations, but other information comes from *reasoning*. Something like a feedforward neural network just computes one big function on sense-data, and can probably be modeled in the way you're suggesting. But something like a memory network has a nontrivial reasoning component. A Bayesian can't handle "updating" on internal calculations it's completed; at best they're treated as if they're black boxes whose outputs are "observations" again.

OB: Ok, I see you're backing me into a corner with logical uncertainty stuff again. I still feel like there should be a Bayesian way to handle it. But what does this have to do with filtered evidence?

SC: *The whole point of the argument we started out discussing* is that if you have this kind of observation/hypothesis divide, and have sufficiently rich ways of predicting sensory experiences, *and remain a classical Bayesian*, then your beliefs about the hidden information are not going to be computable, even if your hypotheses themselves are easy to compute. So we can't realistically reason about the hidden information just by Bayes-conditioning on the observables. The only way to maintain both computability and a rich hypothesis space under these conditions is to be less Bayesian, allowing for more inconsistencies in your beliefs. *Which means,* **reasoning about filtered evidence doesn't reduce to applying Bayes' Law**.

OB: That... seems wrong.

SC: Now we're getting somewhere!

All that being said, reasoning about filtered evidence via Bayes' Law in the orthodox way still seems quite practically compelling. The perspective SC puts forward in the above dialogue would be much more compelling if I had more practical/interesting "failure-cases" for Bayes' Law, and more to say about alternative ways of reasoning which work better for those cases. A real "meta Monty-Hall problem".

Arguably, logical induction *doesn't* use the "condition on the fact that X was observed" solution:

- Rather than the usual sequential prediction model, logical induction accommodates information coming in for any sentence, in any order. So, like the "core of Bayesianism" mentioned by SC, it maintains its good properties without special assumptions about what is being conditioned on. This is in contrast to, e.g., Solomonoff induction, which uses the sequential prediction model.
- In particular, in Monty Hall, although there is a distinction between the sentence "there is a goat behind door 3" and "the LI discovers, at time *t*, that there is a goat behind door 3" (or suitable arithmetizations of these sentences), we can condition on the first rather than the second. A logical inductor would learn to react to this in the appropriate way, since doing otherwise would leave it Dutch-bookable.

One might argue that the traders are implicitly using the standard Bayesian "condition on the fact that X was observed" solution in order to accomplish this. Or that the update an LI performs upon seeing X *is always* that it saw X. But to me, this feels like stretching things. The core of the Bayesian method for handling filtered evidence is to distinguish between X and the observation of X, and update on the latter. A logical

inductor doesn't explicitly follow this, and indeed appears to violate it. Part of the usual idea seems to be that a Bayesian needs to "update on all the evidence" -- but a logical inductor just gets a black-box report of X, without any information on how X was concluded or where it came from. So information can be arbitrarily excluded, and the logical inductor will still do its best (which, in the case of Monty Hall, appears to be sufficient to learn the correct result).

A notable thing about the standard sort of cases, where the Bayesian way of reasoning about filtered evidence is entirely adequate, is that you have a gears-level model of what is going on -- a causal model, which you can turn the crank on. If you run such a model "forward" -- in causal order -- you compute the hidden causes *before* you compute the filtered evidence about them. This makes it sound as if predicting the hidden variables should be *easier* than predicting the sensory observations; and, certainly makes it hard to visualize the situation where it is much much harder.

However, even in cases where we have a nice causal model like that, inferring the hidden variables from what is observed can be intractably computationally difficult, since it requires *reverse-engineering* the computation from its outputs. Forward-sampling causal models is always efficient; running them backwards, not so.

So even with causal models, there can be good reason to engage more directly with logical uncertainty rather than use pure Bayesian methods.

However, I suspect that one could construct a much more convincing example if one were to use partial models explicitly in the construction of the example. Perhaps something involving an "outside view" with strong empirical support, but lacking a known "inside view" (lacking a single consistent causal story).

Unfortunately, such an example escapes me at the moment.

Finally, some notes on further formalisation of my main argument.

The listener is supposed to have probabilistic beliefs of the standard variety -- an event space which is a sigma-algebra, and which has a P(event) obeying the Kolmogorov axioms. In particular, the beliefs are supposed to be perfectly logically consistent in the usual way.

However, in order to allow logical uncertainty, I'm assuming that there is *some embedding of arithmetic;* call it E[]. So, for each arithmetic sentence S, there is an event E[S]. Negation gets mapped to the "star" of an event: $E[\neg S] = (E[S])^*$. This need not be the compliment of the event E[S]. Similarly, the embedding $E[A \vee B]$ need not be $E[A] \cup E[B]$; $E[A \wedge B]$ need not be $E[A] \cap E[B]$; and so on. That's what allows for logical non-omniscience -- the probability distribution doesn't necessarily *know* that $E[A \wedge B]$ should act like $E[A] \cap E[B]$, and so on.

