Consequences of Logical Induction

1. Toward a New Technical Explanation of Technical Explanation
2. In Logical Time, All Games are Iterated Games
3. Do Sufficiently Advanced Agents Use Logic?
4. An Orthodox Case Against Utility Functions
5. What does it mean to apply decision theory?
6. Radical Probabilism [Transcript]
7. Radical Probabilism
8. The Bayesian Tyrant
9. Time Travel Markets for Intellectual Accounting
10. Probability vs Likelihood
11. Reflective Bayesianism

Toward a New Technical Explanation of Technical Explanation

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

A New Framework

(Thanks to Valentine for a discussion leading to this post, and thanks to CFAR for running the CFAR-MIRI cross-fertilization workshop. Val provided feedback on a version of this post. Warning: fairly long.)

Eliezer's *A Technical Explanation of Technical Explanation*, and moreover the sequences as a whole, used the best technical understanding of practical epistemology available at the time* -- the Bayesian account -- to address the question of how humans can try to arrive at better beliefs in practice. The sequences also pointed out several holes in this understanding, mainly having to do with logical uncertainty and reflective consistency.

MIRI's research program has since then made major progress on logical uncertainty. The new understanding of epistemology -- the theory of logical induction -- generalizes the Bayesian account by eliminating the assumption of logical omniscience. Bayesian belief updates are recovered as a special case, but the dynamics of belief change are non-Bayesian in general. While it might not turn out to be the last word on the problem of logical uncertainty, it has a large number of desirable properties, and solves many problems in a unified and relatively clean framework.

It seems worth asking what consequences this theory has for practical rationality. Can we say new things about what good reasoning looks like in humans, and how to avoid pitfalls of reasoning?

First, I'll give a shallow overview of logical induction and possible implications for practical epistemic rationality. Then, I'll focus on the particular question of *A Technical Explanation of Technical Explanation* (which I'll abbreviate TEOTE from now on). Put in CFAR terminology, I'm seeking a gears-level understanding of gears-level understanding. I focus on the intuitions, with only a minimal account of how logical induction helps make that picture work.

Logical Induction

There are a number of difficulties in applying Bayesian uncertainty to logic. No computable probability distribution can give non-zero measure to the logical tautologies, since you can't bound the amount of time you need to think to check whether something is a tautology, so updating on provable sentences always means updating on a set of measure zero. This leads to convergence problems, although there's been recent progress on that front.

Put another way: Logical consequence is deterministic, but due to Gödel's first incompleteness theorem, it is like a stochastic variable in that there is no computable procedure which correctly decides whether something is a logical consequence. This means that any computable probability distribution has infinite Bayes loss on the question of logical consequence. Yet, because the question is actually deterministic, we know how to point in the direction of better distributions by doing more and more consistency checking. This puts us in a puzzling situation where we want to improve the Bayesian probability distribution by doing a kind of non-Bayesian update. This was the two-update problem.

You can think of logical induction as supporting a set of hypotheses which are about ways to shift beliefs as you think longer, rather than fixed probability distributions which can only shift in response to evidence.

This introduces a new problem: how can you score a hypothesis if it keeps shifting around its beliefs? As TEOTE emphasises, Bayesians outlaw this kind of belief shift for a reason: requiring predictions to be made in advance eliminates hindsight bias. (More on this later.) So long as you understand exactly what a hypothesis predicts and what it does not predict, you can evaluate its Bayes score and its prior complexity penalty and rank it objectively. How do you do this if you don't know all the consequences of a belief, and the belief itself makes shifting claims about what those consequences are?

The logical-induction solution is: set up a prediction market. A hypothesis only gets credit for contributing to collective knowledge by moving the market in the right direction early. If the market's odds on prime numbers are currently worse than those which the prime number theorem can provide, a hypothesis can make money by making bets in that direction. If the market has already converged to those beliefs, though, a hypothesis can't make any more money by expressing such beliefs -- so it doesn't get any credit for doing so. If the market has moved on to even more accurate rules of thumb, a trader would only lose money by moving beliefs back in the direction of the prime number theorem.

Mathematical Understanding

This provides a framework in which we can make sense of mathematical labor. For example, a common occurrence in combinatorics is that there is a sequence which we can calculate, such as the catalan numbers, by directly counting the number of objects of some specific type. This sequence is boggled at like data in a scientific experiment. Different patterns in the sequence are observed, and hypotheses for the continuation of these patterns are proposed and tested. Often, a significant goal is the construction of a closed form expression for the sequence.

This looks just like Bayesian empiricism, except for the fact that *we already have a hypothesis which entirely explains the observations.* The sequence is constructed from a definition which mathematicians made up, and which thus assigns 100% probability to the observed data. What's going on? It is possible to partially explain this kind of thing in a Bayesian framework by acting *as if* the true formula were unknown and we were trying to guess where the sequence came from, but this doesn't explain everything, such as why finding a closed form expression would be important.

Logical induction explains this by pointing out how different time-scales are involved. Even if all elements of the sequence are calculable, a new hypothesis can get credit

for calculating them faster than the brute-force method. Anything which allows one to produce correct answers faster contributes to the efficiency of the prediction market inside the logical inductor, and thus, to the overall mathematical understanding of a subject. This cleans up the issue nicely.

What other epistemic phenomena can we now understand better?

Lessons for Aspiring Rationalists

Many of these could benefit from a whole post of their own, but here's some fast-and-loose corrections to Bayesian epistemology which may be useful:

- Hypotheses need not make predictions about everything. Because hypotheses are about how to *adjust* your odds as you think longer, they can leave most sentences alone and focus on a narrow domain of expertise. Everyone was already doing this in practice, but the math of Bayesian probability theory requires each hypothesis to make a prediction about every observation, if you actually look at it. Allowing a hypothesis to remain silent on some issues in standard Bayesianism can cause problems: if you're not careful, a hypothesis can avoid falsification by remaining silent, so you end up incentivising hypotheses to remain mostly silent (and you fail to learn as a result). Prediction markets are one way to solve this problem.
- Hypotheses buy and sell at the *current price*, so they take a hit for leaving a now-unpopular position which they initially supported (but less of a hit than if they'd stuck with it) or coming in late to a position of growing popularity. Other stock-market type dynamics can occur.
- Hypotheses can be like object-level beliefs or meta-level beliefs: you can have a hypothesis about how you're overconfident, which gets credit for smoothing your probabilities (if this improves things on average). This allows you to take into account beliefs about your calibration without getting *too* confused about Hofstadter's-law type paradoxes.

You may want to be a bit careful and Chesterton-fence existing Bayescraft, though, because some things are still better about the Bayesian setting. I mentioned earlier that Bayesians don't have to worry so much about hindsight bias. This is closely related to the problem of old evidence.

Old Evidence

Suppose a new scientific hypothesis, such as general relativity, explains a well-know observation such as the perihelion precession of mercury better than any existing theory. Intuitively, this is a point in favor of the new theory. However, the probability for the well-known observation was already at 100%. How can a previously-known statement provide new support for the hypothesis, as if we are re-updating on evidence we've already updated on long ago? This is known as the problem of old evidence, and is usually levelled as a charge against Bayesian epistemology. However, in some sense, the situation is worse for logical induction.

A Bayesian who endorses Solomonoff induction can tell the following story: Solomonoff induction is the right theory of epistemology, but we can only approximate it, because it is uncomputable. We approximate it by searching for hypotheses, and computing their posterior probability retroactively when we find new ones. It only makes sense

that when we find a new hypothesis, we calculate its posterior probability by multiplying its prior probability (based on its description length) by the probability it assigns to all evidence so far. That's Bayes' Law! The fact that we already knew the evidence is not relevant, since our approximation didn't previously include this hypothesis.

Logical induction speaks against this way of thinking. The hypothetical Solomonoff induction advocate is assuming one way of approximating Bayesian reasoning via finite computing power. Logical induction can be thought of as a different (more rigorous) story about how to approximate intractible mathematical structures. In this new way, *propositions are bought or sold at market prices at the time.* If a new hypothesis is discovered, it can't be given any credit for 'predicting' old information. The price of known evidence is already at maximum -- you can't gain any money by investing in it.

There are good reasons to ignore old evidence, especially if the old evidence has biased your search for new hypotheses. Nonetheless, it doesn't seem right to *totally* rule out this sort of update.

I'm still a bit puzzled by this, but I think the situation is improved by understanding gears-level reasoning. So, let's move on to the discussion of TEOTE.

Gears of Gears

As Valentine noted in his article, it is somewhat frustrating how the overall idea of gears-level understanding seems so clear while remaining only heuristic in definition. It's a sign of a ripe philosophical puzzle. If you don't feel you have a good intuitive grasp of what I mean by "gears level understanding", I suggest reading his post.

Valentine gives three tests which point in the direction of the right concept:

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

I already named one near-synonym for "gears", namely "technical explanation". Two more are "inside view" and Elon Musk's notion of reasoning from first principles. The implication is supposed to be that gears-level understanding is in some sense better than other sorts of knowledge, but this is decidedly not supposed to be valued to the exclusion of other sorts of knowledge. Inside-view reasoning is traditionally supposed to be combined with outside-view reasoning (although Elon Musk calls it "reasoning by analogy" and considers it inferior, and much of Eliezer's recent writing warns of its dangers as well, while allowing for its application to special cases). I suggested the terms gears-level & policy-level in a previous post (which I actually wrote after most of this one).

Although TEOTE gets close to answering Valentine's question, it doesn't quite hit the mark. The definition of "technical explanation" provided there is a theory which strongly concentrates the probability mass on specific predictions and rules out others. It's clear that a model can do this without being "gears". For example, my

model might be that whatever prediction the Great Master makes will come true. The Great Master can make very detailed predictions, but I don't know how they're generated. I lack the understanding associated with the predictive power. I might have a strong outside-view reason to trust the Great Master: their track record on predictions is immaculate, their Bayes-loss miniscule, their calibration supreme. Yet, I lack an inside-view account. I can't derive their predictions from first principles.

Here, I'm siding with David Deutsch's account in the first chapter of *The Fabric of Reality*. He argues that understanding and predictive capability are distinct, and that understanding is about having good explanations. I may not accept his whole critique of Bayesianism, but that much of his view seems right to me. Unfortunately, he doesn't give a *technical* account of what "explanation" and "understanding" could be.

First Attempt: Deterministic Predictions

TEOTE spends a good chunk of time on the issue of making predictions in advance. According to TEOTE, this is a human solution to a human problem: you make predictions in advance so that you can't make up what predictions you could have made after the fact. This counters hindsight bias. An ideal Bayesian reasoner, on the other hand, would never be tempted into hindsight bias in the first place, and is free to evaluate hypotheses on old evidence (as already discussed).

So, is gears-level reasoning just pure Bayesian reasoning, in which hypotheses have strictly defined probabilities which don't depend on anything else? Is outside-view reasoning the thing logical induction adds, by allowing the beliefs of a hypothesis to shift over time and to depend on on the wider market state?

This isn't quite right. An ideal Bayesian can still learn to trust the Great Master, based on the reliability of the Great Master's predictions. Unlike a human (and unlike a logical inductor), the Bayesian will at all times have in mind all the possible ways the Great Master's predictions *could* have become so accurate. This is because a Bayesian hypothesis contains a full joint distribution on all events, and an ideal Bayesian reasons about all hypotheses at all times. In this sense, the Bayesian always operates from an inside view -- it cannot trust the Great Master without a hypothesis which correlates the Great Master with the world.

However, it is possible that this correlation is introduced in a very simple way, by ruling out cases where the Great Master and reality disagree without providing any mechanism explaining how this is the case. This may have low prior probability, but gain prominence due to the hit in Bayes-score other hypotheses are taking for not taking advantage of this correlation. It's not a bad outcome given the epistemic situation, but it's not gears-level reasoning, either. So, being fully Bayesian or not isn't *exactly* what distinguishes whether advanced predictions are needed. What is it?

I suggest it's this: *whether the hypothesis is well-defined, such that anyone can say what predictions it makes without extra information.* In his post on gears, Valentine mentions the importance of "*how deterministically interconnected the variables of the model are*". I'm pointing at something close, but importantly distinct: how deterministic the *predictions* are. You know that a coin is very close to equally likely to land on heads or tails, and from this you can (if you know a little combinatorics) compute things like the probability of getting exactly three heads if you flip the coin five times. Anyone with the same knowledge would compute the same thing. The

model includes probabilities inside it, but how those probabilities flow is perfectly deterministic.

This is a notion of objectivity: a wide variety of people can agree on what probability the model assigns, despite otherwise varied background knowledge.

If a model is well-defined in this way, it is very easy (Bayesian or no) to avoid hindsight bias. You cannot argue about how you could have predicted some result. Anyone can sit down and calculate.

The hypothesis that the Great Master is always correct, on the other hand, does not have this property. Nobody but the Great Master can say what that hypothesis predicts. If I know what the Great Master says about a particular thing, I can evaluate the accuracy of the hypothesis; but, this is special knowledge which I need in order to give the probabilities.

The Bayesian hypothesis which simply forces statements of the Great Master to correlate with the world is somewhat more gears-y, in that there's a probability distribution which can be written down. However, this probability distribution is a complicated mish-mosh of the Bayesian's other hypotheses. So, predicting what it would say requires extensive knowledge of the private beliefs of the Bayesian agent involved. This is typical of the category of non-gears-y models.

Objection: Doctrines

Infortunately, this account doesn't totally satisfy what Valentine wants.

Suppose that, rather than making announcements on the fly, the Great Master has published a set of fixed Doctrines which his adherents memorize. As in the previous thought experiment, the word of the Great Master is infallible; the application of the Doctrines always leads to correct predictions. However, the contents of the Doctrines appears to be a large mish-mosh of rules with no unifying theme. Despite their apparent correctness, they fail to provide any understanding. It is as if a physicist took all the equations in a physics text, transformed them into tables of numbers, and then transported those tables to the middle ages with explanations of how to use the tables (but none of where they come from). Though the tables work, they are opaque; there is no insight as to how they were determined.

The Doctrines are a deterministic tool for making predictions. Yet, they do not seem to be a gears-level model. Going back to Valentine's three tests, the Doctrines fail test three: we could erase any one of the Doctrines and we'd be unable to rederive it by how it fit together with the rest. Hence, the Doctrines have almost as much of a "trust the Great Master" quality as listening to the Great Master directly -- the disciples would not be able to derive the Doctrines for themselves.

Second Attempt: Proofs, Axioms, & Two Levels of Gears

My next proposal is that *having a gears-level model is like knowing the proof*. You might believe a mathematical statement because you saw it in a textbook, or because

you have a strong mathematical intuition which says it must be true. But, you don't have the gears until you can prove it.

This subsumes the "deterministic predictions" picture: a model is an axiomatic system. If we know all the axioms, then we can in theory produce all the predictions ourselves. (Thinking of it this way introduces a new possibility, that the model may be well-defined but we may be unable to *find* the proofs, due to our own limitations.) On the other hand, we don't have access to the axioms of the theory embodied by the Great Master, and so we have no hope of seeing the proofs; we can only observe that the Great Master is always right.

How does this help with the example of the Doctrines?

The concept of "axioms" is somewhat slippery. There are many equivalent ways of axiomatizing any given theory. We can often flip views between what's taken as an axiom vs what's proved as a theorem. However, the most elegant set of axioms tends to be preferred.

So, we *can* regard the Doctrines as one long set of axioms. If we look at them that way, then adherents of the Great Master have a gears-level understanding of the Doctrines if they can successfully apply them as instructed.

However, the Doctrines are not an elegant set of axioms. So, viewing them in this way is very unnatural. It is more natural to see them as a set of assertions which the Great Master has produced by some axioms unknown to us. In this respect, we "can't see the proofs".

In the same way, we can consider flipping any model between the axiom view and the theorem view. Regarding the model as axiomatic, to determine whether it is gears-level we only ask whether its predictions are well-defined. Regarding in in "theorem view", we ask if we know how *the model itself* was derived.

Hence, two of Valentine's desirable properties of a gears-level model can be understood as the same property applied at different levels:

- *Determinism*, which is Val's property #2, follows from requiring that we can see the derivations within the model.
- *Reconstructability,* Val's property #3, follows from requiring that we can see the derivation *of* the model.

We might call the first level of gears "made out of gears", and the second level "made *by* gears" -- the model itself being constructed via a known mechanism.

If we change our view so that a scientific theory is a "theorem", what are the "axioms"? Well, there are many criteria which are applied to scientific theories in different domains. These criteria could be thought of as pre-theories or meta-theories. They encode the hard-won wisdom of a field of study, telling us what theories are likely to work or fail in that field. But, a very basic axiom is: we want a theory to be *the simplest theory consistent with all observations.* The Great Master's Doctrines cannot possibly survive this test.

To give a less silly example: if we train up a big neural network to solve a machine learning problem, the predictions made by the model are deterministic, predictable from the network weights. However, someone else who knew all the principles by which the network was created would nonetheless train up a very different neural

network -- unless they use the very same gradient descent algorithm, data, initial weights, and number and size of layers.

Even if they're the same in all those details, and so reconstruct the same neural network *exactly,* there's a significant sense in which they can't *see* how the conclusion follows inevitably from the initial conditions. It's less doctrine-y than being handed a neural network, but it's more doctrine-y than understanding the structure of the problem and why almost any neural network achieving good performance on the task will have certain structures. Remember what I said about mathematical understanding. There's always another level of "being able to see why" you can ask for. Being able to reproduce the proof is different from being able to explain why the proof has to be the way it is.

Exact Statement?

Gears-y ness is a matter of degree, and there are several interconnected things we can point at, and a slippage of levels of analysis which makes everything quite complicated.

In the ontology of math/logic, we can point at whether you can see the proof of a theorem. There are several slippages which make this fuzzier than it may seem. First: do you derive it only form the axioms, or do you use commonly known theorems and equivalences (which you may or may not be able to prove if put on the spot)? There's a long continuum between what one mathematician might say to another as proof and a formal derivation in logic. Second: how well can you see why the proof has to be? This is the spectrum between following each proof step individually (but seeing them as almost a random walk) vs seeing the proof as an elementary application of a well-known technique. Third: we can start slipping the axioms. There are small changes to the axioms, in which one thing goes from being an axiom to a theorem and another thing makes the opposite transition. There are also large changes, like formalizing number theory via the Peano axioms vs formalizing it in set theory, where the entire description language changes. You need to translate from statements of number theory to statements of set theory. Also, there is a natural ambiguity between taking something as an axiom vs requiring it as a condition in a theorem.

In the ontology of computation, we can point at knowing the output of a machine vs being able to run it by hand to show the output. This is a little less flexible than the concept of mathematical proof, but essentially the same distinction. Changing the axioms is like translating the same algorithm to a different computational formalism, like going between Turing machines and lambda calculus. Also, there is a natural ambiguity between a program vs an input: when you run program XYZ with input ABC on a universal Turing machine, you input XYZABC to the universal turing machine; but, you can also think of this as running program XY on input ZABC, or XYZA on input BC, et cetera.

In the ontology of ontology, we could say "can you see why this has to be, from the structure of the ontology describing things?" "Ontology" is less precise than the previous two concepts, but it's clearly the same idea. A different ontology doesn't necessarily support the same conclusions, just like different axioms don't necessarily give the same theorems. However, the reductionist paradigm holds that the ontologies we use should all be consistent with one another (under some translation between the ontologies). At least, aspire to be eventually consistent. Analogous to axiom/assumption ambiguity and program/input ambiguity, there is ambiguity

between an ontology and the cognitive structure which created and justifies the ontology. We can also distinguish more levels; maybe we would say that an ontology doesn't make predictions directly, but provides a language for stating models, which make predictions. Even longer chains can make sense, but it's all subjective divisions. However, unlike the situation in logic and computation, we can't expect to articulate the full support structure for an ontology; it is, after all, a big mess of evolved neural mechanisms which we don't have direct access to.

Having established that we can talk about the same things in all three settings, I'll restrict myself to talking about ontologies.

Two-level definition of gears: A conclusion is gears-like with respect to a particular ontology to the extent that you can "see the derivation" in that ontology. A conclusion is gears-like without qualification to the extent that you can also "see the derivation" of the ontology itself. This is contiguous with gears-ness relative to an ontology, because of the natural ambiguity between programs and their inputs, or between axioms and assumptions. For a given example, though, it's generally more intuitive to deal with the two levels separately.

Seeing the derivation: There are several things to point at by this phrase.