The more we impose requirements which force the embedding to *act like it should,* the more logical structure we are forcing onto the beliefs. If we impose very much consistency, however, then that would already imply uncomputability and the central result would not be interesting. So, the "minimal consistency" assumption requires

very little of our embedding. Still, it is enough for the embedding of PA to cause trouble in connection with the other assumptions.

In addition to all this, we have a distinguished set of events which count as observations. A first pass on this is that for any event A, there is an associated event obs(A) which is the observation of A. But I do worry that this includes more observation events than we want to require. Some events A do not correspond to sentences; sigma-algebras are closed under countable unions. If we think of the observation events as *claims made by the speaker*, it doesn't make sense to imagine the speaker claiming a countable union of sentences (particularly not the union of an uncomputable collection).

So, more conservatively, we might say that for events E[S], that is *events in the image of the embedding*, we also have an event obs(E[S]). In any case, this is closer to the minimal thing we need to establish the result.

I don't know if the argument works out exactly as I sketched; it's possible that the rich hypothesis assumption needs to be "and also positive weight on a particular enumeration". Given that, we can argue: take one such enumeration; as we continue getting observations consistent with that observation, the hypothesis which predicts it loses no weight, and hypotheses which (eventually) predict other things must (eventually) lose weight; so, the updated probability eventually believes that particular enumeration will continue with probability > 1/2.

On the other hand, that patched definition is certainly less nice. Perhaps there is a better route.

Subtle Forms of Confirmation Bias

There are at least two types of confirmation bias.

The first is ***selective attention:*** a tendency to pay attention to, or recall, that which confirms the hypothesis you are thinking about rather than that which speaks against it.

The second is ***selective experimentation:*** a tendency to do experiments which will confirm, rather than falsify, the hypothesis.

The standard advice for both cases seems to be "explicitly look for things which would falsify the hypothesis". I think this advice is helpful, but it is subtly wrong, especially for the selective-experimentation type of confirmation bias. Selective attention is relatively straightforward, but selective experimentation is much more complex than it initially sounds.

Looking for Falsification

What the standard (Popperian) advice tells you to do is try as hard as you can to falsify your hypothesis. You should think up experiments where your beloved hypothesis really could fail.

What this advice definitely does do is guard against the mistake of making experiments which could not falsify your hypothesis. Such a test is either violating conservation of expected evidence (by claiming to provide evidence one way without having any possibility of providing evidence the other way), or providing only very weak evidence for your claim (by looking much the same whether your claim is true or false). Looking for tests which can falsify your result steers you towards tests which would provide strong evidence, and helps you avoid violating the law of expected evidence.

However, there are more subtle ways in which confirmation bias can act.

Predicting Results in Advance

You can propose a test which would indeed fit your hypothesis if it came out one way, and which would disconfirm your hypothesis if it came out the other way -- but where you can predict the outcome in advance. It's easy to not realize you are doing this. You'll appear to provide significant evidence for your hypothesis, but actually you've cherry-picked your evidence before even looking at it; you knew enough about the world to know where to look to see what you wanted to see.

Suppose Dr. Y studies a rare disease, Swernish syndrome. Many scientists have formed an intuition that Swernish syndrome has something to do with a chemical G-complex. Dr. Y is thinking on this one night, when the intuition crystallizes into G-complex theory, which would provide a complete explanation of how Swernish syndrome develops. G-complex theory makes the novel prediction that G-complex in the bloodstream will spike during early onset of the disease; if this were false, G-complex theory would have to be false. Dr. Y does the experiment, and finds that the

spike does occur. No one has measured this before, nor has anyone else put forward a model which makes that prediction. However, it happens that anyone familiar with the details of Dr. Y's experimental results over the past decade would have strongly suspected the same spike to occur, *whether or not* they endorsed G-complex theory. Does the experimental result constitute significant evidence?

This is a subtle kind of double-counting of evidence. You have enough evidence to know the result of the experiment; also, your evidence has caused you to generate a hypothesis. You cannot then claim the success of the experiment as more evidence for your hypothesis: you already know what would happen, so it can't alter the certainty of your hypothesis.

If we're dealing only with personal rationality, we could invoke conservation of expected evidence again: if you already predict the outcome with high probability, you cannot simultaneously derive much evidence from it. However, in group rationality, there are plenty of cases where you want to predict an experiment in advance and then claim it as evidence. *You* may already be convinced, but you need to convince skeptics. So, we can't criticize someone *just* for being able to predict their experimental results in advance. That would be absurd. The problem is, *the hypothesis isn't what did the work of predicting the outcome.* Dr. Y had general world-knowledge which allowed him to select an experiment whose results would be in line with his theory.

To Dr. Y, it just feels like "if I am right, we will see the spike. If I am wrong, we won't see it." From the outside, we might be tempted to say that Dr. Y is not "trying hard enough to falsify G-complex theory". But how can Dr. Y use this advice to avoid the mistake? A hypothesis is an explicit model of the world, which guides your predictions. When asked to *try to falsify*, though, what's your guide? If you find your hypothesis very compelling, you may have difficulty imagining how it could be false. A hypothesis is solid, definite. The negation of a hypothesis includes *anything else*. As a result, "try to falsify your hypothesis" is very vague advice. It doesn't help that the usual practice is to test against a null hypothesis. Dr. Y tests against the spike not being there, and thinks this sufficient.