- As in TEOTE, we might consider it important that a model make *precise* predictions. This could be seen as a prerequisite of "seeing the derivation": first, we must be saying something *specific*; then, we can ask if we can say why we're saying that particular thing. This implies that models are more gears-like when they are more deterministic, all other things being equal.
- However, I think it is also meaningful and useful to talk about whether the *predictions* of the model are deterministic; the standard way of assigning probabilities to dice is very gears-like, despite placing wide probabilities. I think these are simply two different important things we can talk about.
- Either way, being able to see the derivation is like being able to see the proof or execute the program, with all the slippages this implies. You see the derivation less well to the extent that you rely on known theorems, and more to the extent that you can spell out all the details yourself if need be. You see it less well to the extent that you understand the proof only step-by-step, and more well to the extent that you can derive the proof as a natural application of known principles. You cannot see the derivation if you don't even have access to the program which generated the output, or are missing some important inputs for that program.

Seeing the derivation is about explicitness and external objectivity. You can trivially "execute the program" generating any of your thoughts, in that you thinking *is* the program which generated the thoughts. However, the execution of this program could rely on arbitrary details of your cognition. Moreover, these details are usually not available for conscious access, which means you can't explain the train of thought to others, and even you may not be able to replicate it later. So, a model is more gears-like the more *replicable* it is. I'm not sure if this should be seen as an additional requirement, or an explanation of where the requirements come from.

Conclusion, Further Directions

Obviously, we only touched the tip of the iceberg here. I started the post with the claim that I was trying to hash out the implications of logical induction for practical

rationality, but secretly, the post was about things which logical inductors can only barely begin to explain. (I think these two directions support each other, though!)

We need the framework of logical induction to understand some things here, such as how you still have degrees of understanding when you already have the proof / already have a program which predicts things perfectly (as discussed in the "mathematical understanding" section). However, logical inductors don't look like they care about "gears" -- it's not very close to the formalism, in the way that TEOTE gave a notion of technical explanation which is close to the formalism of probability theory.

I mentioned earlier that logical induction suffers from the old evidence problem more than Bayesianism. However, it doesn't suffer in the sense of losing bets it could be winning. Rather, *we* suffer, when we try to wrap our heads around what's going on. Somehow, logical induction is learning to do the right thing -- the formalism is just not very explicit about how it does this.

The idea (due to Sam Eisenstat, hopefully not butchered by me here) is that logical inductors get around the old evidence problem by learning notions of objectivity.

A hypothesis you come up with later can't gain any credibility by fitting evidence from the past. However, if you register a prediction *ahead of time* that a particular hypothesis-generation process will eventually turn up something which fits the old evidence, you *can* get credit, and use this credit to bet on what the hypothesis claims will happen later. You're betting on a particular school of thought, rather than a known hypothesis. "You can't make money by predicting old evidence, but you may be able to find a benefactor who takes it seriously."

In order to do this, you need to specify a precise prediction-generation process which you are betting in favor of. For example, Solomonoff Induction can't run as a trader, because it is not computable. However, the probabilities which it generates are well-defined (if you believe that halting bits are well-defined, anyway), so you can make a business of betting that its probabilities will have been good in hindsight. If this business does well, then the whole market of the logical inductor will shift toward trying to make predictions which Solomonoff Induction will later endorse.

Similarly for other ideas which you might be able to specify precisely without being able to run right away. For example, you can't find all the proofs right away, but you could bet that all the theorems which the logical inductor observes *have* proofs, and you'd be right every time. Doing so allows the market to start betting it'll see theorems if it sees that they're provable, even if it hasn't yet seen this rule make a successful advance prediction. (Logical inductors start out really ignorant of logic; they don't know what proofs are or how they're connected to theorems.)

This doesn't *exactly* push toward gears-y models as defined earlier, but it seems close. You push toward anything for which you can provide an explicit justification, where "explicit justification" is anything you can name ahead of time (and check later) which pins down predictions of the sort which tend to correlate with the truth.

This doesn't mean the logical inductor converges entirely to gears-level reasoning. Gears were never supposed to be everything, right? The optimal strategy combines gears-like and non-gears-like reasoning. However, it *does* suggest that gears-like reasoning has an advantage over non-gears reasoning: it can gain credibility from old evidence. This will often push gears-y models above competing non-gears considerations.

All of this is still terribly informal, but is the sort of thing which could lead to a formal theory. Hopefully you'll give me credit later for that advanced prediction.

In Logical Time, All Games are Iterated Games

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

Logical Time

The main purpose of this post is to introduce the concept of logical time. The idea was mentioned in Scott's post, Bayesian Probability is for things that are Space-like Separated from You. It was first coined in a conference call with, Daniel Demski, Alex Mennan, and perhaps Corey Staten and Evan Lloyd -- I don't remember exactly who was there, or who first used the term. Logical time is an informal concept which serves as an intuition pump for thinking about logical causality and phenomena in logical decision theory; don't take it too seriously. In particular, I am not interested in anybody trying to formally define logical time (aside from formal approaches to logical causality). Still, it seems like useful language for communicating decision-theory intuitions.

Suppose you are playing chess, and you consider moving your bishop. You play out a hypothetical game which results in your loss in several moves. You decide not to move your bishop as a result of this. The hypothetical game resulting in your loss still exists within logic. You are logically later than it, in that the game you actually play depends on what happened in this hypothetical game.

Suppose you're stuck in the desert in a Parfit's Hitchhiker problem. Paul Ekman is reading your face, deciding whether you're trustworthy. Paul Ekman does this based on experience, meaning that the computation which is *you* has a strong similarity with other computations. This similarity can be used to predict you fairly reliably, based on your facial expressions. What creates this similarity? According to the logical time picture, there is a logical fact much earlier in logical time, which governs the connection between facial expressions and behavior.

To the extent that agents are trying to predict the future, they can be thought of as trying to place themselves later in logical time than the events which they're trying to predict. Two agents trying to predict each other are competing to see who can be later in logical time. This is not necessarily wise; in games like chicken, there is a sense in which you want to be earlier in logical time.

Traditional game theory, especially Nash equilibria, relies on what amounts to loopy logical causality to allow each agent to be after the other in logical time. Whether this is bad depends on your view on logical time travel. Perhaps there is a sense in which logical time can be loopy, due to prediction (which is like logical time travel). Perhaps logical time can't be loopy, and this is a flaw in the models used by traditional game theory.

Iterated Games

In logical time, all games are iterated games. An agent tries to forecast what happens in the decision problem it finds itself in by comparing it to similar decision problems which are small enough for it to look at. This puts it later in logical time than the small examples. "Similar games" includes the exact same game, but in which both players have had less time to think.

This means it is appropriate to use iterated strategies. Agents who are aware of logical time can play tit-for-tat in single-shot Prisoner's Dilemma, and so, can cooperate with each other.

Iterated games are different in character than single-shot games. The folk theorem shows that almost any outcome is possible in iterated play (in a certain sense). This makes it difficult to avoid very bad outcomes, such as nearly always defecting in the prisoner's dilemma, despite the availability of much better equilibria such as tit-for-tat. Intuitively, this is because (as Yoav Shoham et al point out in *If multi-agent learning is the answer, what is the question?*) it is difficult to separate "teaching behavior" from "learning behavior": as in the tit-for-tat strategy, it is generally wise to adopt behavior designed to shape the incentive gradient of the other player, in addition to improving your own score. Unfortunately, it is difficult to say what it means to pursue these two objectives simultaneously.

The subfield most related to this problem is multi-agent learning. Sadly, as discussed in the Shoham et al paper I cited parenthetically in the preceding paragraph, multi-agent learning typically avoids the difficulty by focusing on learning single-shot equilibria via iterated play. Learning single-shot equilibria in an iterated setting is a somewhat weird thing to be doing (hence the title of the paper). However, it is understandable that people might avoid such a difficult problem. The folk theorem illustrates a severe equilibrium selection issue, meaning that traditional tools have little to say about rational play.

One might imagine learning to play single-shot games by playing them over and over. But, what can you do to learn iterated games? You might imagine that you jump up a level again, experiencing the iterated version repeatedly to discover the optimal iterated strategy. However, iterating the game more doesn't really escape the iterated setting; there is no further level!

(You might think meta-iteration involves making the other player forget what it learned in iterated play so far, so that you can re-start the learning process, but that doesn't make much sense if you retain your own knowledge; and if you don't, you can't be learning!)

Toy Models

We can make pictures of logical time using phenomena which we understand more fully. One such picture is based on proofs. If we imagine a theorem prover proving every theorem in some order (such as an ordering based on proof length), we can think of logical time as time-of-proof. We can formulate counterfactuals consistent with this notion of logical time. (As I mentioned before, a picture of logical time is just a picture of logical causality / logical counterfactuals -- the notion of logical time adds nothing, formally.)

We can examine logical time travel in this kind of model by constructing predictors using stronger logics, which allows a predictor to find shorter proofs. This creates

decision-theoretic puzzles, because the agent with the weaker logic can't recognize the loopiness of the situation; it thinks it cannot influence the predictor, because (according to its weaker logic) the predictor has a short proof-length and is therefore earlier in logical time. We, on the other hand, can recognize that agents who act as if they control the predictor could do better in the decision problem.

This weirdness seems to only be possible because of the "two dimensional logical time" which exists in this toy model, in which we can vary both proof length and logical strength. One agent has access to arbitrarily long proofs via oracles, and so is "later" in the length dimension; the other has a stronger logic, and so is "later" in the strength dimension.

However, we can collapse the two dimensions into one via logical induction. Logical induction eventually learns to predict what stronger logics would predict, so computation time and logical strength are more or less the same.

You might expect that the loopy scenario in which an agent and a predictor accurately predict each other becomes impossible in logical induction, but, it does not. Logical-induction agents can predict each other well by examining what similar agents do in similar situations. As a result, LIDT agents converge to playing correlated equilibria with each other, more or less. (This result ignores the iterated aspect of the games, just like the multi-agent learning approaches I was complaining about earlier; despite learning from all the nearby copies within logic, the LIDT agents think only of the utility for their one decision, which paradoxically results in poorer outcomes even for that decision. Asymptotic decision theory does better, but no nice results for game theory have come from it so far.)

So long as an agent eventually settles down to making some reliable pattern of decisions in a situation, there will be relatively young logical inductors which have learned enough to accurately forecast the decisions made by logical-induction agents who reason using much more computational power.

We can think of the purely logical case, with its apparent two dimensions of logical time, as being a degenerate extreme version of the phenomenon in logical induction. In logical induction, the early predictions may be quite accurate, but they are fallible; they always run the risk of being wrong, since we're in the process of learning. In the pure logical case, we *also* run the risk of being wrong: using a stronger logic to make predictions runs the risk of introducing inconsistencies. This is easy to forget, since we are accustomed to the assumption that we can easily add axioms to a consistent system to get a stronger one.

An early predictor predicting a late agent must give up on some accuracy -- a prediction which relies on anything else than actually running the computation to be predicted has some chance of failure. This breaks the loopiness in logical time; the late agent always adds some small amount of information, even if its action is predicted with high reliability.

Do Sufficiently Advanced Agents Use Logic?

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

This is a continuation of a discussion with Vanessa from the MIRIxDiscord group. I'll make some comments on things Vanessa has said, but those should not be considered a summary of the discussion so far. My comments here are also informed by discussion with Sam.

1: Logic as Proxy

1a: The Role of Prediction

Vanessa has said that predictive accuracy is sufficient; consideration of logic is not needed to judge (partial) models. A hypothesis should ultimately ground out to perceptual information. So why is there any need to consider other sorts of "predictions" it can make? (IE, why should we think of it as possessing internal propositions which have a logic of their own?)

But similarly, why should agents use predictive accuracy to learn? What's the argument for it? Ultimately, predicting perceptions ahead of time should only be in service of achieving higher reward.

We could instead learn from reward feedback alone. A (partial) "hypothesis" would really be a (partial) *strategy*, helping us to generate actions. We would judge strategies on (something like) average reward achieved, not even trying to predict precise reward signals. The agent still receives incoming perceptual information, and strategies can use it to update internal states and to inform actions. However, strategies are not asked to produce any predictions. (The framework I'm describing is, of course, model-free RL.)

Intuitively, it seems as if this is missing something. A model-based agent can learn a lot about the world just by watching, taking no actions. However, individual strategies can implement prediction-based learning within themselves. So, it seems difficult to say what benefit model-based RL provides beyond model-free RL, besides a better prior over strategies.

It might be that we can't say anything recommending model-based learning over model free in a standard bounded-regret framework. (I actually haven't thought about it much -- but the argument that model-free strategies can implement models internally seems potentially strong. Perhaps you just can't get much in the AIXI framework because there are no good loss bounds in that framework at all, as Vanessa mentions.) However, if so, this seems like a weakness of standard bounded-regret frameworks. Predicting the world seems to be a significant aspect of intelligence; we should be able to talk about this formally somehow.

Granted, it doesn't make sense for bounded agents to pursue predictive accuracy above all else. There is a computational trade-off, and you don't need to predict something which isn't important. My claim is something like, you should try and predict when you don't yet have an effective strategy. After you have an effective strategy, you don't really need to generate predictions. Before that, you need to generate predictions because you're still grappling with the world, trying to understand what's basically going on.

If we're trying to understand intelligence, the idea that model-free learners can internally manage these trade-offs (by choosing strategies which judiciously choose to learn from predictions when it is efficacious to do so) seems less satisfying than a proper theory of learning from prediction. What is fundamental vs non-fundamental to intelligence can get fuzzy, but learning from prediction seems like something we expect any sufficiently intelligent agent to do (whether it was built-in or learned behavior).

On the other hand, judging hypotheses on their predictive accuracy is kind of a weird thing to do if what you ultimately want a hypothesis to do for you is generate actions. It's like this: You've got two tasks; task A, and task B. Task A is what you really care about, but it might be quite difficult to tackle on its own. Task B is really very different from task A, but you can get a lot of feedback on task B. So you ask your hypotheses to compete on task B, and judge them on that in addition to task A. Somehow you're expecting to get a lot of information about task A from performance on task B. And indeed, it seems you do: predictive accuracy of a hypothesis is somehow a useful proxy for efficacy of that hypothesis in guiding action.

(It should also be noted that a reward-learning framework presumes we get feedback about utility at all. If we get no feedback about reward, then we're forced to *only* judge hypotheses by predictions, and make what inferences about utility we will. A dire situation for learning theory, but a situation where we can still talk about rational agency more generally.)

1b: The Analogy to Logic

My argument is going to be that if achieving high reward is task A, and predicting perception is task B, logic can be task C. Like task B, it is very different from task A. Like task B, it nonetheless provides useful information. Like task B, it seems to me that a theory of (boundedly) rational agency is missing something without it.

The basic picture is this. Perceptual prediction provides a lot of good feedback about the quality of cognitive algorithms. But if you really want to train up some good cognitive algorithms for yourself, it is helpful to do some imaginative play on the side.

One way to visualize this is an agent making up math puzzles in order to strengthen its reasoning skills. This might suggest a picture where the puzzles are always well-defined (terminating) computations. However, there's no special dividing line between decidable and undecidable problems -- any particular restriction to a decidable class might rule out some interesting (decidable but non-obviously so) stuff which we could learn from. So we might end up just going with any computations (halting or no).

Similarly, we might not restrict ourselves to entirely well-defined propositions. It makes a lot of sense to test cognitive heuristics on scenarios closer to life.

Why do I think sufficiently advanced agents are likely to do this?

Well, just as it seems important that we can learn a whole lot from prediction before we ever take an action in a given type of situation, it seems important that we can learn a whole lot by reasoning before we even observe that situation. I'm not formulating a precise learning-theoretic conjecture, but intuitively, it is related to whether we could reasonably expect the agent to get something right on the first try. Good perceptual prediction alone does not guarantee that we can correctly anticipate the effects of actions we have never tried before, but if I see an agent generate an effective strategy in a situation it has never intervened in before (but has had opportunity to observe), I expect that internally it is learning from perception at some level (even if it is model-free in overall architecture). Similarly, if I see an agent quickly pick up a reasoning-heavy game like chess, then I suspect it of learning from hypothetical simulations at some level.

Again, "on the first try" is not supposed to be a formal learning-theoretic requirement; I realize you can't exactly expect anything to work on the first try with learning agents. What I'm getting at has something to do with generalization.

2: Learning-Theoretic Criteria

Part of the frame has been learning-theory-vs-logic. One might interpret my closing remarks from the previous section that way; I don't know how to formulate my intuition learning-theoretically, but I expect that reasoning helps agents in particular situations. It may be that the phenomena of the previous section cannot be understood learning-theoretically, and only amount to a "better prior over strategies" as I mentioned. However, I don't want it to be a learning-theory-vs-logic argument. I would hope that something learning-theoretic can be said in favor of learning from perception, and in favor of learning from logic. Even if it can't, learning theory is still an important component here, regardless of the importance of logic.

I'll try to say something about how I think learning theory should interface with logic.

Vanessa said some relevant things in a comment, which I'll quote in full:

> Heterodox opinion: I think the entire MIRIesque (*and* academic philosophy) approach to decision theory is confused. The basic assumption seems to be, that we can decouple the problem of learning a model of the world from the problem of taking a decision given such a model. We then ignore the first problem, and assume a particular shape for the model (for example, causal network) which allows us to consider decision theories such as CDT, EDT etc. However, in reality the two problems cannot be decoupled. This is because the *type signature* of a world model is only meaningful if it comes with an algorithm for how to learn a model of this type.
>
> For example, consider Newcomb's paradox. The agent makes a decision under the assumption that Omega behaves in a certain way. But, where did the assumption come from? Realistic agents have to *learn* everything they know. Learning normally requires a time sequence. For example, we can consider the *iterated* Newcomb's paradox (INP). In INP, any reinforcement learning (RL) algorithm will converge to one-boxing, simply because one-boxing gives it the money. This is despite RL naively looking like CDT. Why does it happen? Because in the learned

> model, the "causal" relationships are *not* physical causality. The agent comes to believe that taking the one box *causes* the money to appear there.
>
> In Newcomb's paradox EDT succeeds but CDT fails. Let's consider an example where CDT succeeds and EDT fails: the XOR blackmail. The iterated version would be IXB. In IXB, classical RL doesn't guarantee much because the environment is more complex than the agent (it contains Omega). To overcome this, we can use RL with incomplete models. I believe that this indeed solves both INP and IXB.
>
> Then we can consider e.g. counterfactual mugging. In counterfactual mugging, RL with incomplete models doesn't work. That's because the assumption that Omega responds in a way that depends on a counterfactual world is not in the space of models at all. Indeed, it's unclear how can any agent learn such a fact from empirical observations. One way to fix it is by allowing the agent to precommit. Then the assumption about Omega becomes empirically verifiable. But, if we do this, then RL with incomplete models can solve the problem again.
>
> The only class of problems that I'm genuinely unsure how to deal with is game-theoretic superrationality. However, I also don't see much evidence the MIRIesque approach has succeeded on that front. We probably need to start with just solving the grain of truth problem in the sense of converging to ordinary Nash (or similar) equilibria (which might be possible using incomplete models). Later we can consider agents that observe each other's source code, and maybe something along the lines of this can apply.

Besides the MIRI-vs-learning frame, I agree with a lot of this. I wrote a comment elsewhere making some related points about the need for a learning-theoretic approach. Some of the points also relate to my CDT=EDT sequence; I have been arguing that CDT and EDT don't behave as people broadly imagine (often not having the bad behavior which people broadly imagine). Some of those arguments were learning-theoretic while others were not, but the conclusions were similar either way.

In any case, I think the following criterion (originally mentioned to me by Jack Gallagher) makes sense:

A decision problem should be conceived as a sequence, but the algorithm deciding what to do on a particular element of the sequence should not know/care what the whole sequence is.

Asymptotic decision theory was the first major proposal to conceive of decision problems as sequences in this way. Decision-problem-as-sequence allows decision theory to be addressed learning-theoretically; we can't expect a learning agent to necessarily do well in any particular case (because it could have a sufficiently poor prior, and so still be learning in that particular case), but we *can* expect it to *eventually* perform well (provided the problem meets some "fairness" conditions which make it learnable).

As for the second part of the criterion, requiring that the agent is ignorant of the overall sequence when deciding what to do on an instance: this captures the idea of learning from logic. Providing the agent with the sequence is cheating, because you're essentially giving the agent your interpretation of the situation.

Jack mentioned this criterion to me in a discussion of averaging decision theory (AvDT), in order to explain why AvDT was cheating.

AvDT is based on a fairly simple idea: look at the average performance of a strategy so far, rather than its expected performance on this particular problem. Unfortunately, "performance so far" requires things to be defined in terms of a training sequence (counter to the logical-induction philosophy of non-sequential learning).

I created AvDT to try and address some shortcomings of asymptotic decision theory (let's call it AsDT). Specifically, AsDT does not do well in counterlogical mugging. AvDT is capable of doing well in counterfactual mugging. However, it depends on the training sequence. Counterlogical mugging requires the agent to decide on the "probability" of Omega asking for money vs paying up, to figure out whether participation is worth it overall. AvDT solves this problem by looking at the training sequence to see how often Omega pays up. So, the problem of doing well in decision problems is "reduced" to specifying good training sequences. This (1) doesn't obviously make things easier, and (2) puts the work on the human trainers.

Jack is saying that the system should be looking through logic *on its own* to find analogous scenarios to generalize from. When judging whether a system gets counterlogical mugging right, *we* have to define counterlogical mugging as a sequence to enable learning-theoretic analysis; but *the agent* has to figure things out on its own.

This is a somewhat subtle point. A realistic agent experiences the world sequentially, and learns by treating its history as a training sequence of sorts. This is *physical time*. I have no problem with this. What I'm saying is that if an agent is also learning *from analogous circumstances within logic*, as I suggested sophisticated agents will do in the first part, then Jack's condition should come into play. We aren't handed, from on high, a sequence of logically defined scenarios which we can locate ourselves within. We only have regular physical time, plus a bunch of hypothetical scenarios which we can define and whose relevance we have to determine.

This gets back to my earlier intuition about agents having a reasonable chance of getting certain things right on the first try. Learning-theoretic agents don't get things right on the first try. However, agents who learn from logic have "lots of tries" before their first real try in physical time. If you can successfully determine which logical scenarios are relevantly analogous to your own, you can learn what to do just by thinking. (Of course, you still need a lot of physical-time learning to know enough about your situation to do that!)

So, getting back to Vanessa's point in the comment I quoted: can we solve MIRI-style decision problems by considering the iterated problem, rather than the single-shot version? To a large extent, I think so: in logical time, all games are iterated games. However, I don't want to have to set an agent up with a training sequence in which it encounters those specific problems many times. For example, finding good strategies in chess via self-play should come naturally from the way the agent thinks about the world, rather than being an explicit training regime which the designer has to implement. Once the rules for chess are understood, the bottleneck should be thinking time rather than (physical) training instances.

An Orthodox Case Against Utility Functions

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

This post has benefitted from discussion with Sam Eisenstat, Scott Garrabrant, Tsvi Benson-Tilsen, Daniel Demski, Daniel Kokotajlo, and Stuart Armstrong. It started out as a thought about Stuart Armstrong's research agenda.

In this post, I hope to say something about what it means for a rational agent to have preferences. The view I am putting forward is relatively new to me, but it is not *very* radical. It is, dare I say, a conservative view -- I hold close to Bayesian expected utility theory. However, my impression is that it differs greatly from *common impressions* of Bayesian expected utility theory.

I will argue against a particular view of expected utility theory -- a view which I'll call *reductive utility*. I do not recall seeing this view explicitly laid out and defended (except in in-person conversations). However, I expect at least a good chunk of the assumptions are commonly made.

Reductive Utility

The core tenets of reductive utility are as follows:

- The sample space Ω of a rational agent's beliefs is, more or less, the set of possible ways the world could be -- which is to say, the set of possible *physical configurations of the universe*. Hence, each world $\omega \in \Omega$ is one such configuration.
- The preferences of a rational agent are represented by a utility function $U : \Omega \to \mathbb{R}$ from worlds to real numbers.
- Furthermore, the utility function should be a *computable* function of worlds.

Since I'm setting up the view which I'm knocking down, there is a risk I'm striking at a straw man. However, I think there are some good reasons to find the view appealing. The following subsections will expand on the three tenets, and attempt to provide some motivation for them.

If the three points seem obvious to you, you might just skip to the next section.

Worlds Are Basically Physical

What I mean here resembles the standard physical-reductionist view. However, my emphasis is on certain features of this view:

- There is some "basic stuff" -- like like quarks or vibrating strings or what-have-you.
- What there is to know about the world is some set of statements about this basic stuff -- particle locations and momentums, or wave-form function values, or what-have-you.
- These special atomic statements should be logically independent from each other (though they may of course be probabilistically related), and together, fully determine the world.
- These should (more or less) be what beliefs are about, such that we can (more or less) talk about beliefs in terms of the sample space $\omega \in \Omega$ as being the set of worlds understood in this way.

This is the so-called "view from nowhere", as Thomas Nagel puts it.

I don't intend to construe this position as ruling out certain non-physical facts which we may have beliefs about. For example, we may believe indexical facts on top of the physical facts -- there might be (1) beliefs about the universe, and (2) beliefs about where we are in the universe. Exceptions like this violate an extreme reductive view, but are still close enough to count as reductive thinking for my purposes.

Utility Is a Function of Worlds

So we've got the "basically physical" $\omega \in \Omega$. Now we write down a utility function $U(\omega)$. In other words, utility is a random variable on our event space.

What's the big deal?

One thing this is saying is that *preferences are a function of the world*. Specifically, *preferences need not only depend on what is observed.* This is incompatible with standard RL in a way that matters.

But, in addition to saying that utility can depend on more than just observations, we are *restricting* utility to *only* depend on things that are in the world. After we consider all the information in ω, there cannot be any extra uncertainty about utility -- no extra "moral facts" which we may be uncertain of. If there are such moral facts, they have to be present somewhere in the universe (at least, derivable from facts about the universe).

One implication of this: *if utility is about high-level entities, the utility function is responsible for deriving them from low-level stuff.* For example, if the universe is made of quarks, but utility is a function of beauty, consciousness, and such, then $U()$ needs to contain the beauty-detector and consciousness-detector and so on -- otherwise how can it compute utility given all the information about the world?

Utility Is Computable

Finally, and most critically for the discussion here, $U()$ should be a computable function.

To clarify what I mean by this: ω should have some sort of representation which allows us to feed it into a Turing machine -- let's say it's an infinite bit-string which assigns true or false to each of the "atomic sentences" which describe the world. $U()$ should be a computable function; that is, there should be a Turing machine F which takes a rational number $\epsilon > 0$ and takes ω, prints a rational number within ϵ of $U(\omega)$, and halts. (In other words, we can compute $U(\omega)$ to any desired degree of approximation.)

Why should $U()$ be computable?

One argument is that $U()$ should be computable because the agent has to be able to use it in computations. This perspective is especially appealing if you think of $U()$ as a black-box function which you can only optimize through search. If you can't evaluate $U()$, how are you supposed to use it? If $U()$ exists as an actual module somewhere in the brain, how is it supposed to be implemented? (If you don't think this sounds very convincing, great!)

Requiring $U()$ to be computable may also seem easy. What is there to lose? Are there preference structures we really care about being able to represent, which are fundamentally not computable?

And what would it even mean for a computable agent to have non-computable preferences?

However, the computability requirement is more restrictive than it may seem.

There is a sort of continuity implied by computability: $U()$ must not depend too much on "small" differences between worlds. The computation $F(\epsilon, \omega)$ only accesses finitely many bits of ω before it halts. All the rest of the bits in ω must not make more than ϵ difference to the value of $U(\omega)$.

This means some seemingly simple utility functions are not computable.

As an example, consider the procrastination paradox. Your task is to push a button. You get 10 utility for pushing the button. You can push it any time you like. However, if you never press the button, you get -10. On any day, you are fine with putting the button-pressing off for one more day. Yet, if you put it off forever, you lose!

We can think of ω as a string like 000000100.., where the "1" is the day you push the button. To compute the utility, we might look for the "1", outputting 10 if we find it.

But what about the all-zero universe, 0000000...? The program must loop forever. We can't tell we're in the all-zero universe by examining any finite number of bits. You don't know whether you will eventually push the button. (Even if the universe also gives your source code, you can't necessarily tell from that -- the logical difficulty of determining this about yourself is, of course, the original point of the procrastination paradox.)

Hence, a preference structure like this is not computable, and is not allowed according to the reductive utility doctrine.

The advocate of reductive utility might take this as a victory. The procrastination paradox has been avoided, and other paradoxes with a similar structure. (The St. Petersburg Paradox is another example.)

On the other hand, if you think this is a *legitimate preference structure*, dealing with such 'problematic' preferences motivates abandonment of reductive utility.

Subjective Utility: The Real Thing

We can strongly oppose all three points without leaving orthodox Bayesianism. Specifically, I'll sketch how the Jeffrey-Bolker axioms enable non-reductive utility. (The title of this section is a reference to Jeffrey's book *Subjective Probability: The Real Thing.*)

However, the *real* position I'm advocating is more grounded in logical induction rather than the Jeffrey-Bolker axioms; I'll sketch that version at the end.

The View From Somewhere

The reductive-utility view approached things from the starting-point of the universe. Beliefs are for what is real, and what is real is basically physical.

The non-reductive view starts from the standpoint of the agent. Beliefs are *for things you can think about*. This doesn't rule out a physicalist approach. What it *does* do is give high-level objects like tables and chairs an equal footing with low-level objects like quarks: both are inferred from sensory experience by the agent.

Rather than assuming an underlying set of *worlds*, Jeffrey-Bolker assume only a set of events. For two events P and Q, the conjunction $P \wedge Q$ exists, and the disjunction $P \vee Q$, and the negations $\neg P$ and $\neg Q$. However, unlike in the Kolmogorov axioms, these are not assumed to be intersection, union, and complement of an underlying set of worlds.

Let me emphasize that: *we need not assume there are "worlds" at all.*

In philosophy, this is called situation semantics -- an alternative to the more common possible-world semantics. In mathematics, it brings to mind pointless topology.

In the Jeffrey-Bolker treatment, a world is just a maximally specific event: an event which describes everything completely. But there is no requirement that maximally-

specific events exist. Perhaps any event, no matter how detailed, can be further extended by specifying some yet-unmentioned stuff. (Indeed, the Jeffrey-Bolker axioms assume this! Although, Jeffrey does not seem philosophically committed to that assumption, from what I have read.)

Thus, there need not be any "view from nowhere" -- no semantic vantage point from which we see the whole universe.

This, of course, deprives us of the objects which utility was a function of, in the reductive view.

Utility Is a Function of Events

The reductive-utility makes a distinction between utility -- the random variable itself -- and *expected* utility, which is the subjective estimate of the random variable which we use for making decisions.

The Jeffrey-Bolker framework does not make a distinction. Everything is a subjective preference evaluation.

A reductive-utility advocate sees the expected utility of an event $E \subseteq \Omega$ as ***derived from*** the utility of the worlds within the event. They start by defining $U(\omega)$; then, we define the expected utility of an event as $E[U|E] := \sum_\omega U(\omega)P(\omega)$ -- or, more generally, the corresponding integral.

In the Jeffrey-Bolker framework, we instead define $U(E)$ *directly* on events. These preferences are required to be *coherent with* breaking things up into sums, so $U(E) = \frac{U(F \wedge A) \cdot P(F \wedge A) + U(F \wedge \neg A) \cdot P(F \wedge \neg A)}{P(E)}$ - but we do not define one from the other.

We don't have to know how to evaluate entire worlds in order to evaluate events. All we have to know is how to evaluate events!

I find it difficult to really believe "humans have a utility function", even approximately -- but I find it *much easier* to believe "humans have expectations on propositions". Something like that could even be true at the *neural* level (although of course we would not obey the Jeffrey-Bolker axioms in our neural expectations).

Updates Are Computable

Jeffrey-Bolker doesn't say anything about computability. However, if we do want to address this sort of issue, it leaves us in a different position.

Because *subjective expectation is primary*, it is now more natural to require that the agent can evaluate events, without any requirement about a function on worlds. (Of course, we *could* do that in the Kolmogorov framework.)

Agents don't need to be able to compute the utility of a whole world. All they need to know is how to update expected utilities as they go along.

Of course, the subjective utility can't be just *any* way of updating as you go along. It needs to be ***coherent***, in the sense of the Jeffrey-Bolker axioms. And, maintaining coherence can be very difficult. But it can be quite easy even in cases where the random-variable treatment of the utility function is not computable.

Let's go back to the procrastination example. In this case, to evaluate the expected utility of each action at a given time-step, the agent does not need to figure out whether it ever pushes the button. It just needs to have some probability, which it updates over time.

For example, an agent might initially assign probability $2^{-(t+1)}$ to pressing the button at time t, and 1/2 to never pressing the button. Its probability that it would ever press the button, and thus its utility estimate, would decrease with each observed time-step in which it didn't press the button. (Of course, such an agent would press the button immediately.)

Of course, this "solution" doesn't touch on any of the tricky logical issues which the procrastination paradox was originally introduced to illustrate. This isn't meant as a solution to the procrastination paradox -- only as an illustration of how to coherently update discontinuous preferences. This simple U() is ***uncomputable*** by the definition of the previous section.

It also doesn't address computational tractability in a very real way, since if the prior is very complicated, computing the subjective expectations can get extremely difficult.

We can come closer to addressing logical issues and computational tractability by considering things in a logical induction framework.

Utility Is Not a Function

In a logical induction (LI) framework, the central idea becomes "*update your subjective expectations in any way you like, so long as those expectations aren't (too easily) exploitable to Dutch-book.*" This clarifies what it means for the updates to be "coherent" -- it is somewhat more elegant than saying "... any way you like, so long as they follow the Jeffrey-Bolker axioms."

This replaces the idea of "utility function" entirely -- there isn't any need for a *function* any more, just a logically-uncertain-variable (LUV, in the terminology from the LI paper).

Actually, there are different ways one might want to set things up. I hope to get more technical in a later post. For now, here's some bullet points:

- In the simple procrastination-paradox example, you push the button if you have any uncertainty at all. So things are not that interesting. But, at least we've solved the problem.

- In more complicated examples -- where there is some real benefit to procrastinating -- a LI-based agent could totally procrastinate forever. This is because LI doesn't give any guarantee about converging to correct beliefs for uncomputable propositions like whether Turing machines halt or whether people stop procrastinating.
- Believing you'll stop procrastinating even though you won't is *perfectly coherent* -- in the same way that believing in nonstandard numbers is perfectly logically consistent. Putting ourselves in the shoes of such an agent, this just means we've examined our own decision-making to the best of our ability, and have put significant probability on "we don't procrastinate forever". This kind of reasoning is necessarily fallible.
- Yet, if a system we built were to do this, we might have strong objections. So, this can count as an alignment problem. How can we give feedback to a system to avoid this kind of mistake? I hope to work on this question in future posts.

What does it mean to apply decision theory?

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

Based on discussions with Stuart Armstrong and Daniel Kokotajlo.

There are two conflicting ways of thinking about foundational rationality arguments such as the VNM theorem.

1. As direct arguments for normative principles. The axioms are supposed to be premises which you'd actually accept. The axioms imply theories of rationality such as probability theory and utility theory. These are supposed to apply in practice: if you accept the axioms, then you should be following them.
2. As idealized models. Eliezer compares Bayesian reasoners to a Carnot engine: an idealized, thermodynamically perfect engine which can never be built. To the extent that any real engine works, it approximates a Carnot engine. To the extent that any cognition really works, it approximates Bayes. Bayes sets the bounds for what is possible.

The second way of thinking is very useful. Philosophers, economists, and others have made some real progress thinking in this way. However, I'm going to argue that we should push for the first sort of normative principle. We should not be satisfied with normative principles which remain as unachievable ideals, giving upper bounds on performance without directly helping us get there.

This implies dealing with problems of bounded rationality. But it's not the sort of "bounded rationality" where we set out to explicitly model irrationality. We don't want to talk about partial rationality; we want notions of rationality which bounded agents can *fully* satisfy.

Approximating Rationality

In order to apply an idealized rationality, such as Bayesian superintelligence, we need to have a concept of *what it means to approximate it*. This is more subtle than it may seem. You can't necessarily try to minimize some notion of distance between your behavior and the ideal behavior. For one thing, you can't compute the ideal behavior to find the distance! But, for another thing, simple imitation of the ideal behavior can go wrong. Adopting one part of an optimal policy without adopting all the other parts might put you in a *much worse* position than the one you started in.

Wei Dai discusses the problem in a post about Hanson's pre-rationality concept:

> [...] 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 [...] rationality, but it's troubling that little attention has been paid to them, relative to defining [...] 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?)

Clearly, we have *some* idea of which moves toward rationality are correct vs incorrect. Think about the concept of cargo-culting: pointless and ineffective imitation of a more capable agent. The problem is the absence of a formal theory.

Examples

One *possible* way of framing the problem: the VNM axioms, the Kolmogorov probability axioms, and/or other rationality frameworks give us a **notion of consistency**. We can check our behaviors and opinions for inconsistency. But what do we do when we *notice* an inconsistency? Which parts are we supposed to change?

Here are some cases where there is at least a *tendency* to update in a particular direction:

- Suppose we value an event E at 4.2 expected utils. We then unpack E into two mutually exclusive sub-events, $E_1 \cup E_2 = E$. We notice that we value E_1 at 1.1 utils and E_2 at 3.4 utils. This is inconsistent with the evaluation of E. We usually trust E less than the unpacked version, and would reset the evaluation of E to $P(E_1) \cdot 1.1 + P(E_2) \cdot 3.4$.
- Suppose we notice that we're doing things in a way that's not optimal for our goals. That is, we notice some new way of doing things which is better for what we believe our goals to be. We will tend to change our behavior rather than change our beliefs about what our goals are. (Obviously this is not always the case, however.)
- Similarly, suppose we notice that we are acting in a way which is inconsistent with our beliefs. There is a tendency to correct the action rather than the belief. (Again, not as surely as my first example, though.)
- If we find that a belief was subject to base-rate neglect, there is a tendency to multiply by base-rates and renormalize, rather than adjust our beliefs about base rates to make them consistent.
- If we notice that X and Y are equivalent, but we had different beliefs about X and Y, then we tend to pool information from X and Y such that, for example, if we had a very sharp distribution about X and a very uninformative distribution about Y, the sharp distribution would win.

If you're like me, you might have read some of those and immediately thought of a Bayesian model of the inference going on. Keep in mind that this is *supposed* to be about noticing *actual inconsistencies*, and what we want is a model which deals directly with that. It might turn out to be a kind of meta-Bayesian model, where we approximate a Bayesian superintelligence by way of a much more bounded Bayesian view which attempts to reason about what a truly consistent view would look like. But don't fool yourself into thinking a standard one-level Bayesian picture is sufficient, just

because you can look at some of the bullet points and imagine a Bayesian way to handle it.

It would be quite interesting to have a general "theory of becoming rational" which had something to say about how we make decisions in cases such as I've listed.

Logical Uncertainty

Obviously, I'm pointing in the general direction of logical uncertainty and bounded notions of rationality (IE notions of rationality which can apply to bounded agents). Particularly in the "noticing inconsistencies" framing, it sounds like this might *entirely* reduce to logical uncertainty. But I want to point at the broader problem, because (1) an example of this might not immediately look like a problem of logical uncertainty; (2) a theory of logical uncertainty, such as logical induction, might not entirely solve this problem; (3) logical uncertainty is an epistemic issue, whereas this problem applies to instrumental rationality as well; (4) even setting all that aside, it's worth pointing at the distinction between ideal notions of rationality and applicable notions of rationality as a point in itself.

The Ideal Fades into the Background

So far, it sounds like my suggestion is that we should keep our idealized notions of rationality, but also develop "theories of approximation" which tell us what it means to approach the ideals in a good way vs a bad way. However, I want to point out an interesting phenomenon: sometimes, when you get a really good notion of "approximation", the idealized notion of rationality you started with fades into the background.

Example 1: Logical Induction

Start with the Demski Prior, which was supposed to be an idealized notion of rational belief much like the Solomonoff prior, but built for logic rather than computation. I designed the prior with approximability in mind, because I thought it should be a constraint on a normative theory that we actually be able to approximate the ideal. Scott and Benya modified the Demski prior to make it nicer, and noticed that when you do so, the approximation itself has a desirable property. The line of research called asymptotic logical uncertainty focused on such "good properties of approximations", eventually leading to logical induction.