Implicit Knowledge

Part of the problem here is that it should be very clear what could and could not have been predicted. There's an interaction between your general world knowledge, which is not explicitly articulated, and your scientific knowledge, which is.

If all of your knowledge was explicit scientific knowledge, many biases would disappear. You couldn't possibly have hindsight bias; each hypothesis would predict the observation with a precise probability, which you can calculate.

Similarly, the failure mode I'm describing would become impossible. You could easily notice that it's not really your new hypothesis doing the work of telling you which experimental result to expect; you would know exactly what other world-knowledge you're using to design your experiment.

I think this is part of why it is useful to orient toward gear-like models. If our understanding of a subject is explicit rather than implicit, we can do a lot more to correct our reasoning. However, we'll always have large amounts of implicit, fuzzy

knowledge coming in to our reasoning process; so, we have to be able to deal with that.

Is "Sufficient Novelty" The Answer?

In some sense, the problem is that Dr. Y's experimental result isn't novel enough. It might be a "novel prediction" in the sense that it hasn't been explicitly predicted by anyone, but it is a prediction that *could* have been made without Dr. Y's new hypothesis. Extraordinary claims require extraordinary evidence, right? It isn't enough that a hypothesis makes a prediction which is new. The hypothesis should make a prediction which is *really surprising.*

But, this rule wouldn't be any good for practical science. How surprising something is is too subjective, and it is too easy for hindsight bias to make it feel as if the result of the experiment could have been predicted. Besides: if you want science to be able to provide compelling evidence to skeptics, you can't throw out experiments as unscientific just because most people can predict their outcome.

Method of Multiple Hypotheses

So, how could Dr. Y have avoided the mistake?

It is meaningless to confirm or falsify a hypothesis in isolation; all you can really do is provide evidence which helps distinguish *between* hypotheses. This will guide you away from "mundane" tests where you actually could have predicted the outcome without your hypothesis, because there will likely be many other hypotheses which would be able to predict the outcome of that test. It guides you toward corner cases, where otherwise similar hypotheses make very different predictions.

We can unpack "try to falsify" as "come up with as many plausible alternative hypotheses as you can, and look for experiments which would rule out the others." But actually, "come up with alternative hypotheses" is more than an unpacking of "try to falsify"; it shifts you to trying to distinguish between many hypotheses, rather than focusing on "your" hypothesis as central.

The *actual, exactly correct* criteria for an experiment is its value-of-information. "Try to falsify your hypothesis" is a lousy approximation of this, which judges experiments by how likely they are to provide evidence against your hypothesis, or the likelihood ratio against your hypothesis in the case where the experiment doesn't go as your hypothesis predicts, or something. Don't optimize for the wrong metric; things'll tend to go poorly for you.

Some might object that trying-to-falsify is a good heuristic, since value of information is too difficult to compute. I'd say that a much better heuristic is to pretend distinguishing the right hypothesis is equally valuable in all cases, and look for experiments that allow you to maximally differentiate between them. Come up with as many possibilities as you can, and try to differentiate between the most plausible ones.

Given that the data was already very suggestive of a G-complex spike, Dr. Y would most likely generate other hypotheses which also involve a G-complex spike. This

would make the experiment which tests for the spike uninteresting, and suggest other more illuminating experiments.

I think "coming up with alternatives" is a somewhat underrated debiasing technique. It is discussed more in Heuer's *Psychology of Intelligence Analysis* and Chamberlin's *Method of Multiple Working Hypotheses*.

Gears Level & Policy Level

Inside view vs outside view has been a fairly useful intuition-pump for rationality. However, the dichotomy has a lot of shortcomings. We've just gotten a whole sequence about failures of a cluster of practices called *modest epistemology*, which largely overlaps with what people call outside view. I'm not ready to stop championing what *I* think of as the outside view. However, I *am* ready for a name change. The term *outside view* doesn't exactly have a clear definition; or, to the extent that it does have one, it's "reference class forecasting", which is not what I want to point at. Reference class forecasting has its uses, but many problems have been noted.

I propose *gears level & policy level.* But, before I discuss why these are appropriate replacements, let's look at my motives for finding better terms.