A logical inductor is a sequence of improving belief assignments. The beliefs do converge to a probability distribution, which will have some resemblance to the modified Demski prior (and to Solomonoff's prior). However, the concept of logical induction gives a much richer theory of rationality, in which this limit plays a minor role. Furthermore, the theory of logical induction comes much closer to applying to realistic agents than "rational agents approximate a Bayesian reasoning with [some prior]".

Example 2: Game-Theoretic Equilibria vs MAL

Game-theoretic equilibrium concepts, such as Nash equilibrium and correlated equilibrium, provide a rationality concept for games: rational agents who know that each other are rational are supposed to be in equilibrium with each other. However, most games have multiple Nash equilibria, and even more correlated equilibria. How is a rational agent supposed to decide which of these to play? Assuming only the rationality of the other players is not enough to choose one equilibrium over another. If rational agents play an equilibrium, how do they get there?

One approach to this conundrum has been to introduce refined equilibrium concepts, which admit some Nash equilibria and not others. Trembling Hand equilibrium is one such concept. This introduces a notion of "stable" equilibria, pointing out that it is implausible that agents play "unstable" equilibria. However, while this narrows things down to a single equilibrium solution in some cases, it does not do so for all cases. Other refined equilibrium concepts may leave no equilibria for some games. To get rid of the problem, one would need an equilibrium concept which (a) leaves one and only one equilibrium for every game, and (b) follows from plausible rationality assumptions. Such things have been proposed, most prominently by Harsanyi & Selten A General Theory of Equilibrium Selection in Games, but so far I find them unconvincing.

A very different approach is represented by multi-agent learning (MAL), which asks the question: can agents learn to play equilibrium strategies? In this version, agents must interact over time in order to converge to equilibrium play. (Or at least, agents simulate dumber versions of each other in an effort to figure out how to play.)

It turns out that, in MAL, there are somewhat nicer stories about how agents converge to *correlated* equilibria than there are about converging to Nash equilibria. For example, Calibrated Learning and Correlated Equilibrium (Foster & Vohra) shows that agents with a calibrated learning property will converge to correlated equilibrium in repeated play.

These new rationality principles, which come from MAL, are then much more relevant to the design and implementation of game-playing agents than the equilibrium concepts which they support. Equilibrium concepts, such as correlated equilibria, tell you something about what agents converge to in the limit; the learning principles which let them accomplish that, however, tell you about the *dynamics* -- what agents do at finite times, in response to non-equilibrium situations. This is more relevant to agents "on the ground", as it were.

And, to the extent that requirements like calibrated learning are NOT computationally feasible, this *weakens our trust in equilibrium concepts as a rationality notion* -- if there isn't a plausible story about how (bounded-) rational agents can get into equilibrium, why should we think of equilibrium as rational?

So, we see that the bounded, dynamic notions of rationality are more fundamental than the unbounded, fixed-point style equilibrium concepts: if we want to deal with realistic agents, we should be more willing to adjust/abandon our equilibrium concepts in response to how nice the MAL story is, than vice versa.

Counterexample: Complete Class Theorems

This doesn't always happen. The complete class theorems give a picture of rationality in which we *start* with the ability and willingness to take Pareto-improvements. Given

this, we *end up* with an agent being classically rational: having a probability distribution, and choosing actions which maximize expected utility.

Given this argument, we become more confident in the usefulness of probability distributions. But why should this be the conclusion? A different way of looking at the argument could be: we don't need to think about probability distributions. All we need to think about is Pareto improvements.

Somehow, probability still seems very useful to think about. We don't switch to the "dynamic" view of agents who haven't yet constructed probabilistic beliefs, taking Pareto improvements on their way to reflective consistency. This just doesn't seem like a realistic view of bounded agents. **Yes,** bounded agents are still engaged in a search for the best policy, which may involve finding new strategies which are strictly better along every relevant dimension. But bounded agency **also** involves making trade-offs, when no Pareto improvement can be found. This necessitates thinking of probabilities. So it doesn't seem like we want to erase that from our picture of practical agency.

Perhaps this is because, in some sense, the complete class theorems are not very good -- they don't really end up explaining a less basic thing in terms of a more basic thing. After all, when can you realistically find a pure Pareto improvement?

Conclusion

I've suggested that we move toward notions of rationality that are fundamentally bounded (applying to agents who lack the resources to be rational in more classical senses) and dynamic (fundamentally involving learning, rather than assuming the agent already has a good picture of the world; breaking down equilibrium concepts such as those in game theory, and instead looking for the dynamics which can converge to equilibrium).

This gives us a picture of "rationality" which is more like "optimality" in computer science: in computer science, it's more typical to come up with a notion of optimality which *actually applies* to some algorithms. For example, "optimal sorting algorithm" usually refers to big-O optimality, and many sorting algorithms are optimal in that sense. Similarly, in machine learning, regret bounds are mainly interesting when they are achievable by some algorithm. (Although, it could be interesting to know a lower bound on achievable regret guarantees.)

Why should notions of rationality be so far from notions of optimality? Can we take a more computer-science flavored approach to rationality?

Barring that, it should at least be of critical importance to investigate in what sense idealized notions of rationality are normative principles for bounded agents like us. What constitutes cargo-culting rationality, vs really becoming more rational? What kind of adjustments should an irrational agent make when irrationality is noticed?

Radical Probabilism [Transcript]

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

(Talk given on Sunday 21st June, over a zoom call with 40 attendees. Abram Demski is responsible for the talk, Ben Pace is responsible for the transcription)

Talk

Abram Demski: I want to talk about this idea that, for me, is an update from the logical induction result that came out of MIRI a while ago. I feel like it's an update that I wish the entire LessWrong community had gotten from logical induction but it wasn't communicated that well, or it's a subtle point or something.

Abram Demski: But hopefully, this talk isn't going to require any knowledge of logical induction from you guys. I'm actually going to talk about it in terms of philosophers who had a very similar update starting around, I think, the '80s.

Abram Demski: There's this philosophy called 'radical probabilism' which is more or less the same insight that you can get from thinking about logical induction. Radical probabilism is spearheaded by this guy Richard Jeffrey who I also like separately for the Jeffrey-Bolker axioms which I've written about on LessWrong.

Abram Demski: But, after the Jeffrey-Bolker axioms he was like, well, we need to revise Bayesianism even more radically than that. Specifically he zeroed in on the consequences of Dutch book arguments. So, the Dutch book arguments which are for the Kolmogorov axioms, or alternatively the Jeffrey-Bolker axioms, are pretty solid. However, you may not immediately realize that this does not imply that Bayes' rule should be an update rule.

Abram Demski: You have Bayes' rule as a fact about your static probabilities, that's fine. As a fact about conditional probabilities, Bayes' rule is just as solid as all the other probability rules. But for some reason, Bayesians take it that you start with these probabilities, you make an observation, and then you have now these probabilities. These probabilities should be updated by Bayes' rule. And the argument for that is not super solid.

Abram Demski: There are two important flaws with the argument which I want to highlight. There is a Dutch book argument for using Bayes' rule to update your probabilities, but it makes two critical assumptions which Jeffrey wants to relax. Assumption one is that updates are always and precisely accounted for by propositions which you learn, and everything that you learn and moves your probabilities is accounted for in this proposition. These are usually thought of as sensory data. Jeffrey said, wait a minute, my sensory data isn't so certain. When I see something, we don't have perfect introspective access to even just our visual field. It's not like we get a pixel array and know exactly how everything is. So, I want to treat the things that I'm updating on as, themselves, uncertain.

Abram Demski: Difficulty two with the Dutch book argument for Bayes' rule as an update rule, is that it assumes you know already how you would update,

hypothetically, given different propositions you might observe. Then, given that assumption, you can get this argument that you need to use Bayes' rule. Because I can Dutch-book you based on my knowledge of how you're going to update. But if I don't know how you're updating, if your update has some random element, subjectively random, if I can't predict it, then we get this radical treatment of how you're updating. We get this picture where you believe things one day and then you can just believe different things the next day. And there's no Dutch book I can make to say you're irrational for doing that. "I've thought about it more and I've changed my mind."

Abram Demski: This is very important for logical uncertainty (which Jeffrey didn't realize because he wasn't thinking about logical uncertainty). That's why we came up with this philosophy, thinking about logical uncertainty. But Jeffrey came up with it just by thinking about the foundations and what we can argue a rational agent must be.

Abram Demski: So, that's the update I want to convey. I want to convey that Bayes' rule is not the only way that a rational agent can update. You have this great freedom of how you update.

Q&A

Ben Pace: Thank you very much, Abram. You timed yourself excellently.

Ben Pace: As I understand it, you need to have inexploitability in your belief updates and so on, such that people cannot reliably Dutch book you?

Abram Demski: Yeah. I say radical freedom meaning, if you have belief X one day and you have beliefs Y the next day, any pair of X and Y are justifiable, or potentially rational (as long as you don't take something that has probability zero and now give it positive probability or something like that).

Abram Demski: There are rationality constraints. It's not that you can do anything at all. The most concrete example of this is that you can't change your mind back and forth forever on any one proposition, because then I can money-pump you. Because I know, eventually, your beliefs are going to drift up, which means I can buy low and eventually your beliefs will drift up and then I can sell the bet back to you because now you're like, "That's a bad bet," and then I've made money off of you.

Abram Demski: If I can predict anything about how your beliefs are going to drift, then you're in trouble. I can make money off of you by buying low and selling high. In particular that means you can't oscillate forever, you have to eventually converge. And there's lots of other implications.

Abram Demski: But I can't summarize this in any nice rule is the thing. There's just a bunch of rationality constraints that come from non-Dutch-book-ability. But there's no nice summary of it. There's just a bunch of constraints.

Ben Pace: I'm somewhat surprised and shocked. So, I shouldn't be able to be exploited in any obvious way, but this doesn't constrain me to the level of Bayes' rule. It doesn't constrain me to clearly knowing how my updates will be affected by future evidence.

Abram Demski: Right. If you do know your updates, then you're constrained. He calls that the rigidity condition. And even that doesn't imply Bayes' rule, because of the

first problem that I mentioned. So, if you do know how you're going to update, then you don't want to change your conditional probabilities as a result of observing something, but you can still have these uncertain observations where you move a probability but only partially. And this is called a Jeffrey update.

Ben Pace: Phil Hazelden has a question. Phil, do you want to ask your question?

Phil Hazelden: Yeah. So, you said if you don't know how you'd update on an observation, then you get pure constraints on your belief update. I'm wondering, if someone else knows how you'd update on an observation but you don't, does that for example, give them the power to extract money from you?

Abram Demski: Yeah, so if somebody else knows, then they can extract money if you're not at least doing a Jeffrey update. In general, if a bookie knows something that you don't, then a bookie can extract money from you by making bets. So this is not a proper Dutch book argument, because what we mean by a Dutch book argument is that a totally ignorant bookie can extract money.

Phil Hazelden: Thank you.

Ben Pace: I would have expected that if I was constrained to not be exploitable then this would have resulted in Bayes' rule, but you're saying all it actually means is there are some very basic arguments about how you shouldn't be exploited but otherwise you can move very freely between. You can update upwards on Monday, down on Tuesday, down again on Wednesday, up on Thursday and then stay there and as long as I can't predict it in advance, you get to do whatever the hell you like with your beliefs.

Abram Demski: Yep, and that's rational in the sense that I think rational should mean.

Ben Pace: I do sometimes use Bayes' rule in arguments. In fact, I've done it not-irregularly. Do you expect, if I fully propagate this argument I will stop using Bayes' rule in arguments? I feel it's very helpful for me to be able to say, all right, I was believing X on Monday and not-X on Wednesday, and let me show you the shape of my update that I made using certain probabilistic updates.

Abram Demski: Yeah, so I think that if you propagate this update you'll notice cases where your shift simply cannot be accounted for as Bayes' rule. But, this rigidity condition, the condition of "I already know how I would update hypothetically on various pieces of information", the way Jeffrey talks about this (or at least the way some Jeffrey-interpreters talk about this), it's like: if you have considered this question ahead of time, of how you would update on this particular piece of information, then your update had better be either a Bayes' update or at least a Jeffrey update. In the cases where you think about it, it has this narrowing effect where you do indeed have to be looking more like Bayes.

Abram Demski: As an example of something that's non-Bayesian that you might become more comfortable with if you fully propagate this: you can notice that something is amiss with your model because the evidence is less probable than you would have expected, without having an alternative that you're updating towards. You update down your model without updating it down because of normalization constraints of updating something else up. "I'm less confident in this model now." And somebody asks what Bayesian update did you do, and I'm like "No, it's not a Bayesian update, it's just that this model seems shakier.".

Ben Pace: It’s like the thing where I have four possible hypotheses here, X, Y, Z, and “I do not have a good hypothesis here yet”. And sometimes I just move probability into “the hypothesis is not yet in my space of considerations”.

Abram Demski: But it's like, how do you do that if “I don't have a good hypothesis” doesn't make any predictions?

Ben Pace: Interesting. Thanks, Abram.

Radical Probabilism

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

This is an expanded version of my talk . I assume a high degree of familiarity with Bayesian probability theory.

Toward a New Technical Explanation of Technical Explanation -- an attempt to convey the practical implications of logical induction -- was one of my most-appreciated posts, but I don't really get the feeling that very many people have received the update. Granted, that post was speculative, sketching what a new technical explanation of technical explanation *might* look like. I think I can do a bit better now.

If the implied project of that post had really been completed, I would expect new practical probabilistic reasoning tools, explicitly violating Bayes' law. For example, we might expect:

- A new version of information theory.
    - An update to the "prediction=compression" maxim, either repairing it to incorporate the new cases, or explicitly denying it and providing a good intuitive account of why it was wrong.
    - A new account of concepts such as mutual information, allowing for the fact that variables have behavior over thinking time; for example, variables may initially be very correlated, but lose correlation as our picture of each variable becomes more detailed.
- New ways of thinking about epistemology.
    - One thing that my post did manage to do was to spell out the importance of "making advanced predictions", a facet of epistemology which Bayesian thinking does not do justice to.
    - However, I left aspects of the problem of old evidence open, rather than giving a complete way to think about it.
- New probabilistic structures.
    - Bayesian Networks are one really nice way to capture the structure of probability distributions, making them much easier to reason about. Is there anything similar for the new, wider space of probabilistic reasoning which has been opened up?

Unfortunately, I still don't have any of those things to offer. The aim of this post is more humble. I think what I originally wrote was too ambitious for didactic purposes. Where the previous post aimed to communicate the insights of logical induction by sketching broad implications, I here aim to communicate the insights *in themselves*, focusing on the detailed differences between classical Bayesian reasoning and the new space of ways to reason.

Rather than talking about logical induction directly, I'm mainly going to explain things in terms of a very similar philosophy which Richard Jeffrey invented -- apparently starting with his phd dissertation in the 50s, although I'm unable to get my hands on it or other early references to see how fleshed-out the view was at that point. He called this philosophy **radical probabilism**. Unlike logical induction, radical probabilism appears not to have any roots in worries about logical uncertainty or bounded rationality. Instead it appears to be motivated simply by a desire to generalize, and a refusal to accept unjustified assumptions. Nonetheless, it carries most of the same insights.

Radical Probabilism has not been very concerned with computational issues, and so constructing an actual algorithm (like the logical induction algorithm) has not been a focus. (However, there have been some developments -- see historical notes at the end.) This could be seen as a weakness. However, for the purpose of communicating the core insights, I think this is a strength -- there are fewer technical details to communicate.

A terminological note: I will use "radical probabilism" to refer to the new theory of rationality (treating logical induction as merely a specific way to flesh out Jeffrey's theory). I'm more conflicted about how to refer to the older theory. I'm tempted to just use the term "Bayesian", implying that the new theory is non-Bayesian -- this highlights its rejection of Bayesian updates. However, radical probabilism is Bayesian in the most important sense. Bayesianism is not about Bayes' Law. Bayesianism is, at core, about the subjectivist interpretation of probability. Radical probabilism is, if anything, *much more* subjectivist.

However, this choice of terminology makes for a confusion which readers (and myself) will have to carefully avoid: confusion between Bayesian probability theory and Bayesian updates. The way I'm using the term, a Bayesian need not endorse Bayesian updates.

In any case, I'll default to Jeffrey's term for the opposing viewpoint: ***dogmatic probabilism***. (I will occasionally fall into calling it "classical Bayesianism" or similar.)

What Is Dogmatic Probabilism?

Dogmatic Probabilism is the doctrine that the conditional probability $P(B|A)$is *also* how we *update* probabilistic beliefs: any rational change in beliefs should be explained by a Bayesian update.

We can unpack this a little:

1. (**Dynamic Belief:**) A rational agent is understood to have different beliefs over time -- call these $P_1, P_2, P_3, \ldots$
2. (**Static Rationality:**) At any one time, a rational agent's beliefs P_n are probabilistically coherent (obey the Kolmogorov axioms, or a similar axiomatization of probability theory).
3. (**Empiricism:**) Reasons for changing beliefs *across* time are *given entirely by observations* -- that is, propositions which the agent learns.
4. (**Dogmatism of Perception:**) Observations are believed with probability one, once learned.
5. (**Rigidity:**) Upon observing a proposition A, conditional probabilities $P(B|A)$ are unmodified.

The assumptions minus *empiricism* imply that an update on observing A is a Bayesian update: if we start with P_n and update on A to get P_{n+1}, then $P_{n+1}(A)$ must equal 1, and $P_{n+1}(B|A) = P_n(B|A)$. So we must have $P_{n+1}(B) = P_n(B|A)$. Then, *empiricism* says that this is the *only* kind of update we can possibly have.

What Is Radical Probabilism?

Radical probabilism accepts assumptions #1 and #2, but rejects the rest. (Logical Induction need not follow axiom #2, either, since beliefs at any given time only approximately follow the probability laws -- however, it's not necessary to discuss this complication here. Jeffrey's philosophy did not attempt to tackle such things.)

Jeffrey seemed uncomfortable with updating to 100% on anything, making *dogmatism of perception* untenable. A similar view is already popular on LessWrong, but it seems that no

one here took the implication and denied Bayesian updates as a result. (Bayesian updates have been questioned for other reasons, of course.) This is a bit of an embarassment. But fans of Bayesian updates reading this are more likely to accept that zero and one are probabilities, rather than give up Bayes.

Fortunately, this isn't actually the crux. Radical probabilism is a pure generalization of orthodox Bayesianism; you can have zero and one as probabilities, and still be a radical probabilist. The real fun begins not with the rejection of *dogmatism of perception*, but with the rejection of *rigidity* and *empiricism*.

This gives us a view in which a rational update from P_n to P_{n+1} can be almost anything. (You still can't update from $P_n(A) = 0$ to $P_{n+1}(A) > 0$.) Simply put, *you are allowed to change your mind*. This doesn't make you irrational.

Yet, there are still *some* rationality constraints. In fact, we can say a lot about how rational agents think in this model. In place of assumptions #3-#5, we assume *rational agents cannot be Dutch Booked*.

Radical Probabilism and Dutch Books

Rejecting the Dutch Book for Bayesian Updates

At this point, if you're familiar with the philosophy of probability theory, you might be thinking: wait a minute, isn't there a Dutch Book argument for Bayesian updates? If radical probabilism accepts the validity of Dutch Book arguments, shouldn't it thereby be forced into Bayesian updates?

No!

As it turns out, there is a major flaw in the Dutch Book for Bayesian updates. The argument *assumes that the bookie knows how the agent will update*. (I encourage the interested reader to read the SEP section on diachronic Dutch Book arguments for details.) Normally, a Dutch Book argument requires the bookie to be *ignorant*. It's no surprise if a bookie can take our lunch money by getting us to agree to bets *when the bookie knows something we don't know*. So what's actually established by these arguments is: ***if you know how you're going to update, then your update had better be Bayesian.***

Actually, that's not quite right: the argument for Bayesian updates also still assumes *dogmatism of perception*. If we relax that assumption, all we can really argue for is *rigidity*: ***if you know how you are going to update, then your update had better be rigid.***

This leads to a generalized update rule, called **Jeffrey updating** (or Jeffrey conditioning).

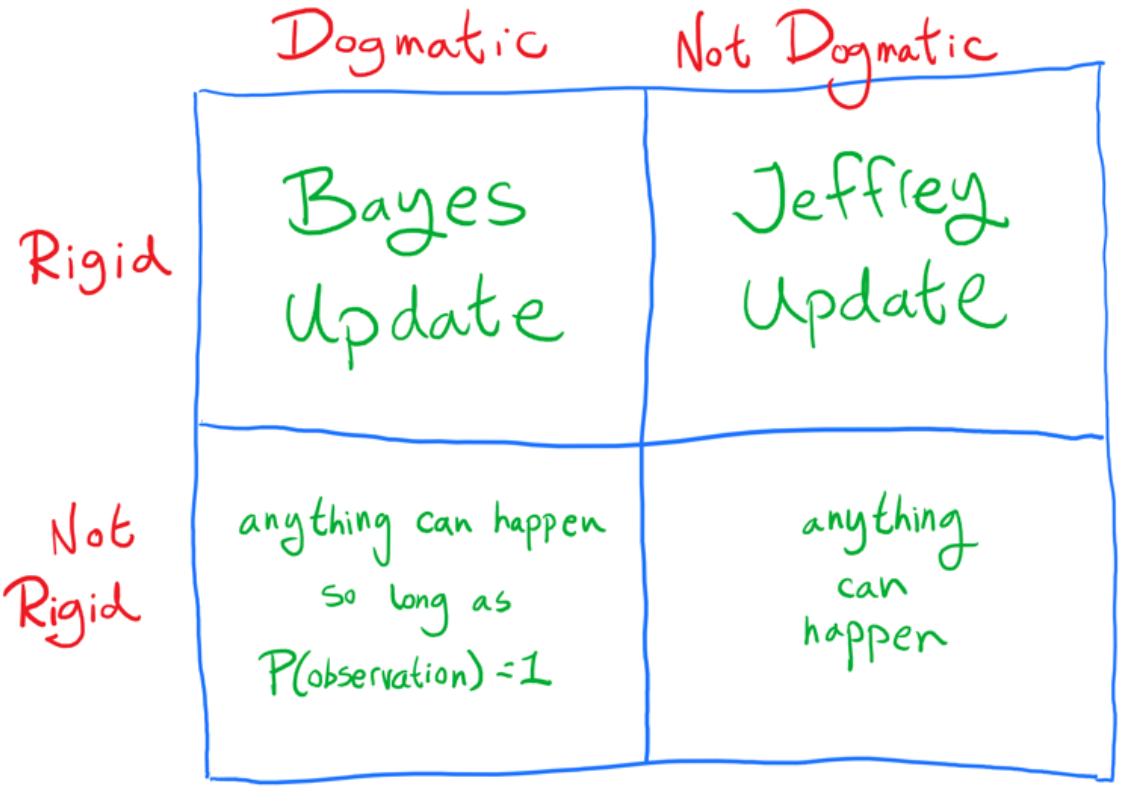
Generalized Updates

Jeffrey updates keep the rigidity assumption, but reject *dogmatism of perception*. So, we're changing the probability of some sentence A to P(A) = c, without changing any P(B|A). There's only one way to do this:

$$P_{n+1}(B) = c \cdot P_n(B \mid A) + (1 - c) \cdot P_n(B \mid \neg A)$$

In other words, the Jeffrey update interpolates linearly between the Bayesian update on A and the Bayesian update on ¬A. This generalizes Bayesian updates to allow for uncertain evidence: we're not sure we just saw someone duck behind the corner, but we're 40% sure.

If this way of updating seems a bit arbitrary to you, Jeffrey would agree. It offers only a small generalization of Bayes. Jeffrey wants to open up much broader space:



Classifying updates by assumptions made.

As I've already said, the rigidity assumption can only be justified *if the agent knows how it will update*. Philosophers like to say the agent *has a plan* for updating: "If I saw a UFO land in my yard and little green men come out, I would believe I was hallucinating." This is something we've worked out ahead of time.

A non-rigid update, on the other hand, means you don't know how you'd react: "If I saw a convincing proof of P=NP, I wouldn't know what to think. I'd have to consider it carefully." I'll call non-rigid updates ***fluid updates***.

For me, fluid updates are primarily about *having longer to think, and reaching better conclusions as a result.* That's because *my* main motivation for accepting a radical-probabilist view is logical uncertainty. Without such a motivation, I can't really imagine being very interested. I boggle at the fact that Jeffrey arrived at this view without such a motivation.

Dogmatic Probabilist: *All I can say is: why??*

Richard Jeffrey: *I've explained to you how the Dutch Book for Bayesian updates fails. What more do you want? My view is simply what you get when you remove the faulty assumptions and keep the rest.*

Dogmatic Probabilist (DP): *I understand that, but why should anyone be interested in this theory? OK, sure, I CAN make Jeffrey updates without getting Dutch Booked. But why ever would I? If I see a cloth in dim lighting, and update to 80% confident the cloth is red, I update in that way* ***because of the evidence which I've seen, which is itself fully confident****. How could it be any other way?*

Richard Jeffrey (RJ): *Tell me one peice of information you're absolutely certain of in such a situation.*

DP: *I'm certain I had that experience, of looking at the cloth.*

RJ: *Surely you aren't 100% sure you were looking at cloth. It's merely very probable.*

DP: *Fine then. The experience of looking at ... what I was looking at.*

RJ: *I'll grant you that tautologies have probability one.*

DP: *It's not a tautology... it's the fact that I had an experience, rather than none!*

RJ: *OK, but you are trying to defend the position that* ***there is some observation, which you condition on, which explains your 80% confidence the cloth is red.*** *Conditioning on "I had an experience, rather than none" won't do that. What proposition are you confident in, which explains your less-confident updates?*

DP: *The photons hitting my retinas, which I directly experience.*

RJ: *Surely not. You don't have any detailed knowledge of that.*

DP: *OK, fine, the individual rods and cones.*

RJ: *I doubt that. Within the retina, before any message gets sent to the brain, these get put through an opponent process which sharpens the contrast and colors. You're not perceiving rods and cones directly, but rather a probabilistic guess at light conditions based on rod and cone activation.*

DP: *The output of that process, then.*

RJ: *Again I doubt it. You're engaging in* ***inner-outer hocus pocus.**** *There is no clean dividing line before which a signal is external, and after which that signal has been "observed". The optic nerve is a noisy channel, warping the signal. And the output of the optic nerve itself gets processed at V1, so the rest of your visual processing doesn't get direct access to it, but rather a processed version of the information. And all this processing is noisy. Nowhere is anything certain. Everything is a guess. If, anywhere in the brain, there were a sharp 100% observation, then the nerves carrying that signal to other parts of the brain would rapidly turn it into a 99% observation, or a 90% observation...*

DP: *I begin to suspect you are trying to describe human fallibility rather than ideal rationality.*

RJ: *Not so! I'm describing how to rationally deal with uncertain observations. The source of this uncertainty could be anything. I'm merely giving human examples to establish that the theory has practical interest for humans. The theory itself only throws out unnecessary assumptions from the usual theory of rationality* **--** *as we've already discussed.*

DP: *(sigh...) OK. I'm still never going to design an artificial intelligence to have uncertain observations. It just doesn't seem like something you do on purpose. But let's grant, provisionally, that rational agents could do so and still be called rational.*

RJ: *Great.*

DP: *So what's this about giving up rigidity??*

RJ: *It's the same story: it's just another assumption we don't need.*

DP: *Right, but then how do we update?*

RJ: *However we want.*

DP: *Right, but how? I want a constructive story for where my updates come from.*

RJ: *Well, if you precommit to update in a predictable fashion, you'll be Dutch-Bookable unless it's a rigid fashion.*

DP: *So you admit it! Updates need to be rigid!*

RJ: *By no means!*

DP: *But updates need to come from somewhere. Whether you know it or not, there's some mechanism in your brain which produces the updates.*

RJ: *Whether you know it or not is a critical factor. Updates you can't anticipate need not be Bayesian.*

DP: *Right, but... the point of epistemology is to give guidance about forming rational beliefs. So you should provide some formula for updating. But any formula is predictable. So a formula has to satisfy the rigidity condition. So it's got to be a Bayesian update, or at least a Jeffrey update. Right?*

RJ: *I see the confusion. But epistemology does not have to reduce things to a strict formula in order to provide useful advice. Radical probabilism can still say many useful things. Indeed, I think it's **more** useful, since it's closer to real human experience. Humans can't always account for **why** they change their minds. They've updated, but they can't give any account of where it came from.*

DP: *OK... but... I'm sure as hell never designing an artificial intelligence that way.*

I hope you see what I mean. It's all terribly uninteresting to a typical Bayesian, especially with the design of artificial agents in mind. Why consider uncertainty about evidence? Why study updates which don't obey any concrete update rules? What would it even *mean* for an artificial intelligence to be designed with such updates?

In the light of logical uncertainty, however, it all becomes well-motivated. Updates are unpredictable not because there's no rule behind them -- nor because we lack knowledge of what exactly that rule is -- but because we can't always anticipate the results of computations before we finish running them. There are updates without corresponding evidence because we can think longer to reach better conclusions, and doing so does not reduce to Bayesian conditioning on the output of some computation. This doesn't imply uncertain evidence in exactly Jeffrey's sense, but it does give us cases where we update specific propositions to confidence levels other than 100%, and want to know how to move other beliefs in response. For example, we might apply a heuristic to determine that some number is very very likely to be prime, and update on this information.

Still, I'm very impressed with Jeffrey for reaching so many of the right conclusions without this motivation.

Other Rationality Properties

So far, I've emphasized that fluid updates "can be almost anything". This makes it sound as if there are essentially no rationality constraints at all! However, this is far from true. We can establish some very important properties via Dutch Book.

Convergence

No *single* update can be condemned as irrational. However, if you keep changing your mind again and again without ever settling down, *that* is irrational. Rational beliefs are required to eventually move less and less, converging to a single value.

Proof: If there exists a point p which your beliefs forever oscillate around (that is, your belief falls above $p + c$ infinitely often, and falls below $p - c$ infinitely often, for some $c > 0$) then a bookie can make money off of you as follows: when your belief is below $p - c$, the bookie makes a bet in favor of the proposition in question, at $p : (1 - p)$ odds. When your belief is above $p + c$, the bookie offers to cancel that bet for a small fee. The bookie earns the fee with certainty, since your beliefs are sure to swing down eventually (allowing the bet to be placed) and are sure to swing up some time after that (allowing the fee to be collected). What's more, the bookie can do this again and again and again, turning you into a money pump.

If there exists no such p, then your beliefs must converge to some value. □

Caveat: this is the proof in the context of logical induction. There are other ways to establish convergence in other formalizations of radical probabilism.

In any case, this is *really important.* This isn't just a nice rationality property. *It's a nice rationality property which dogmatic probilists don't have.* Lack of a convergence guarantee is **one of the main criticisms Frequentists make of Bayesian updates.** And it's a good critique!

Consider a simple coin-tossing scenario, in which we have two hypotheses: $h_{\frac{1}{3}}$ posits that the probability of heads is $\frac{1}{3}$, and $h_{\frac{2}{3}}$ posits that the probability of heads is $\frac{2}{3}$. The prior places probability $\frac{1}{2}$ on both of these hypotheses. The only problem is that the true coin probability is $\frac{1}{2}$. What happens? The probabilities $P(h_{\frac{1}{3}})$ and $P(h_{\frac{2}{3}})$ will oscillate forever without converging.

Proof: The quantity heads − tails will take a random walk as we keep flipping the fair coin. A random walk returns to zero infinitely often (a phenomenon known as gambler's ruin). At each such point, evidence is evenly balanced between the two hypotheses, so we've returned to the prior. Then, the next flip is either heads or tails. This results in a probability of $\frac{1}{3}$ for one of the hypotheses, and $\frac{2}{3}$ for the other. This sequence of events happens infinitely often, so $P(h_{\frac{1}{3}})$ and $P(h_{\frac{2}{3}})$ keep experiencing changes of size at least $\frac{1}{6}$, never settling down. □

Now, the objection to Bayesian updates here isn't *just* that oscillating forever looks irrational. Bayesian updates are *supposed to help us predict the data well*; in particular, you might think they're *supposed to help us minimize log-loss*. But here, we would be doing much better if beliefs would converge toward $P(h_{\frac{1}{3}}) = P(h_{\frac{2}{3}}) = \frac{1}{2}$. The problem is, Bayes takes each new bit of evidence *just as seriously* as the last. Really, though, a rational agent in this situation should be saying: "Ugh, this again! If I send my probability up, it'll come crashing

right back down some time later. I should skip all the hassle and keep my probability close to where it is."

In other words, a rational agent should be looking out for Dutch Books against itself, including the non-convergence Dutch Book. Its probabilities should be adjusted to avoid such Dutch Books.

DP: *Why should I be bothered by this example? If my prior is as you describe it, I assign* ***literally zero probability*** *to the world you describe -- I* ***know*** *the coin isn't fair. I'm fine with my inference procedure displaying pathological behavior in a universe I'm absolutely confident I'm not in.*

RJ: *So you're fine with an inference procedure which performs abysmally in the real world?*

DP: *What? Of course not.*

RJ: *But the real world cannot possibly be in your hypothesis space. It's too big. You can't explicitly write it down.*

DP: *Physicists seem to be making good progress.*

RJ: *Sure, but those aren't hypotheses which you can directly use to anticipate your experiences. They require too much computation. Anything that can fit in your head, can't be the real world.*

DP: *You're dealing with human frailty again.*

RJ: *On the contrary. Even idealized agents can't fit inside a universe they can perfectly predict. To see the contradiction, just let two of them play rock-paper-scissors with each other. Anything that can anticipate what you expect, and then do something else, can't be in your hypothesis space. But let me try a different angle of attack. Bayesianism is supposed to be the philosophy of subjective probability. Here, you're arguing as if the prior represented an objective fact about how the universe is. It isn't, and can't be.*

DP: *I'll deal with both of those points at once. I don't really need to assume that the* ***actual universe*** *is within my hypothesis space. Constructing a prior over a set of hypotheses guarantees you this:* ***if there is a best element in that class, you will converge to it.*** *In the coin-flip example, I don't have the objective universe in my set of hypotheses unless I can perfectly predict every coin-flip. But the subjective hypothesis which treats the coin as fair is the best of its kind. In the rock-paper-scissors example, rational players would similarly converge toward treating each other's moves as random, with* $\frac{1}{3}$ *probability on each move.*

RJ: *Good. But you've set up the punchline for me:* ***if there is no best element, you lack a convergence guarantee.***

DP: *But it seems as if good priors usually do have a best element. Using Laplace's rule of succession , I can predict coins of any bias without divergence.*

RJ: *What if the coin lands as follows: 5 heads in a row, then 25 tails, then 125 heads, and so on, each run lasting for the next power of five. Then you diverge again.*

DP: *Ok, sure... but if the coin flips might not be independent, then I should have hypotheses like that in my prior.*

RJ: *I could keep trying to give examples which break your prior, and you could keep trying to patch it. But we have agreed on the important thing: good priors should have the convergence property. At least you've agreed that this is a desirable property not always achieved by Bayes.*

DP: *Sure.*

In the end, I'm not sure who would win the counterexample/patch game: it's quite possible that there general priors with convergence guarantees. No computable prior has convergence guarantees for "sufficiently rich" observables (ie, observables including logical combinations of observables). However, that's a theorem with a lot of caveats. In particular, Solomonoff Induction isn't computable, so might be immune to the critique. And we can certainly get rid of the problem by restricting the observables, EG by conditioning on their sequential order rather than just their truth. Yet, I suspect all such solutions will either be really dumb, or uncomputable.

So there's work to be done here.

But, in general (ie *without any special prior which does guarantee convergence for restricted observation models*), a Bayesian relies on a realizability (aka grain-of-truth) assumption for convergence, as it does for some other nice properties. Radical probabilism demands these properties without such an assumption.

So much for technical details. Another point I want to make is that convergence points at a notion of "objectivity" for the radical probabilist. Although the individual updates a radical probabilist makes can go all over the place, the beliefs must eventually settle down to something. The goal of reasoning is to settle down to that answer as quickly as possible. Updates may appear arbitrary from the outside, but internally, they are always moving toward this goal.

This point is further emphasized by the next rationality property: conservation of expected evidence.

Conservation of Expected Evidence

The law of conservation of expected evidence is a dearly beloved Bayesian principle. You'll be glad to hear that it survives unscathed:

$$P_n(X) = E_n P_m(X)$$

In the above, $P_n(X)$ is your current belief in some proposition X; $P_m(X)$ is some future belief about X (so I'm assuming $m > n$); and E_n is the expected value operator according to your current beliefs. So what the equation says is: your current beliefs equal your expected value of your future beliefs. This is just like the usual formulation of no-expected-net-update, except we no longer take the expectation *with respect to evidence*, since a non-Bayesian update may not be grounded in evidence.

Proof: Suppose $P_n(X) \neq E_n P_m(X)$. One of the two numbers is higher, and the other lower. Suppose $E_n P_m(X)$ is the lower number. Then a bookie can buy a certificate paying $\$P_m(X)$ on day m; we will willingly sell the bookie this for $\$E_n P_m(X)$. The bookie can also sell us a certificate paying \$1 if X, for a price of $\$P_n(X)$. At time m, the bookie gains $\$P_m(X)$ due to the first certificate. It can then buy the second certificate back from us for $\$P_m(X)$, using the

winnings. Overall, the bookie has now paid $\$E_nP_m(X)$ to us, but we have paid the bookie $\$P_n(X)$, which we assumed was greater. So the bookie profits the difference.

If $P_n(X)$ is the lower number instead, the same strategy works, reversing all buys and sells. □

The key idea here is that both a direct bet on X and a bet on $P_m(X)$ will be worth $P_m(x)$ later, so they'd better have the same price now, too.

I see this property as being even more important for a radical probabilist than it is for a dogmatic probabilist. For a dogmatic probabilist, it's a consequence of Bayesian conditional probability. For a radical probabilist, it's a basic condition on rational updates. With updates being so free to go in any direction, it's an important anchor-point.

Another name for this law is *the martingale property*. This is a property of many stochastic processes, such as Brownian motion. From wikipedia:

> In probability theory, a **martingale** is a sequence of random variables (i.e., a stochastic process) for which, at a particular time, the conditional expectation of the next value in the sequence, given all prior values, is equal to the present value.

It's important that a sequence of rational beliefs have this property. Otherwise, future beliefs are different from current beliefs in a predictable way, and we would be better off updating ahead of time.

Actually, that's not immediately obvious, right? The bookie in the Dutch Book argument doesn't make money by updating to the future belief faster than the agent, but rather, by playing the agent's beliefs off of each other.

This leads me to a stronger property, which has the martingale property as an immediate consequence (**strong self trust**):

$$P_n(X \mid P_m(X) = y) = y$$

Again I'm assuming m > n. The idea here is supposed to be: *if you knew your own future belief, you would believe it already*. Furthermore, you believe X and $P_m(X)$ are perfectly correlated: the only way you'd have high confidence in X would be if it were very probably true, and the only way you'd have low confidence would be for it to be very probably false.

I won't try to prove this one. In fact, be wary: this rationality condition is a bit too strong. The condition holds true in the radical-probabilism formalization of *Diachronic Coherence and Radical Probabilism* by Brian Skyrms, so long as $P_n(P_m(X) = y) > 0$ (see section 6 for statement and proof). However, *Logical Induction* argues persuasively that this condition is undesirable in specific cases, and replaces it with a slightly weaker condition (see section 4.12).

Nonetheless, for simplicity, I'll proceed as if *strong self trust* were precisely true.

At the end of the previous section, I promised that the current section would further illuminate my remark:

> The goal of reasoning is to settle down to that answer as quickly as possible. Updates may appear arbitrary from the outside, but internally, they are always moving toward this goal.

The way radical probabilism allows *just about any change* when beliefs shift from P_n to P_{n+1} may make its updates seem irrational. How can the update be *anything*, and still be called rational? Doesn't that mean a radical probabilist is open to garbage updates?

No. A radical probabilist doesn't *subjectively* think all updates are equally rational. A radical probabilist *trusts the progression of their own thinking*, and also *does not yet know the outcome of their own thinking*; this is why I asserted earlier that a fluid update can be just about anything (barring the transformation of a zero into a positive probability). However, this does not mean that a radical probabilist would accept a psychedelic pill which arbitrarily modified their beliefs.