Issues with Inside vs Outside

Problems with the concept of outside view as it currently exists:

- Reference class forecasting tends to imply stopping at base-rate reasoning, rather than *starting* at base-rate reasoning. I want a concept of outside view which helps overcome base-rate neglect, but which more obviously connotes combining an outside view with an inside view (by analogy to combining a prior probability with a likelihood function to get a posterior probability).
- Reference class forecasting lends itself to reference class tennis, IE, a game of choosing the reference class which best makes your point for you. (That's a link to the same article as the previous bullet point, since it originated the term, but this Stuart Armstrong article also discusses it. Paul Christiano discusses rules and ettiquete of reference class tennis, because of course he does.) Reference class tennis is both a pretty bad conversation to have, which makes reference class forecasting a poor choice for productive discussion, and a ponentially big source of bias if you do it to yourself. It's closely related to the worst argument in the world.
- Reference class forecasting is specified at the object level: you find a class fitting the prediction you want to make, and you check the statistics for things in that class to make your prediction. However, central examples of the usefulness of the outside view occur at the meta level. In examples of planning-fallacy correction, you don't just note how close you usually get to the deadline before finishing something. You *compare* it to how close to the deadline you usually *expect* to get. Why would you do that? *To correct your inside view!* As I mentioned before, the *type* of the outside view should be such that it *begs* combination with the inside view, rather than standing on its own.
- Outside view has the connotation of stepping back and ignoring some details. However, we'd like to be able to use all the information at our disposal -- so long as we can use it in the right way. Taking base rates into account can *look* like ignoring information: walking by the proverbial hundred-dollar bill on the gorund in times square, or preparing for a large flood despite there being none in living memory. However, while accounting for base rates does indeed tend to smooth out behavior and make it depend less on evidence, that's because we're working with *more* information, not less. A concept of outside view which connotes bringing in more information, rather than less, would be an improvement.

The existing notion of inside view is also problematic:

- The inside-view vs outside-view distinction does double duty as a descriptive dichotomy and a prescriptive technique. This is especially harmful in the case of inside view, which gets belittled as the naive thing you do before you learn to move to outside view. (We *could* similarly malign the outside view as what you have before you have a true inside-view understanding of a thing.) On the contrary, there are significant skills in forming a high-quality inside view. I primarily want to point at those, rather than the descriptive cluster.

The Gears Level and the Policy Level

Gears-level understanding is a term from CFAR, so you can't blame me for it. Well, I'm endorsing it, so I suppose you can blame me a little. In any case, I like the term, and I think it fits my purposes. Some features of gears-level reasoning:

- Dishing out probability mass precisely, so as to have the virtue of precision.
- Having the properties of a good explanation, along the lines of David Deutsch: being pinned down on all sides by the evidence, and providing understanding, not only predictive accuracy. (Contrast this concept with a big neural-net model which classifies images extremely well but is difficult to analyse.)
- Reasoning from first principles, rather than analogy.
- Making a prediction with a well-defined model, such that anyone who understood your model could calculate the same prediction independently.

The policy level is not a CFAR concept. It is similar to the CFAR concept of *the strategic level*, which I suspect is based on Nate Soares' Staring Into Regrets. In any case, here are some things which point in the right direction:

- Placing yourself as an instance of a class.
- Accounting for knock-on effects, including consistency effects. Choosing an action really is a lot like setting your future policy.
- What game theorists mean by policy: a function from observations to actions, which is (ideally) in equilibrium with the policies of all other agents. A good policy lets you coordinate sucessfully with yourself and with others. Choosing a policy illustrates the idea of choosing at the meta level: you aren't selecting an action, but rather, a function from situations to actions.
- Timeless decision theory / updateless decision theory / functional decision theory. Roughly, choosing a policy from behind a Rawlsian veil of ignorance. As I mentioned with accounting for base rates, it might seem from one perspective like this kind of reasoning is throwing information away; but actually, it is much more powerful. It allows you to set up arbitrary functions *from* information states *to* strategies. You are not actually throwing information away; you always have the option of responding to it as usual. You are *gaining* the option of ignoring it, or reacting to it in a different way, based on larger considerations.
- Cognitive reductions, in Jessica Taylor's sense (points five and six here). Taking the outside view should not entail giving up on having a gears-level model. The virtues of good models at the gears level are still virtues at the policy level. Rather, *the policy level asks you to make a gears-level model of your own cognitive process.* When you go to the policy level, you take your normal way of thinking and doing as an object. You think about the causes and effects of your normal ways of being.

Most of the existing ideas I can point to are about *actions:* game theory, decision theory, the planning fallacy. That's probably the worst problem with the terminology choice. Policy-level thinking has a very instrumental character, because it is about *process*. However, at its core, it is epistemic. Gears level thinking is the *practice* of good map-making. The output is a high-quality map. Policy-level thinking, on the other hand, is the *theory* of map-making. The output is a refined strategy for making maps.

The standard example with the planning fallacy illustrates this: although the goal is to improve planning, which sounds instrumental, the key is noticing the miscalibration of time estimates. The same trick works for any kind of mental miscalibration: if you know about it, you can adjust for it.