Suppose a radical probabilist has a sequence of beliefs $P_1, P_2, P_3, P_4, \dots, P_n$. If they thought hard for a while, they could update to P_{n+1}. On the other hand, if they took the psychedelic pill, their beliefs would be modified to become Q. The sequence would be abruptly disrupted, and go off the rails: $P_1, P_2, P_3, \dots, P_n, Q, R, S, \dots$

The radical probabilist does not trust *whatever they believe next*. Rather, the radical probabilist has a concept of *virtuous epistemic process*, and is willing to believe the next output of such a process. Disruptions to the epistemic process do not get this sort of trust without reason. (For those familiar with *The Abolition of Man*, this concept is very reminiscent of his "Tao".)

On the other hand, a radical probabilist *could* trust a different process. One person, P, might trust that another person, Q, is better-informed about any subject:

$$P_n(X \mid Q_n(X) = y) = y$$

This says that P trusts Q on any subject if they've had the same amount of time to think. This leaves open the question of what P thinks if Q has had longer to think. In the extreme case, it might be that P thinks Q is better *no matter how long* P *has to think:*

$$\forall_{m,n} P_m(X \mid Q_n(X) = y) = y$$

On the other hand, P and Q can both be perfectly rational by the standards of radical probabilism *and not trust each other at all*. P might not trust Q's opinion no matter how long Q thinks.

(Note, however, that you *do* get eventual agreement on matters where good feedback is available -- much like in dogmatic Bayesianism, it's difficult for two Bayesians to disagree about *empirical predictions* for long.)

This means you can't necessarily replace one "virtuous epistemic process" with another. $P_1, P_2, P_3, \ldots$ and $Q_1, Q_2, Q_3, \ldots$ might both be perfectly rational by the standards of radical probabilism, and yet the disrupted sequence $P_1, P_2, P_3, Q_4, Q_5, Q_6, \ldots$ would not be, because P_3 does not necessarily trust Q_4 or subsequent Qs.

Realistically, we can be in this kind of position *and not even know what constitutes a virtuous reasoning process by our standards*. We generally think that we can "do philosophy" and reach better conclusions. But we don't have a clean specification of our own thinking process. We don't know exactly what counts as a virtuous continuation of our thinking vs a disruption.

This has some implications for AI alignment, but I won't try to spell them out here.

Calibration

One more rationality property before we move on.

One could be forgiven for reading Eliezer's A Technical Explanation of Technical Explanation and coming to believe that Bayesian reasoners are calibrated. Eliezer goes so far as to suggest that we *define* probability in terms of calibration, so that *what it means* to say "90% probability" is that, in cases where you say 90%, the thing happens 9 out of 10 times.

However, the truth is that calibration is a neglected property in Bayesian probability theory. Bayesian updates do not help you learn to be calibrated, any more than they help your beliefs to be convergent.

We can make a sort of Dutch Book argument for calibration: if things happen 9-out-of-ten times when the agent says 80%, then a bookie can place bets with the agent at 85:15 odds and profit in the long run. (Note, however, that this is a bit different from typical Dutch Book arguments: it's a strategy in which the bookie risks some money, rather than just getting a sure gain. What I can say is that Logical Induction treats this as a valid Dutch Book, and so, we get a calibration property in that formalism. I'm not sure about other formalisations of Radical Probabilism.)

The intuition is similar to convergence: even lacking a hypothesis to explain it, a rational agent should eventually notice "hey, when I say 80%, the thing happens 90% of the time!". It can then improve its beliefs in future cases by adjusting upwards.

This illustrates "meta-probabilistic beliefs": a radical probabilist can have informed opinions *about the beliefs themselves*. By default, a classical Bayesian doesn't have beliefs-about-beliefs except as a result of learning about the world and reasoning about themselves as a part of the world, which is problematic in the classical Bayesian formalism. It is possible to add second-order probabilities, third-order, etc. But calibration is a case which collapses all those levels, illustrating how the radical probabilist can handle all of this more naturally.

I'm struck by the way calibration *is something Bayesians obviously want*. The set of people who advocate applying Bayes Law and the set of people who look at calibration charts for their own probabilities has a *very significant overlap*. Yet, Bayes' Law does not give you calibration. It makes me feel like more people should have noticed this sooner and made a bigger deal about it.

Bayes From a Distance

Before any more technical details about radical probabilism, I want to take a step back and give one intuition for what's going on here.

We can see radical probabilism as *what a dogmatic Bayesian looks like if you can't see all the details*.

The Rationality of Acquaintances

Imagine you have a roommate who is perfectly rational in the dogmatic sense: this roommate has low-level observations which are 100% confident, and performs a perfect Bayesian update on those observations.

However, observing your roommate, you can't track all the details of this. You talk to your roommate about some important beliefs, but you can't track every little Bayesian update -- that would mean tracking every sensory stimulus.

From your perspective, your roommate has constantly shifting beliefs, which can't quite be accounted for. If you are particularly puzzled by a shift in belief, you can discuss reasons. "I updated against getting a cat because I observed a hairball in our neighbor's apartment." Yet, none of the evidence discussed is itself 100% confident -- it's at least a little bit removed from low-level sense-data, and at least a little uncertain.

Yet, this is not a big obstacle to viewing your roommate's beliefs as rational. You can evaluate these beliefs on their own merits.

I've heard this model called *Bayes-with-a-side-channel*. You have an agent updating via Bayes, but part of the evidence is hidden. You can't give a formula for changes in belief over time, but you can still assert that they'll follow conservation of expected evidence, and some other rationality conditions.

What Jeffrey proposes is that we allow these dynamics without necessarily positing a side-channel to explain the unpredictable updates. This has an anti-reductionist flavor to it: updates do not have to reduce to observations. But why should we be reductionist in that way? Why would subjective belief updates *need* to reduce to observations?

(Note that Bayes-with-a-side-channel does not imply conditions such as convergence and calibration; so, Jeffrey's theory of rationality is more demanding.)

Wetware Bayes

Of course, Jeffrey would say that our relationship with ourselves is much like the roommate in my story. Our beliefs move around, and while we can often give some account of why, we can't give a full account in terms of things we've learned with 100% confidence. And it's not simply because we're a Bayesian reasoner who lacks introspective access to the low-level information. The nature of our wetware is such that there isn't really any place you can point to and say "this is a 100% known observation". Jeffrey would go on to point out that there's no clean dividing line between external and internal, so you can't really draw a boundary between external event and internal observation-of-that-event.

(I would remark that Jeffrey doesn't exactly give us a way to *handle* that problem; he just offers an abstraction which doesn't chafe on that aspect of reality so badly.)

Rather than imagining that there are perfect observations somewhere in the nervous system, we can instead imagine that a sensory stimulus exerts a kind of "evidential pressure" which can be less than 100%. These evidential pressures can also come from within the brain, as is the case with logical updates.

But Where Do Updates Come From?

Dogmatic probabilism raises the all-important question "where do priors come from?" -- but once you answer that, everything else is supposed to be settled. There have been many debates about what constitutes a rational prior.

> **Q. How can I find the priors for a problem?**
> **A.** Many commonly used priors are listed in the *Handbook of Chemistry and Physics.*
>
> **Q. Where do priors *originally* come from?**
> **A.** Never ask that question.
>
> **Q. Uh huh. Then where do scientists get their priors?**
> **A.** Priors for scientific problems are established by annual vote of the AAAS. In recent years the vote has become fractious and controversial, with widespread acrimony, factional polarization, and several outright assassinations. This may be a front for infighting within the Bayes Council, or it may be that the disputants have too much spare time. No one is really sure.
>
> **Q. I see. And where does everyone else get their priors?**
> **A.** They download their priors from Kazaa.
>
> **Q. What if the priors I want aren't available on Kazaa?**
> **A.** There's a small, cluttered antique shop in a back alley of San Francisco's Chinatown. *Don't ask about the bronze rat.*
>
> -- Eliezer Yudkowsky, *An Intuitive Explanation of Bayes' Theorem*

Radical probabilists put less emphasis on the prior, since a radical probabilist can effectively "decide to have a different prior" (updating their beliefs as if they'd swapped out one prior for another). However, they face a similarly large problem of where *updates* come from.

We are given a picture in which beliefs are like a small particle in a fluid, reacting to all sorts of forces (some strong and some weak). Its location gradually shifts as a result of Brownian motion. Presumably, the interesting work is being done behind the scenes, by whatever is *generating* these updates. Yet, Jeffrey's picture seems to mainly be about the dance of the particle, while the fluid around it remains a mystery.

A full answer to that question is beyond the scope of this post. (Logical Induction offers *one* fully detailed answer to that question.) However, I do want to make a few remarks on this problem.

- It might at first seem strange for beliefs to be so radically malleable to external pressures. But, actually, this is already the familiar Bayesian picture: everything happens due to externally-driven updates.
- Bayesian updates don't really answer the question of where updates come from, either. They take it as given that there are some "observations". Radical probabilism simply allows for a *more general* sort of feedback for learning.
- An orthodox probabilist might answer this challenge by saying something like: when we design an agent, we design sensors for it. These are connected in such a way as to feed in sensory observations. A radical probabilist can similarly say: when we design an agent, we get to decide what sort of feedback it uses to improve its beliefs.

The next section will give some practical, human examples of non-Bayesian updates.

Virtual Evidence

Bayesian updates are path-independent: it does not matter in what order you apply them. If you first learn A and then learn B, your updated probability distribution is

$P_3(X) = P_2(X|B) = P_1(X|A\&B)$. If you learn these facts the other way around, it's still

$P_3(X) = P_2(X|A) = P_1(X|A\&B)$.

Jeffrey updates are path-dependent. Suppose my probability distribution is as follows:

	A	¬A
B	30%	20%
¬B	20%	30%

I then apply the Jeffrey update P(B)=60%:

	A	¬A
B	36%	24%
¬B	16%	24%

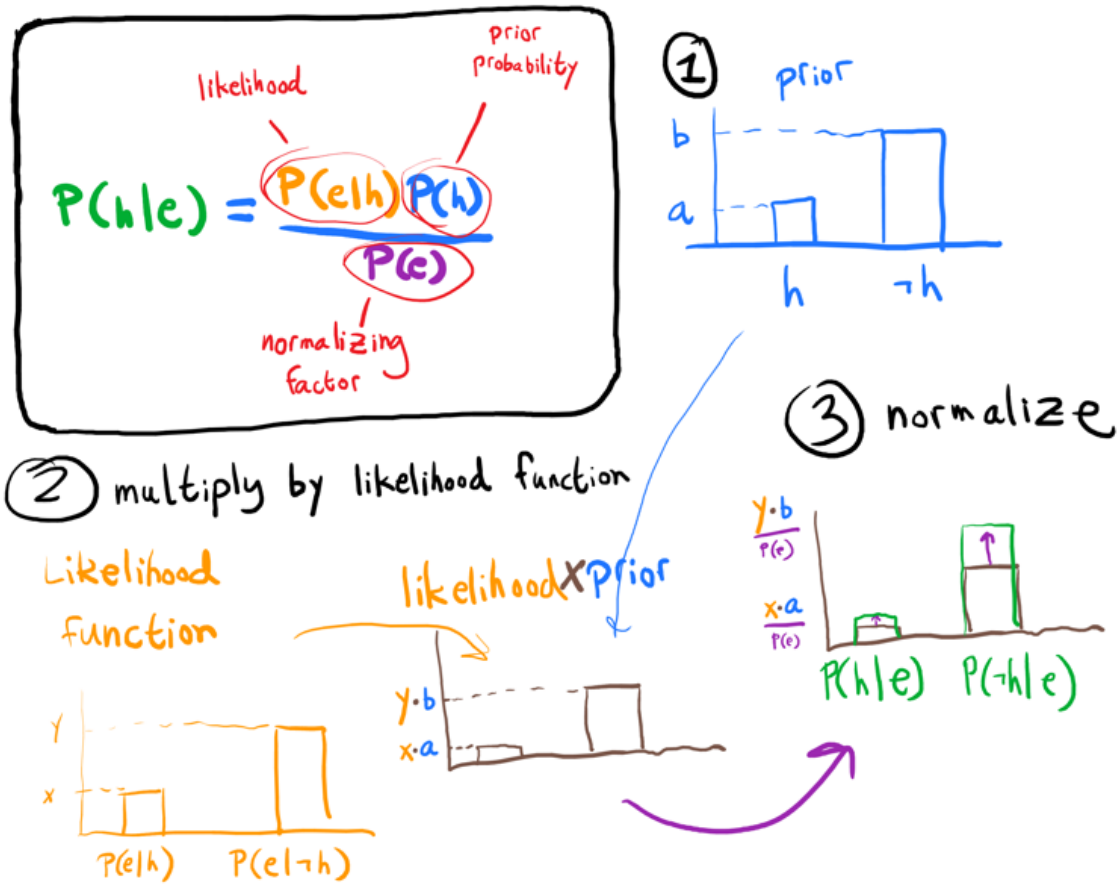
Now I apply P(A)=60%:

	A	¬A
B	41.54%	20%
¬B	18.46%	20%

Since this is asymmetric, but the initial distribution was symmetric, obviously this would turn out differently if we had applied the Jeffrey updates in a different order.

Jeffrey considered this to be a bug -- although he seems fine with path-dependence under some circumstances, he used examples like the above to motivate a *different* way of handling uncertain evidence, which I'll call ***virtual evidence***. (Judea Pearl strongly advocated virtual evidence over Jeffrey's rule near the beginning of Probabilistic Reasoning in Intelligent Systems (Section 2.2.2 and 2.3.3), in what can easily be read as a critique of Jeffrey's theory -- if one does not realize that Jeffrey is largely in agreement with Pearl. I thoroughly recommend Pearl's discussion of the details.)

Recall the basic anatomy of a Bayesian update:



The idea of virtual evidence is to use evidence 'e' which is not an event in our event space. We're just acting *as if* there were evidence 'e' which justifies our update. Terms such as P(e), P(e&h), P(e|h), P(h|e), and so on are not given the usual probabilistic interpretation; they just stand as a convenient notation for the update. ***All we need to know is the likelihood function for the update.*** We then multiply our probabilities by the likelihood function as usual, and normalize. P(e) is easy to find, since it's just whatever factor makes everything sum to one at the end. This is good, since it isn't clear what P(e) would mean for a virtual event.

Actually, we can simplify even further. All we *really* need to know is the likelihood *ratio:* the ratio between the two numbers in the likelihood function. (I will illustrate this with an example soon). However, it may sometimes be easier to find the whole likelihood function in practice.

Let's look at the path-dependence example again. As before, we start with:

	A	¬A
B	30%	20%
¬B	20%	30%

I want to apply a Jeffrey update which makes P(B)=60%. However, let's represent the update via virtual evidence this time. Currently, P(B)=50%. To take it to 60%, we need to see virtual evidence with a 60:40 likelihood ratio, such as P(B|E)=60%, P(¬B|E)=40%. This gives us the same update as before:

	A	¬A

B	36%	24%
¬B	16%	24%

(Note that we would have gotten the same result with a likelihood function of P(B|E)=3%, P(¬B|E)=2%, since 60:40 is the same as 3:2. That's what I meant when I said that only the ratio matters.)

But now we want to apply the same update to A as we did to B. So now we update on virtual evidence P(A|E)=60%, P(¬A|E)=40%. This gives us the following (approximately):

	A	¬A
B	43%	19%
¬B	19%	19%

As you can see, the result is quite symmetric. In general, virtual evidence updates will be path-independent, because multiplication is commutative (and the normalization step of updating doesn't mess with this commutativity).

So, virtual evidence is a reformulation of Jeffrey updates with a lot of advantages:

- Unlike raw Jeffrey updates, virtual evidence is path-independent.
- You don't have to decide right away what you're updating *to*; you just have to decide the strength and direction of the update.
- I don't fully discuss this here, but Pearl argues persuasively that it's easier to tell when a virtual-evidence update is appropriate than when a Jeffrey update is appropriate.

Because of these features, virtual evidence is much more useful for *integrating information from multiple sources.*

Integrating Expert Opinions

Suppose you have an ancient artefact. You want to know whether this artefact was made by ancient aliens. You have some friends who are also curious about ancient aliens, so you enlist their help.

You ask one friend who is a metallurgist. After performing experiments (the details of which you don't understand), the metallurgist isn't sure, but gives 80% that the tests would turn out that way if it were of terrestrial origin, and 20% for metals of non-terrestrial origin. (Let's pretend that ancient aliens would 100% use metals of non-Earth origin, and that ancient humans would 100% use Earth metals.)

You then ask a second friend, who is an anthropologist. The anthropologist uses cultural signs, identifying the style of the art and writing. Based on that information, the anthropologist estimates that it's half as likely to be of terrestrial origin as alien.

How do we integrate this information? According to Jeffrey and Pearl, we can apply the virtual evidence formula *if we think the two expert judgements are independent*. What 'independence' means for virtual evidence is a bit murky, since the evidence is not part of our probability calculus, so we can't apply the usual probabilistic definition. However, Pearl argues persuasively that this condition is easier to evaluate in practice than the *rigidity* condition which governs the applicability of Jeffrey updates. (He also gives an example where rigidity is violated, so a naive Jeffrey update gives a nonsensical result but where virtual evidence can still be easily applied to get a correct result.)

The information provided by the anthropologist and the metallurgist seem to be quite independent types of information (at least, if we ignore the fact that both experts are biased by an interest in ancient aliens), so let's apply the virtual evidence rule. The likelihood ratio

from the metallurgist was 80:20, which simplifies to 4:1. The likelihood ratio from the anthropologist was 1:2. That makes the combined likelihood vector 2:1 in favor of terrestrial origin. We would then combine this with our prior; for example, if we had a prior of 3:1 in favor of a terrestrial origin, our posterior would be 6:1 in favor.

(Note that we also have to think that the virtual evidence is independent *of our prior information.)*

So, virtual evidence offers a practical way to integrate information when we cannot quantify exactly what the evidence was -- a condition which is especially likely when consulting experts. This illustrates the utility of the bayes-with-a-side-channel model mentioned earlier; we are able to deal effectively with evidence, even when the exact nature of the evidence is hidden to us.

A few notes on how we gathered expert information in our hypothetical example.

- We asked for likelihood ratios, rather than posterior probabilities. This allows us to combine the information as virtual evidence.
- In the case of the metallurgist, it makes sense to ask for likelihood ratios, since the metallurgist is unlikely to have good prior information about the artefact. Asking only for likelihoods allows us to factor out any effect from this poor prior (and instead use our own prior, which may still be poor, but has the advantage of being ours).
- In the case of the anthropologist, however, it doesn't make as much sense -- if we trust their expertise, we're likely to think the anthropologist has a good prior about artefacts. It might have made more sense to ask for the anthropologist's posterior, take it as our own, and *then* apply a virtual-evidence update to integrate the metallurgist's report. (However, if we weren't able to properly communicate our own prior information to the anthropologist, it would be ignored in such an approach.)
- In the case of the metallurgist, it felt more natural to give a full likelihood function, rather than a likelihood ratio. It makes sense to know the probability of test result given a particular substance. It would have made even more sense if the likelihood function were *a function of each metal the artefact could be made of*, rather than just "terrestrial" or "extraterrestrial" -- using broad categories allows the metallurgist's prior about specific substances to creep in, which might be unfortunate.
- In the case of the anthropologist, however, it didn't make sense to give a full likelihood function. "The probability that the artefact would look exactly the way it looks assuming that it's made by humans" is very very low, and seems quite difficult and unnatural to evaluate. It seems much easier to come up with a likelihood *ratio,* comparing the probability of terrestrial and extraterrestrial origin.

Why did Pearl devote several sections to virtual evidence, in a book which is otherwise a bible for dogmatic probabilists? I think the main reason is the close analogy to the mathematics of Bayesian networks. The message-passing algorithm which makes Bayesian networks efficient is almost exactly the virtual evidence procedure I've described. If we think of each node as an expert trying to integrate information from its neighbors, then the efficiency of Bayes nets comes from the fact that they can use virtual evidence to update on likelihood functions rather than needing to know about the evidence in detail. This may have even been one source of inspiration for Pearl's belief propagation algorithm?

Can Dogmatic Probabilists Use Virtual Evidence?

OK, so we've put Jeffrey's radical updates into a more palatable form -- one which borrows the structure and notation of classical Bayesian updates.

Does this mean orthodox Bayesians can join the party, and use virtual evidence to accomplish everything a radical probabilist can do?

No.

Virtual evidence abandons the ratio formula.

One of the longstanding axioms of classical Bayesian thought is the ratio formula for conditional probability that Bayes himself introduced:

$$P(A \mid B) = \frac{P(A \& B)}{P(B)}$$

Virtual evidence, as an updating practice, holds that $P(A|B)$ can be usefully defined in cases where the ratio $P(A\&B)/P(B)$ ***cannot*** be usefully defined. Indeed, virtual evidence treats Bayes' Law (which is usually a derived theorem) as more fundamental than the ratio formula (which is usually taken as a definition).

Granted, dogmatic probabilism *as I defined it at the beginning of this post* does not explicitly assume the ratio formula. But the assumption is so ingrained that I assume most readers took $P(A|B)$ to mean the ratio.

Still, even so, we can *consider* a version of dogmatic probabilism which rejects the ratio formula. Couldn't they use virtual evidence?

Virtual evidence requires probability functions to take arguments which aren't part of the event space.

Even abandoning the ratio formula, still, it's hard to see how a dogmatic probabilist could use virtual evidence without abandoning the Kolmogorov axioms as the foundation of probability theory. The Kolmogorov axioms make probabilities a function of events; and events are taken from a pre-defined event space. Virtual evidence constructs new events at will, and does not include them in an overarching event space (so that, for example, virtual evidence V can be defined -- so that $P(X|V)$ is meaningful for all X from the event space --without events like $X\&V$ being meaningful, as would be required for a sigma-algebra).

I left some wiggle room in my definition, saying that a dogmatic probabilist might endorse the Kolmogorov axioms "or a similar axiomatization of probability theory". But even the Jeffrey-Bolker axioms, which are pretty liberal, don't allow enough flexibility for this!

Representing Fluid Updates

A final point about virtual evidence and Jeffrey updates.

Near the beginning of this essay, I gave a picture in which Jeffrey updates generalize Bayesian updates, but *fluid* updates generalize things even further, opening up the space of possibilities when rigidity does not hold.

However, I should point out that *any update is a Jeffrey update on a sufficiently fine partition.*

So far, for simplicity, I've focused on binary partitions: we're judging between H and $\neg H$, rather than a larger set such as H_1, H_2, H_3. However, we can generalize everything to arbitrarily sized partitions, and will often want to do so. I noted that a larger set might have

been better when asking the metallurgist about the artefact, since it's easier to judge the probability of test results given specific metals rather than broad categories.

If we make a partition large enough to cover every possible combination of events, then a Jeffrey update is now just a completely arbitrary shift in probability. Or, alternatively, we can represent arbitrary shifts via virtual evidence, by converting to likelihood-ratio format.

So, these updates are completely general after all.

Granted, there might not be any *point* to seeing things that way.

Non-Sequential Prediction

One advantage of radical probabilism is that it offers a more general framework for statistical learning theory. I already mentioned, briefly, that it allows one to do away with the realizability/grain-of-truth assumption. This is very important, but not what I'm going to dwell on here. Instead I'm going to talk about non-sequential prediction, which is a benefit of logical induction which I think has been under-emphasized so far.

Information theory -- in particular, algorithmic information theory -- in particular, Solomonoff induction -- is restricted to a *sequential prediction* frame. This means there's a very rigid observation model: observations are a sequence of tokens and you always observe the *nth* token after observing tokens one through *n-1*.

Granted, you can fit lots of things into a sequential prediction model. However, it is a flaw the otherwise close relationship between Bayesian probability and information theory. You'll run into this if you try to relate information theory and logic. Can you give an information-theoretic intuition for the laws of probability that deal with logical combinations, such as P(A or B) + P(A and B) = P(A) + P(B)?

I've complained about this before, offering a theorem which (somewhat) problematizes the situation, and suggesting that people should notice whether or not they're making sequential-prediction style assumptions. I almost included related assumptions in my definition of dogmatic probabilism at the beginning of this post, but ultimately it makes more sense to contrast radical probabilism to the more general doctrine of Bayesian updates.

Sequential prediction cares only about the accuracy of beliefs *at the moment of observation*; the accuracy of the full distribution over the future is reduced to the accuracy about each next bit as it is observed.

If information is coming in "in any old way" rather than according to the assumptions of sequential prediction, then we can construct problematic cases for Solomonoff induction. For example, if we condition the *nth* bit to be 1 (or 0) when a theorem prover proves (or refutes) the *nth* sentence of Peano arithmetic, then Solomonoff induction will never assign positive probability to hypotheses consistent with Peano arithmetic, and will therefore do poorly on this prediction task. This is despite the fact that there are *computable* programs which do better at this prediction task; for example, the same theorem prover running just a little bit faster can have highly accurate beliefs at the moment of observation.

In non-sequential prediction, however, we care about accuracy *at every moment*, rather than just at the moment of observation. Running the same theorem prover, just one step faster, doesn't do very well on that metric. It allows you to get things right just in time, but you won't have any clue about what probabilities to assign before that. We don't just want the right conclusion; we want to get there as fast as possible, and (in a subtle sense) via a rational path

Part of the difficulty of non-sequential prediction is how to score it. Bayes loss applied to your predictions at the moment of observation, in a sequential prediction setting, seems quite useful. Bayes loss *applied to all your beliefs, at every moment* does not seem very useful.

Radical probabilism gives us a way to evaluate the rationality of non-sequential predictions -- namely, how vulnerable the sequence of belief distributions was to losing money via some sequence of bets.

Sadly, I'm not yet aware of any appropriate generalization of information theory -- at least not one that's very interesting. (You can index information by time, to account for the way probabilities stift over time... but that does not come with a nice theory of communication or compression, which are fundamental to classical information theory.) This is why I objected to prediction=compression in the discussion section of Alkjash's talk.

To summarize, sequential prediction makes three critical assumptions which may not be true in general:

- It assumes observations will always inform us about one of a set of observable variables. In general, Bayesian updates can instead inform us about any event, including complex logical combinations (such as "either the first bit is 1, or the second bit is 0").
- It assumes these observations will be made in a specific sequence, whereas in general updates could come in in any order.
- It assumes that what we care about is the accuracy of belief *at the time of observation;* in general, we may care about the accuracy of beliefs at other times.

The only way I currently know how to get theoretical benefits similar to those of Solomonoff induction while avoiding all three of these assumptions is radical probabilism (in particular, as formalized by logical induction).

(The connection between this section and radical probabilism is notably weaker than the other parts of this essay. I think there is a lot of low-hanging fruit here, fleshing out the space of possible properties, the relationship between various problems and various assumptions, trying to generalize information theory, clarifying our concept of observation models, et cetera.)

Making the Meta-Bayesian Update

In *Pascal's Muggle* (long version, short version) Eliezer discusses situations in which he would be forced to make a non-Bayesian update:

> But if I actually see strong evidence for something I previously thought was super-improbable, I don't just do a Bayesian update, I should also question whether I was right to assign such a tiny probability in the first place - whether the scenario was really as complex, or unnatural, as I thought. In real life, you are not ever supposed to have a prior improbability of 10^{-100} for some fact distinguished enough to be written down, and yet encounter strong evidence, say 10^{10} to 1, that the thing has actually happened. If something like that happens, you don't do a Bayesian update to a posterior of 10^{-90}. Instead you question both whether the evidence might be weaker than it seems, *and* whether your estimate of prior improbability might have been poorly calibrated, because rational agents who actually have well-calibrated priors should not encounter situations like that until they are ten billion days old. Now, this may mean that I end up doing some non-Bayesian updates: I say some hypothesis has a prior probability of a quadrillion to one, you show me evidence with a likelihood ratio of a billion to one, and I say 'Guess I was wrong about that quadrillion to one thing' rather than being a Muggle about it.

At the risk of being too cutesy, I want to make two related points:

- At the object level, radical probabilism offers a framework in which we can make these sorts of non-Bayesian updates. We can encounter something which makes us question our whole way of thinking. It also allows us to significantly revise that way of thinking, without modeling the situation as something extreme like self-modification (or even something very out of the ordinary).
- At the meta level, updating to radical probabilism *is itself* one of these non-Bayesian updates. Of course, if you were really a hard-wired dogmatic probabilist at core, you would be unable to make such an update (except perhaps if we model it as self-modification). But, since you *are already* using reasoning which is actually closer in spirit to radical probabilism, you can start to model yourself in this way and using radical-probabilist ideas to guide future updates.

So, I wanted to use this penultimate section for some advice about making the leap.

It All Adds Up to Normality

Radical Probabilism is not a license to update however you want, nor even an invitation to massively change the way you update. It is primarily a new way to understand what you are already doing. Yes, it's possible that viewing things through this lense (rather than the more narrow lense of dogmatic probabilism) will change the way you see things, and as a consequence, change the way you do things. However, you are not (usually) making some sort of mistake by engaging in the sort of Bayesian reasoning you are familiar with -- there is no need to abandon large portions of your thinking.

Instead, try to notice ordinary updates you make which are not perfectly understood as Bayesian updates.

- Calibration corrections are not well-modeled as Bayesian updates. If you say to yourself "I've been overconfident in similar situations", and lower your probability, your shift is better-understood as a fluid update.
- Many instances of "outside view" are not well-modeled in a Bayesian update framework. You've probably seen outside view explained as prior probability. However, you often take the outside view on one of your own arguments, e.g. "I've often made arguments like this and been wrong". This kind of reflection doesn't fit well in the framework of Bayesian updates, but fits fine in a radical-probabilist picture.
- It is often warranted to downgrade the probability of a hypothesis without having an alternative in mind to upgrade. You can start to find a hypothesis suspicious without having any better way of predicting observations. For example, a sequence of surprising events might stick out to you as evidence that your hypothesis is wrong, even though your hypothesis is still the best way that you know to try and predict the data. This is hard to formalize as a Bayesian update. Changes in probability between hypotheses always remain balanced. It's true that you move the probability to a "not the hypotheses I know" category which balances the probability loss, but *it's not true that this category earned the increased probability by predicting the data better*. Instead, you used a set of heuristics which have worked well in the past to decide when to move probabilities around.

Don't Predictably Violate Bayes

Again, this is not a license to violate Bayes' Rule whenever you feel like it.

A radical probabilist should obey Bayes' Law in expectation, in the following sense:

If some evidence E or ¬E is bound to be observed by time m>n, then the following should hold:

$$E_n(P_m(H) \mid E) = P_n(H \mid E)$$

And the same for ¬E. In other words, you should not expect your updated beliefs to differ from your conditional probabilities on average.

(*You should suspect from the fact that I'm not proving this one that I'm playing a bit fast and loose -- whether this law holds may depend on the formalization of radical probabilism, and it probably needs some extra conditions I haven't stated, such as P(E)>0.*)

And remember, every update is a Bayesian update, with the right virtual evidence.

Exchange Virtual Evidence

Play around with the epistemic practice Jeffrey suggests. I suspect some of you already do something similar, just not necessarily calling it by this name or looking so closely at what you're doing.

Don't Be So Realist About Your Own Utility Function

Note that the picture here is quite compatible with what I said in *An Orthodox Case Against Utility Functions*. Your utility function need not be computable, and there need not be something in your ontology which you can think of your utility as a function of. All you need are utility *expectations*, and the ability to update those expectations. Radical Probabilism adds a further twist: you don't need to be able to predict those updates ahead of time; indeed, you probably can't. Your values aren't tied to a function, but rather, are tied to your trust in the ongoing process of reasoning which refines and extends those values (very much like the self-trust discussed in the section on conservation of expected evidence).

Not So Radical After All

And remember, every update *is* a Bayesian update, with the right virtual evidence.

Recommended Reading

Diachronic Coherence and Radical Probabilism, Brian Skyrms

- This paper is really nice in that it constructs Radical Probabilism from the ground up, rather than starting with regular probability theory and relaxing it. It provides a view in which diachronic coherence is foundational, and regular one-time-slice probabilistic coherence is derived. Like logical induction, it rests on a market metaphor. It also briefly covers the argument that radical-probabilism beliefs must have a convergence property.

Radical Probabilism and Bayesian Conditioning, Richard Bradley

- This is a more thorough comparison of radical probabilism to standard bayesian probabilism, which breaks down the departure carefully, while covering the fundamentals of radical probabilism. In addition to Bayesian conditioning and Jeffrey

conditioning, it introduces Adams conditioning, a new type of conditioning which will be valid in many cases (for the same sort of reason as why Jeffrey conditioning or Bayesian conditioning can be valid). He contends that there are, nonetheless, many more ways to update beyond these; and, he illustrates this with a purported example where none of those updates seems to be the correct one.

Epistemology Probabilized, Richard Jeffrey

- The man himself. This essay focuses mainly on how to update on likelihood ratios rather than directly performing Jeffrey updates (what I called virtual evidence). The motivations are rather practical -- updating on expert advice when you don't know precisely what observations lead to that advice.

I was a Teenage Logical Positivist (Now a Septuagenarian Radical Probabilist), Richard Jeffrey.

- Richard Jeffrey reflects on his life and philosophy.

Probabilistic Reasoning in Intelligent Systems, Judea Pearl.

- See especially chapter 2, especially 2.2.2 and 2.3.3.

Logical Induction, Garrabrant et al.

*: Jeffrey actually used this phrase. See *I was a Teenage Logical Positivist*, linked above.

The Bayesian Tyrant

Long ago and far away, there was a kingdom called Estimor in a broad green valley surrounded by tall grey mountains. It was an average kingdom in most respects, until the King read the works of Robin Hanson and Eliezer Yudkowsy, and decided to institute a Royalist Futarchy.

(This is a parable about the differences between Bayesian updates and logical induction. See also: Radical Probabilism .)

The setup was very simple. It followed the futarchic motto, "Vote Values, But Bet Beliefs" -- the only special consideration being that there was just one voting constituent (that being the King). A betting market would inform the King of everything He needed to know in order to best serve His interests and the interests of Estimor (which were, of course, one and the same).

The Seer's Hall -- a building previously devoted to religious prophecy -- was repurposed to serve the needs of the new betting market. (The old prophets were, of course, welcome to participate -- so long as they were willing to put money on the line.)

All went well at first. The new betting market allowed the King to set the revenue-maximizing tax rate, via the Laffer curve. An early success of the market was the forecasting of a grain shortage based on crop growth early in the season, which allowed ample time for grain to be purchased from neighboring lands.

Being an expert Bayesian Himself, the King would often wander the Seer's Hall, questioning and discussing with the traders at the market. Sometimes the King would be shocked by what he learned there. For example, many of the traders were calculating the Kelly betting criterion to determine how much to invest in a single bet. However, they then proceeded to invest *only a set fraction* of the Kelly amount (such as 80%). When questioned, traders replied that they were hedging against mistakes in their own calculations, or reducing volatility, or the like.

One day, the King noticed a man who would always run in and out of the Hall, making bets hastily. This man did particularly well at the betting tables -- he ended the day with a visibly heavy purse. However, when questioned by The King as to the source of his good luck, the man had no answers. This man will subsequently be referred to as the Fool.

The King ordered spies to follow the Fool on his daily business. That evening, spies returned to report that The Fool was running back and forth between the Seer's Hall and Dragon's Foot, a local tavern. The Fool would consult betting odds at Dragon's Foot, and return to the Seer's Hall to bet using those odds.

Evidently, Dragon's Foot had become an unlicensed gambling den. But were they truly doing better than the Seer's Hall, so that this man could profit simply by using their information?

The King had the Fool brought in for questioning. As it turned out, the Fool was turning a profit by *arbitrage* between the two markets: whenever there was a difference in prices, the Fool would bet in favor at the location where prices were low, and bet

against at the location where prices were high. In this way, he was making guaranteed money.

The King was disgusted at this way of making money without bringing valuable information to the market. He ordered all other gambling in the Kingdom shut down, requiring it to all take place at the Seer's Hall.

Soon after that, the Fool showed his face again. Once again, he did well in the market. The King had his spiel follow the Fool, but this time, he went nowhere of significance.

Questioning the Fool a second time, he learned that this time the Fool was making use of *calibration charts*. The Fool would make meticulous records of the true historical frequency of events given their probabilistic judgement -- for example, he had recorded that when the market judges an event to be 90% probable, that event actually occurs about 85% of the time. The Fool had made these records about individual traders as well as the market as a whole, and would place bets accordingly.

The King was once again disgusted by the way the Fool made money off of the market without contributing any external information. But this time, He felt that He needed a more subtle solution to the problem. Thinking back to his first days of reading about Bayes' Law, the King realized the huge gap between His vision of perfected reasoning and the reality of the crowded, noisy, irrational market. The iron law of the market was *buy low sell high*. It did not follow rational logic. The Fool had proved it: the individual traders were poorly calibrated, and so was the market itself.

What the King needed to do was reform the market, making it a more rational place.

And so it was that the King instituted the Bayesian Law: *all bets on the market are required to be Kelly bets made on valid probability estimates. Valid probability estimates are required to be Bayesian updates of previously registered probability estimates.*

All traders on the market would now proceed according to Bayes' Law. They would pre-register their probability distributions, pre-specifying what kind of information would update them, and by how much it would update them.

The new ordinance proved burdensome. Only a few traders continued to visit the Seer's Hall. They spent their days in meticulous calculation, updating detailed prior models of grain and weather with all the data which poured in.

Surprisingly, the Fool was amongst the hangers-on, and continued to make a tidy profit, even to the point of driving out some of the remaining traders -- they simply couldn't compete with him.

The King examined the registered probability distribution the Fool was using. It proved puzzling. The Fool's entire probability distribution was based on numbers which were to be posted to a particular tree out by Mulberry road. Updating on these numbers, the Fool was somehow a tidy profit. But where were the numbers coming from?

The King's spies found that the numbers were being posted by a secretive group, whose meetings they were unable to infiltrate.

The King had all the attendees arrested, accusing them of running an illegal gambling ring. The Fool was brought in for questioning once more.

"But it wasn't a gambling ring!" the Fool protested. "They merely got together and compiled *odds* for gambling. They were quite addicted when the Bayesian Law shut their sort out of the Seer's Hall, after all. And I took those odds and used them to bet in the Seer's Hall, perfectly legally."

"And redistributed the winnings?" accused the King.

"As is only fair," agreed the Fool. "But that is not gambling. I simply paid them as consultants."

"You took money from honest Bayesians, and drove them out of my Hall!"

"As is the advantage of Bayesianism, no?" The Fool cocked an eyebrow. "The money flows to he who can give the best odds."

"Take him away!" the king bellowed, waving a hand for the guards.

At that moment, the guards removed their helmets, revealing themselves to be comrades-in-arms with the Fool. The outcasts of the Seer's Hall had foreseen that the King would move against them, and with the power of Futarchy, had prepared well -- they staged a bloodless revolution that day.

The King, his family, and his most loyal staff were forced into exile. They went to stay with a distant cousin of the King, who ruled the nation Ludos, in the next valley over.

The King of Ludos had, upon seeing Estimor's success with prediction markets, set up His own. Unlike the Seer's Hall of Estimor, that of Ludos continued to thrive.

The King in Exile asked his cousin: "What did I do wrong? All I wanted was to serve Estimor. The prediction market worked so well at first. And I only tried to improve it."

The King of Ludos sat in thought for a time, and then spoke. "Cousin, We cannot tell You that You did anything wrong. Revolutions will happen. But We will say this: the many prediction markets of Ludos strengthen each other. Runners go back and forth between them, profiting from arbitrage, and this only makes them stronger. Calibration traders correct any bias of the market, ensuring unbiased results. You tried to outlaw the irrational at your market -- but remember, the wise gambler profits off the foolhardy gambler. Without any novices throwing away their money, none would profit."

"But most of all, cousin, We think You lost sight of the power of betting. It is a truth more fundamental than Bayes' Law that money will flow from the unclever to the clever. You lost Your trust in that system. Even if You had enforced Kelly betting, but left it to each individual trader to set his probability however he liked -- rather than updating via Bayes' Law alone -- you would have been fine. If Bayes' Law were truly the correct way, money would have flowed to those who excelled at it. If not, then money would have flowed elsewhere. But instead you overwhelmed them with the bureaucracy of Bayes -- requiring them to record every little bit of information they used to reach a conclusion."

The arbitrage between different betting halls represented outside view / modest epistemology, trying to reach agreement between different reasoners. It's a questionable thing to include, in terms of the point I'm making, since this is not exactly a thing that happens in logical induction. However, it fits in the allegory so well

that I felt I couldn't not include it. One argument for the common prior assumption (an assumption which underpins the Aumann Agreement Theorem, and is closely related to modest-epistemology arguments) is that a bookie can Dutch Book any group of agents who do not have a common prior, via performing arbitrage on their various beliefs.