This is not just reference class forecasting, though. You don't adjust your time estimates for projects upward and stop there. The *fact* that you normally underestimate how long things will take makes you *think* about your model. "Hm, that's interesting. My plans almost never come out as stated, but I always believe in them when I'm making them." You shouldn't be satisfied with this state of affairs! You *can* slap on a correction factor and keep planning like you always have, but this is a sort of paradoxical mental state to maintain. If you do manage keep the disparity between your past predictions and actual events actively in mind, I think it's more natural to start considering which parts of your plans are most likely to go *wrong.*

If I had to spell it out in steps:

1. Notice that a thing is happening. In particular, notice that a thing is happening *to you,* or that you're *doing* a thing. This step is skipped in experiments on the planning fallacy; experimenters frame the situation. In some respects, though, it's the most important part; naming the situation *as* a situation is what lets you jump outside of it. This is what lets you go off-script, or be anti-sphexish.
2. Make a model of the input-output relations involved. Why did you say what you just said? Why did you think what you just thought? Why did you do what you just did? What are the typical effects of these thoughts, words, actions? This step is most similar to reference class forecasting. Figuring out the input-output relation is a combination of refining the reference class to be the most relevant one, and thinking of the base-rates of outcomes in the reference class.
3. Adjust your policy. Is there a systematic bias in what you're currently doing? Is there a risk you weren't accounting for? Is there an extra variable you could use to differentiate between two cases you were treating as the same? Chesterton-fencing your old strategy is important here. Be gentle with policy changes -- you don't want to make a bucket error or fall into a hufflepuff trap. If you notice resistence in yourself, be sure to leave a line of retreat by visualizing possible worlds. (Yes, I think all those links are actually relevant. No, you don't have to read them to get the point.)

I don't know quite what I can say here to convey the importance of this. There is a *skill* here; a very important skill, which can be done in a split second. It is the skill of going meta.

Gears-Leves and Policy-Level Are Not Opposites

The second-most confusing thing about my proposed terms is probably that they are not opposites of each other. They'd be snappier if they were; "inside view vs outside view" had a nice sound to it. On the other hand, I don't want the concepts to be opposed. I don't want a dichotomy that serves as a discriptive clustering of ways of thinking; I want to point at *skills* of thinking. As I mentioned, the virtuous features of gears-level thinking are still present when thinking at the policy level; unlike in reference class forecasting, the ideal is still to get a good causal model of what's going on (IE, a good causal model of what is producing systematic bias in your way of thinking).

The opposite of gears-level thinking is un-gears-like thinking: reasoning by analogy, loose verbal arguments, rules of thumb. Policy-level thinking will often be like this when you seek to make simple corrections for biases. But, remember, these are error models in the errors-vs-bugs dichotomy; real skill improvement relies on bug models (as studies in deliberate practice suggest).

The opposite of policy-level thinking? Stimulus-response; reinforcement learning; habit; scripted, sphexish behavior. This, too, has its place.

Still, like inside and outside view, gears and policy thinking are made to work together. Learning the principles of strong gears-level thinking helps you fill in the intricate structure of the universe. It allows you to get past social reasoning about who said what and what you were taught and whay you're supposed to think and believe, and instead, get at *what's true.* Policy-level thinking, on the other hand, helps you to not get lost in the details. It provides the rudder which can keep you moving in the right direction. It's better at cooperating with others, maintaining sanity before you figure out how it all adds up to normality, and optimizing your daily life.

Gears and policies both constitute moment-to-moment ways of looking at the world which can change the way you think. There's no simple place to go to learn the skillsets behind each of them, but if you've been around LessWrong long enough, I suspect you know what I'm gesturing at.

Placing Yourself as an Instance of a Class

There's an intuition that I have, which I think informs my opinion on subjective probabilities, game theory, and many related matters: that part of what separates a foolish decision from a wise one is whether you treat it as an isolated instance or as one of a class of similar decisions.

A simple case: someone doing something for the first time (first date, first job interview, etc) vs someone who has done it many times and "knows how these things go". Surprise events to the greenhorn are tired stereotypes for the old hand. But, sometimes, we can short-circuit this process and react wisely to a situation without long experience and hard knocks.

For example, if a person is trying to save money but sees a doodad they'd like to buy, the fool reason as follows: "It's just this one purchase. The amount of money isn't very consequential to my overall budget. I can just save a little more in other ways and I'll meet my target." The wise person reasons as follows: "If I make this purchase now, I will similarly allow myself to make exceptions to my money-saving rule later, until the exception becomes the rule and I spend all my money. So, even though the amount of money here isn't so large, I prefer to follow a general policy of saving, which implies saving in this particular case." A *very* wise person may reason a bit more cleverly: "I can make impulse purchases if they pass a high bar, such that I actually only let a few dollars of unplanned spending past the bar every week on average. How rare is it that a purchase opportunity costing this much is at least this appealing?" ***does a quick check and usually doesn't buy the thing, but sometimes does, when it is worth it***

One way to get this kind of wisdom is by spending money for a long time, and calibrating willingness-to-spend based on what seems to be happening with your bank account. This doesn't work very well for a lot of people, because the reinforcement happens too slowly to be habit-forming. A different way is to notice the logic of the situation, and think *as though* you had lived it.

Similarly, although people are heavily biased to treat vivid examples from the news (airplane crashes involving celebrity deaths) or personal anecdotes from friends and family (an uncle who fell ill when a high-voltage power line was installed near his home) or worse, from strangers on the internet (the guy who got "popcorn lung" from smoking an e-cig), actually, statistics are far more powerful. (It's true that medical anecdotes from family might be more relevant than statistics due to genetic factors shared with family members, but even so, taking the "statistical view" -- looking at the statistics available, including information about genetic conditions and heritability if available, and then making reasonable guesses about yourself based on your family -- will be better than just viewing your situation in isolation.)

I won't lecture much on the implications of that -- most readers will be familiar with the availability heuristic, the base-rate fallacy, and scope insensitivity. Here, I just want to point out that putting more credence in numbers than vivid examples is an instance of the pattern I'm pointing at: placing your decision as an instance of a class, rather than seeing it in isolation.

In an abstract sense, the statistical view of any given decision -- the view of the decision as part of a class of relevantly similar decisions -- is "more real" than an attempt to view it in isolation as a unique moment. Not because it's actually more real, but because this view is closer to your decision. Humans may exist in a single, deterministic universe -- but we *decide* in statistical space, where outcomes come in percentages.

When the weatherman says "70% chance of rain" in your area, but it is *already* raining outside, you know you're one of the 70% -- he's speaking to an area, and giving the percent of viewers in the area who will experience rain. He can't just say 100% for those viewers who will experience rain and 0% for those who won't. Similarly, when you make a decision, you're doing it in a general area -- you can't just decide to drive when you won't crash and avoid driving when you will (though you can split up your decision by some relevant factors and avoid driving when risk is high).

This exact same intuition, of course, supports timeless/updateless decision theory even more directly than it supports probabilism. Perhaps some future generation will regard the idea of timeless/updateless decision-making as more fundamental and obvious than the doctrine of probabilism, and wonder why subjective probability was invented first.

The Problematic Third Person Perspective

[Epistemic status: I now endorse this again. Michael pointed out a possibility for downside risk with losing mathematical ability, which initially made me update away from the view here. However, some experience noticing what it is like to make certain kinds of mathematical progress made me return to the view presented here. Maybe don't take this post as inspiration to engage in extreme rejection of objectivity.]

There are a number of conversational norms based on the idea of an imaginary impartial observer who needs to be convinced. It's the adversarial courtroom model of conversation. Better norms, such as common crux, can be established by recognizing that a conversation is taking place between two people.

Burden-of-proof is one of these problematic ideas. The idea that there is some kind of standard which would put the burden on one person or another would only make sense if there were a judge to convince. If anything, it would be better to say the burden of proof is on both people in any argument, in the sense that they are responsible for conveying their own views to the other person. If burden-of-proof is about establishing that they "should" give in to your position, it accomplishes nothing; you need to convince *them* of that, not yourself. If burden-of-proof is about establishing that you don't have to believe them until they say more... well, that was true anyway, but perhaps speaks to a lack of curiosity on your part.

More generally, this external-judge intuition promotes the bad model that there are objective standards of logic which must be adhered to in a debate. There *are* epistemic standards which it is *good* to adhere to, including logic and notions of probabilistic evidence. But, if the other person has different standards, then you have to either work with them or discuss the differences. There's a failure mode of the overly rationalistic where you just get angry that *their* arguments are illogical and they're not accepting *your* perfectly-formatted arguments, so you try to get them to bow down to your standards by force of will. (The same failure mode applies to treating definitions as objective standards which must be adhered to.) What good does it do to continue arguing with them via standards you already know differ from theirs? Try to understand and engage with their real reasons rather than replacing them with imaginary things.

Actually, it's even worse than this, because you don't know your own standards of evidence completely. So, the imaginary impartial judge is also interfering with your ability to get in touch with your real reasons, what you really think, and what might sway you one way or the other. If your mental motion is to reach for justifications which the impartial judge would accept, you are rationalizing rather than finding your true rejection. You have to realize that you're using standards of evidence that you yourself don't fully understand, and live in that world -- otherwise you rob yourself of the ability to improve your tools.

This happens in two ways, that I can think of.

- Maybe your explicit standards are good, but not perfect. You notice beliefs that are not up to your standards, and you drop them reflexively. This might be a good idea most of the time, but there are two things wrong with the policy. First,

you might have dropped a good belief. You could have done better by checking which you trusted more in this instance: the beliefs, or your standards of belief. Second, you've missed an opportunity to improve your explicit standards. You could have explored your reasons for believing what you did, and compared them to your explicit standards for belief.
- Maybe you don't notice the difference between your explicit standards and the way you actually arrive at your beliefs. You assume implicitly that if you believe something strongly, it's because there are strong reasons of the sort you endorse. This is especially likely if the beliefs pattern-match to the sort of thing your standards endorse; for example, being very sciency. As a result, you miss an opportunity to notice that you're rationalizing something. You would have done better to first look for the reasons you *really* believed the thing, and then check whether they meet your explicit standards and whether the belief still seems worth endorsing.

So far, I've argued that the imaginary judge creates problems in two domains: navigating disagreements with other people, and navigating your own epistemic standards. I'll note a third domain where the judge seems problematic: judging your own actions and decisions. Many people use an imaginary judge to guide their actions. This leads to pitfalls such as moral self-licensing, in which doing good things gives you a license to do more bad things (setting up a budget makes you feel good enough about your finances that you can go on a spending spree, eating a salad for lunch makes you more likely to treat yourself with ice cream after work, etc). Getting rid of the internal judge is an instance of Nate's Replacing Guilt, and carries similar risks: if you're currently using the internal judge for a bunch of important things, you have to either make sure you replace it with other working strategies, or be OK with kicking those things to the roadside (at least temporarily).

Similarly with the other two categories I mentioned. Noticing the dysfunctions of the imaginary-judge perspective should not make you immediately remove it; invoke Chesterton's Fence. However, I would encourage you to experiment with removing the imaginary third person from your conversations, and seeing what you do when you remind yourself that there's no one looking over your shoulder in your private mental life. I think this relates to a larger ontological shift which Val was also pointing toward in In Praise of Fake Frameworks. There is no third-person perspective. There is no view from nowhere. This isn't a rejection of reductionism, but a reminder that we haven't finished yet. This isn't a rejection of the principles of rationality, but a reminder that we are created already in motion, and there is no argument so persuasive it would move a rock.

And, more basically, it is a reminder that the map is not the territory, because humans confuse the two by default. The picture in your head isn't what's there to be seen. Putting pieces of your judgement inside an imaginary impartial judge doesn't automatically make it true. Perhaps it does really make it more trustworthy -- you "promote" your better heuristics by wrapping them up inside the judge, giving them authority over the rest. But, this system has its problems. It can create perverse incentives on the other parts of your mind, to please the judge in ways that let them get away with what they want. It can make you blind to other ways of being. It can make you *think* you've avoided map-territory confusion once and for all -- "See? It's written right there on my soul: DO NOT CONFUSE MAP AND TERRITORY. It is simply something I don't do." -- while really passing the responsibility to a special part of your map which is now almost *always* confused for the territory.

So, laugh at the judge a little. Look out for your real reasons for thinking and doing things. Notice whether your arguments seem tailored to convince your judge rather than the person in front of you. See where it leads you.

Confusions Concerning Pre-Rationality

Robin Hanson's Uncommon Priors Require Origin Disputes is a short paper with, according to me, a surprisingly high ratio of does-something-interesting-there per character. It is not clearly *right,* but it merits some careful consideration. If it *is* right, it offers strong reason in support of the common prior assumption, which is a major crux of certain modest-epistemology flavored arguments.

Wei Dai wrote two posts reviewing the concepts in the paper and discussing problems/implications. I recommend reviewing those before reading the present post, and possibly the paper itself as well.

Robin Hanson's notion of pre-rationality is: an agent's counterfactual beliefs should treat the details of its creation process like an update. If the agent is a Bayesian robot with an explicitly programmed prior, then the agent's distribution after conditioning on any event "programmer implements prior ***p***" should be exactly ***p***.

These beliefs are "counterfactual" in that agents are typically assumed to know their priors already, so that the above conditional probability is not well-defined for any choice of ***p*** other than the agent's true prior. This fact leads to a major complication in the paper; the pre-rationality condition is instead stated in terms of hypothetical "pre-agents" which have "pre-priors" encoding the agent's counterfactual beliefs about what the world would have been like if the agent had had a different prior. (I'm curious what happens if we drop that assumption, so that we can represent pre-rationality within the agent's same prior.)

Wei Dai offers an example in which a programmer flips a coin to determine whether a robot believes coin-flips to have probability 2/3rds or 1/3rd. Pre-rationality seems like an implausible constraint to put on this robot, because the programmer's coin-flip is not good reason to form such expectations about other coins.

Wei Dai seems to be arguing against a position which Robin Hanson isn't quite advocating. Wei Dai's accusation is that pre-rationality implies a belief that the process which created you was itself a rational process, which is not always plausible. Indeed, it's easy to see this interpretation in the math. However, Robin Hanson's response indicates that he doesn't see it:

> I just don't see pre-rationality being much tied to whether you in fact had a rational creator. The point is, as you say, to consider the info in the way you were created.

Unfortunately, the discussion in the comments doesn't go any further on this point. However, we can make some inferences about Robin Hanson's position from the paper itself.

Robin Hanson does not discuss the robot/programmer example; instead, he discusses the possibility that people have differing priors due to genetic factors. Far from claiming people are obligated by rationality principles to treat inherited priors as rational, Robin Hanson says that *because* we know some randomness is involved in Mendelian inheritance, we can't *both* recognize the arbitrariness of our prior's origin *and* stick with that prior. Quoting the paper on this point:

> Mendel’s rules of genetic inheritance, however, are symmetric and random between siblings. If optimism were coded in genes, you would not acquire an optimism gene in situations where optimism was more appropriate, nor would your sister’s attitude gene track truth any worse than your attitude gene does.
>
> Thus it seems to be a violation of pre-rationality to, conditional on accepting Mendel’s rules, allow one’s prior to depend on individual variations in genetically-encoded attitudes. Having your prior depend on species-average genetic attitudes may not violate pre-rationality, but this would not justify differing priors within a species.

Robin Hanson suggests that pre-rationality is only plausible conditional on some knowledge we have gained throughout our lifetime about our own origins. He posits a sentence ***B*** which contains this knowledge, and suggests that the pre-rationality condition can be relativized to ***B***. In the above-quoted case, ***B*** would consist of Mendelian inheritance and the genetics of optimism. Robin Hanson is not saying that genetic inheritance of optimism or pessimism is a rational process, but rather, he is saying that once we know about these genetic factors, we should adjust our pessimism or optimism toward the species average. *After* performing this adjustment, we are pre-rational: we consider any *remaining* influences on our probability distribution to have been rational.

Wei Dai's argument might be charitably interpreted as objecting to this position by offering a concrete case in which a rational agent does not update to pre-rationality in this way: the robot has no motivation to adjust for the random noise in its prior, despite its recognition of the irrationality of the process by which it inherited this prior. However, I agree with Robin Hanson that this is intuitively quite problematic, even if no laws of probability are violated. There is *something wrong* with the robot's position, even if the robot lacks cognitive tools to escape this epistemic state.

However, Wei Dai does offer a significant response to this: he complains that Robin Hanson says too little about what the robot should do to become pre-rational from its flawed state. The pre-rationality condition provides no guidance for the robot. As such, what guidance can pre-rationality offer to humans? Robin Hanson's paper admits that we have to condition on ***B*** to become pre-rational, but offers no account whatsoever about the structure of this update. What normative structure should we require of priors so that an agent *becomes* pre-rational when conditioned on the appropriate ***B***?

Here is the text of Wei Dai's sage complaint:

> Assuming that we do want to be pre-rational, how do we move from our current non-pre-rational state to a pre-rational one? This is somewhat similar to the question of how do we move from our current non-rational (according to ordinary rationality) state to a rational one. Expected utility theory says that we should act as if we are maximizing expected utility, but it doesn't say what we should do if we find ourselves lacking a prior and a utility function (i.e., if our actual preferences cannot be represented as maximizing expected utility).
>
> The fact that we don't have good answers for these questions perhaps shouldn't be considered fatal to pre-rationality and rationality, but it's troubling that little attention has been paid to them, relative to defining pre-rationality and rationality. (Why are rationality researchers more interested in knowing what rationality is, and less interested in knowing how to be rational? Also, BTW, why are there so

few rationality researchers? Why aren't there hordes of people interested in these issues?)

I find myself in the somewhat awkward position of agreeing strongly with Robin Hanson's intuitions here, but also having no idea how it should work. For example, suppose that we have a robot whose probabilistic beliefs are occasionally modified by cosmic rays. These modification events can be thought of as the environment writing a new "prior" into the agent. We cannot perfectly safeguard the agent against this, but we can write the agent's probability distribution such that so long as it is not *too* damaged, it can self-repair when it sees evidence that its beliefs have been modified by the environment. This seems like an updating-to-pre-rationality move, with "a cosmic ray hit you in this memory cell" playing the role of ***B***.

Similarly, it seems reasonable to do something like average beliefs with someone if you discover that your differing beliefs are due only to genetic chance. Yet, it does not seem similarly reasonable to average values, despite the distinction between beliefs and preferences being somewhat fuzzy.

This is made even more awkward by the fact that Robin Hanson has to create the whole pre-prior framework in order to state his new rationality constraint.

The idea seems to be that a pre-prior is not a belief structure which an actual agent has, but rather, is a kind of plausible extrapolation of an agent's belief structure which we layer on top of the true belief structure in order to reason about the new rationality constraint. If so, how could this kind of rationality constraint be compelling to an agent? The agent itself doesn't have any pre-prior. Yet, if we have an intuition that Robin Hanson's argument implies something about humans, then *we ourselves* are agents who find arguments involving pre-priors to be relevant.

Alternatively, pre-priors could be capturing information about counterfactual beliefs which the agent itself has. This seems less objectionable, but it brings in tricky issues of counterfactual reasoning. I don't think this is likely to be the right path to properly formalizing what is going on, either.

I see two clusters of approaches:

- What rationality conditions might we impose on a Bayesian agent such that it updates to pre-rationality given "appropriate" ***B***? Can we formalize this purely within the agent's own prior, without the use of pre-priors?
- What can we say about agents becoming rational from irrational positions? What should agents do when they notice Dutch Books against their beliefs, or money-pumps against their preferences? (Logical Induction is a somewhat helpful story about the former, but not the latter.) Can we characterize the *receiver* of decision-theoretic arguments such as the VNM theorem, who would find such arguments interesting? If we can produce anything in this direction, can it say anything about Robin Hanson's arguments concerning pre-rationality? Does it give a model, or can it be modified to give a model, of updating to pre-rationlity?

It seems to me that there is something interesting going on here, and I wish that there were more work on Hansonian pre-rationality and Wei Dai's objection.