[Edit: actually, what we can conclude from the analogy is that bets on different markets should converge to the same thing ***if they ever pay out****, which is also true in logical induction.]*

The calibration-chart idea, clearly, represented calibration properties .

The idea of the Bayesian Law represented requiring all hypotheses/traders to update in a Bayesian manner. Starting from Bayesian hypothesis testing, one step we can take in the direction of logical induction is to allow hypotheses to themselves make non-Bayesian updates. The overall update ***between*** *hypotheses would remain Bayesian, but an individual hypothesis could change its mind in a non-Bayesian fashion. A hypothesis would still be required to have a coherent probability distribution at any given time; just, the updates could be non-Bayesian. A fan of Bayes' Law might suppose that, in such a situation, the hypotheses which update according to Bayes' Law would dominate -- in other words, a meta-level Bayesian would learn to also be an object-level Bayesian. But I see no reason to suspect this to be the case. Indeed, in situations where logical uncertainty is relevant, non-Bayesian updates can be used to continue improving one's probability distribution over and above the explicit evidence which comes in. It was this idea -- that we could take one step toward logical induction by being a meta-level Bayesian without being an object-level Bayesian -- which inspired this post (although the allegory didn't end up having such a strong connection with this idea).*

The main point of this post, anyway, is that ***Bayes' Law would be a bad law.*** *Don't institute a requirement that everyone reason according to it.*

Time Travel Markets for Intellectual Accounting

In a comment to Can we hold intellectuals to similar public standards as athletes? I said that one of the largest problems with using prediction accuracy (EG prediction markets) as the ***exclusive*** standard by which we judge intellectuals would be *undervaluing contributing to the thought process*.

Here, I propose a modification of prediction markets which values more types of contributions.

Arguments as Models

In many cases, a thought/argument can be reformulated as a *model*. What I mean by this is a formally described way of making predictions.

Adding models to prediction markets could increase their transparency; we want to create an incentive for traders to *explain themselves.*

A model has a *complexity* (description length; the negative log of its prior probability). We generally have more respect for predictions which come from models, in contrast to those which don't. And, given two models, we generally have more respect for the simpler one.

So, to adjust prediction markets to account for models, we would want something which:

- Values making good predictions over everything else;
- Values models over opaque predictions;
- Values simple models more.

But how can we trade off between modelling and predictive accuracy?

Time Travel Markets

As I've discussed before, one of the major differences between prediction markets and Bayesianism is that prediction-market traders only get credit for *moving the market in the right direction,* which introduces a dependence on *when* a prediction is made: if you make an accurate prediction at a time when everyone else has already arrived at that conclusion, you don't get any credit; whereas, if you make an accurate prediction at a time when it's unpopular, you'll get a lot of credit for that.

That's a problem for model-makers. An *explanation* is still useful, even after all the facts needing explanation are agreed upon.

This is a generalization of the problem of old evidence.

Solomonoff induction solves the problem of old evidence for Bayesians as follows: a new hypothesis which explains old evidence is evaluated as if it were in our set of

hypotheses from the beginning. (After all, it *would have been* in the Solomonoff prior from the beginning; it was only absent from our calculations because we are forced to approximate Solomonoff.) The prior probability is based on the complexity of the hypothesis; the updates are whatever they would have been.

(Critics of Bayesianism rightly point out that adding a hypothesis in this way is not a Bayesian update, making this a problem for Bayesianism; but let's set aside that problem for the moment. We can think of the adherents of Solomonoff induction as endorsing *two* types of update; the Bayesian update, and a non-Bayesian update to account for logical uncertainty by adding hypotheses as we improve our approximation of Solomonoff.)

Similarly, we could imagine adding models to our market in this way: we score them *as if* they were there from the beginning, with a starting weight based on their complexity.

This gives models an advantage over regular trades. Regular trades, made by humans, are always scored at the time they're submitted. But models can get credit for predicting things *even after those things are known by the market*, because their predictions are assessed retroactively.

On the other hand, it satisfies our desire to favor models without compromising predictive accuracy:

- Unless the prior probability of a model is overwhelmingly high (due to very low complexity), a high posterior probability simply means a strong agreement with the market. So, this doesn't warp the market to reduce predictive accuracy in favor of modeling; for the most part, it will keep the market the same, but give credit to the model for *explaining* the market.
- If the prior probability *is* quite high, *and* later evidence isn't overwhelming, it *will* warp the market; but this basically makes sense, because in the face of uncertainty we *do* favor simple hypotheses.

So this mostly satisfies the desiderata I set down earlier.

A Virtual Currency

This is more of an implementation detail, but I want to discuss how the prediction market currency works.

Prediction markets can of course use real money, and this has some advantages. However:

- If we're trying to use a prediction market to keep track of the accuracy of public intellectuals, it doesn't make so much sense to use real money. We want a non-transferable currency of accuracy.
- Using real money also constrains things; we can't create or destroy it at will, but we might want to do so for our virtual currency.

Starting Budgets

One problem: how much currency is given to people when they start on the market?

If we give the same starting currency to each person who joins, then the amount of currency in the market just keeps growing and growing as people join. This seems problematic. For example, a large number of newcomers could warp the market in any direction they liked despite having no track record; and, people with a poor track record are highly incentivised to abandon their old accounts and go to new ones.

One solution would be to use a sequence which sums to a finite number. For example, the starting budget of each new account could be cut in half... but this seems too harsh, favoring early accounts over later ones to a very high degree. Any coefficient $c < 1$ could be used, with the nth account getting c^n starting funds. (This always sums to a finite number.) Selecting c very close to 1 makes things more fair (only slightly decreasing the value of new accounts over time), although putting things too close to 1 brings back the same weird dynamics we're trying to avoid. So the value of c depends on how much we value fairness to new users vs robustness to exploits.

Another proposal could be to assign starting currency 1/n, which would have an *infinite* sum, but which would go to infinity very very slowly. This is like increasing c slowly over time (more and more slowly as it approaches 1). My justification for this is than as the size of the market grows, it can handle more incoming newbies with bad opinions. Under this rule, we get 1 more unit of currency in the system for about every factor of 2.7 increase in the user base.

Rewards for Modelers

More importantly to the present discussion, how exactly do we reward people for adding models?

I've discussed how a model is scored as if it were present in the market from the beginning. I basically want to give that currency to the person who created the model. But there could be several different ways to do this. Here is my proposal:

The model has its own starting currency, based on its prior probability. This is new currency, not taken from anyone (one reason it's important to use a virtual currency). The model's creator doesn't have access to this currency; we want the model to have autonomy once created, so that it can continue to gain credibility based on evidence even if its creator loses faith in it. (In other words, we don't want to let people create models and then drain their accounts.)

But we *do* want to reward the creator. We do this by giving the creator extra money at the start of time, based on the prior probability of the model (IE based on the model's complexity). This extra money is *assumed to be traded by the creator, but in exactly the way the model prescribes.*

Once we get up to the current time, the model's creator is free to use that money as they wish.

Note that there is only a finite supply of model money to earn, since the prior probability sums to 1. However, finding the good models to reap those rewards will probably be pretty challenging even with automated enumeration of models.

Also note: we don't want to *un*reward everyone *else* for their contributions to the market. So we're not really going back in time and re-doing the entire market calculation as if a new trader was present. Instead, we just imagine what profit a model could have made if it had been present and if the market prices had all been the same. (This might lead to some problems, but I'm not thinking of any right now, so I'll leave this as my current proposal.)

It's possible that we also want to punish people for bad models, but I don't really think so -- it's simple enough to run a model before adding it to the market, to check if it does poorly, so an actual punishment for bad models could simply be avoided.

How do we compare human and virtual time?

A big implementation question is how to line up the timelines of (a) the actual human traders, and (b) the computation of the models.

One extreme would be to let models do all their computation up-front, at the "beginning of time". This is like trying to approximate Solomonoff induction rather than logical induction: we're doing nothing to encourage computationally simple models.

It seems like we instead want to allow models more time to improve their forecasts as we go. However, I don't see any clear principle to guide the trade-off. Human time and computational time seem incomparable.

This seems like a basic conceptual problem, and a solution might be quite interesting.

The Problem of Proofs

I was initially optimistic that, in an "arguments as models" paradigm, mathematical proofs could be handled as particularly strong arguments. I'm now less optimistic.

Imagine there's one big prediction market for mathematics, with MathBucks as the intellectual currency -- a professional is judged by their accumulation of MathBucks. The justification for this is supposed to be the same as the justification for using these markets in any other area: to efficiently pool all the information.

Now, proofs will of course be how outstanding bets get decided. But is there enough of an incentive to produce proofs? Is the reward for proofs close enough to the "true deserved intellectual credit"?

Certainly finding proofs will sometimes be quite profitable. If you prove something new, you could often make some money on the market by betting in favor of that conjecture before revealing your proof.

The problem, however, is that conjectures may be quite confident long before they're proven one way or the other. A 98% confident conjecture only enables a small profit to someone who proves it. I think it's fair to say that the size of the intellectual contribution may be much larger than the amount of credit earned.

This could be indirectly mitigated though bets on details which are hard to know without the proof in hand; for example, related conjectures likely to be decided by the proof, details of the method of proof, and so on. However, these provide no strong sense of proportional rewards for a contribution, and require more work for the contributor to profit.

At first, I thought a time-travel market would fix this. A proof could be turned into a model which bets before everyone.

The problem is that this is equivalent to a simple model which bets on the proposition at the beginning of time. This model could be claimed by someone long before a proof is found.

We could somehow promote the prior probability of correct proofs over other models. This would be somewhat useful, but feels like a hack which might not end up rewarding proof-writers to the desired extent.

Conclusion

The question I'm directly addressing in this post -- how to assign credit for intellectual labor -- is only part of my interest here. I'm also using this as a way to think about radical probabilism, particularly its relationship to models and transparent reasoning.

Philosophically, this is about the radical probabilist's response to the problem of old evidence. When do we require predictions to be made ahead of time for credit, and when do we give credit to predictions made after the fact? The answer proposed here is: we give credit for after-the-fact predictions to the extent that they're generated from simple models.

If this post seems to have the flavor of rough rules of thumb rather than solid mathematical epistemology, I'd say this is because we're in relatively unexplored territory here -- we've traveled from the old world of Bayes' Rule to the new continent of Radical Probabilism, and we're still at the stage of drawing rough maps with no proper scale. More formal versions of all of this seem within reach.

Probability vs Likelihood

This expands on some issues I mentioned in Radical Probabilism.

A rationality pet-nitpick of mine, which is shared by *almost no one,* is the probable/likely distinction. I got my introduction to Bayesian thinking, in part, from *Probabilistic Reasoning in Intelligent Systems* by Judea Pearl. In the book, Pearl makes a simple distinction between probability and likelihood which I find to be quite wonderful: the ***likelihood of X given Y*** is just the probability of Y given X!

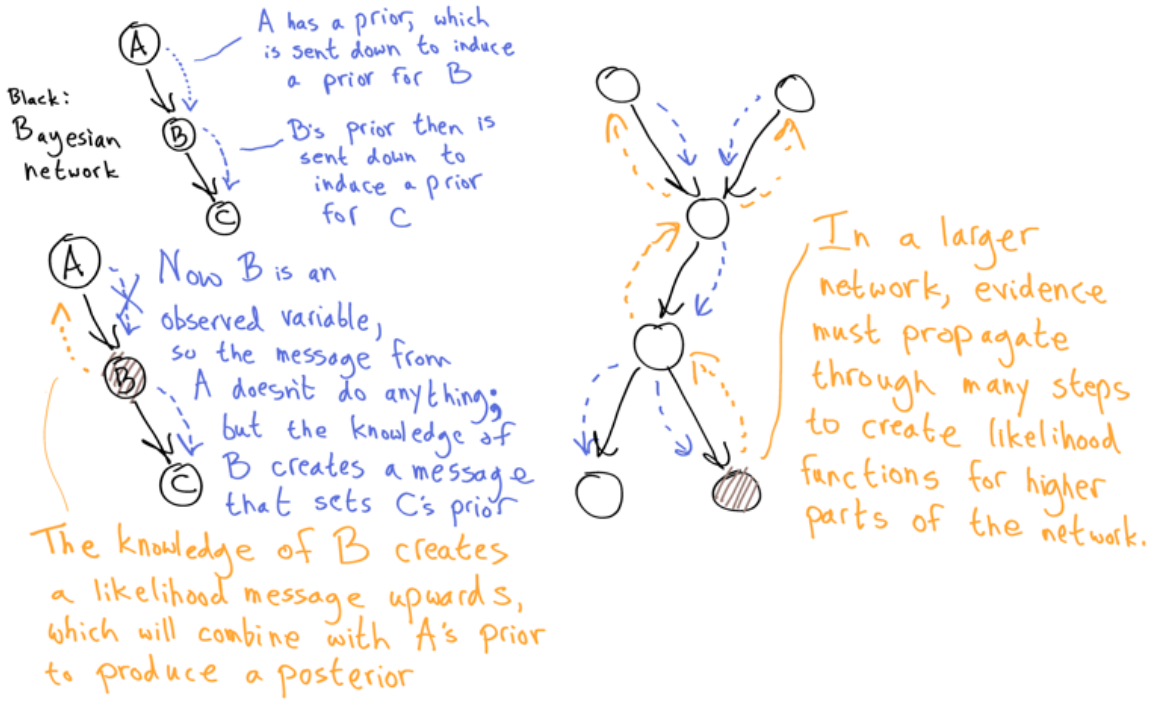
Probability of X given Y: $P(X|Y)$

Likelihood of X given Y: $L(X|Y) := P(Y|X)$

Why invent new terminology for something so simple?

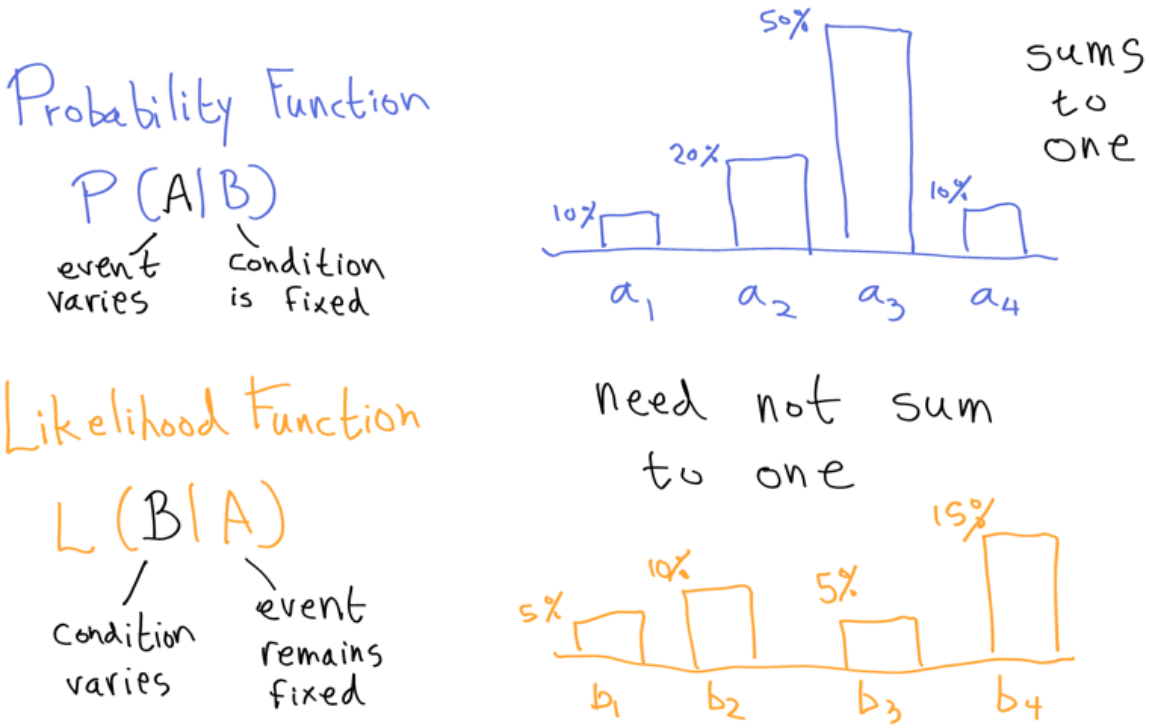
What we're basically doing here is making a *type* distinction between probability and likelihoods, to help us remember that *likelihoods can't be converted to probabilities without combining them with a prior.*

In the context of the book, it's because Bayesian networks pass *two types* of messages: "probability" type messages pass *prior*-like information *down* the Bayesian network, and "likelihood" type messages pass *evidence*-like information *up* the network.



At each individual node, we need to combine a prior with evidence in order to get a posterior. However, we only start with prior information for the top nodes of the network, and we only start with evidence information for the observed nodes. In order to get the information we need for other nodes, we have to propagate information around the network via the two types of messages.

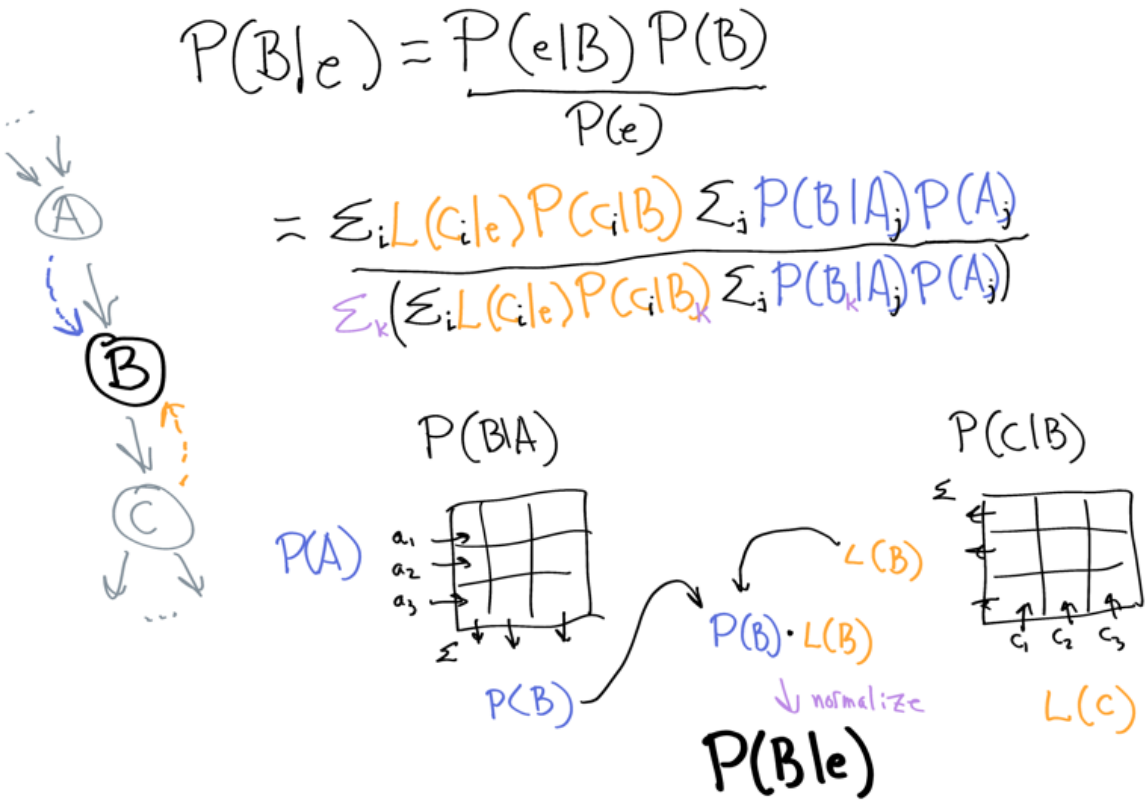
Another important aspect of the probability/likelihood distinction is that *we swap our views of what varies and what remains fixed.* When we regard P(A|B) as a probability *function,* rather than just a single probability judgement, we generally think of varying A. But when we use P(A|B) to construct a likelihood function for hypothesis testing, we think of A as the fixed evidence, and more readily vary B, which is now thought of as the hypothesis. Pearl's notation helps us track which thing we're holding steady vs which thing we're varying: we think of the first argument as what we're varying, and the second argument as what we're holding steady. Critically, *probability functions sum to 1, while likelihood functions need not do so.*



(Likelihood functions can sum to less than 1 if the event was improbable according to all the hypotheses, or more than 1 if the event was quite probable according to all the hypotheses.)

The downward messages in a Bayesian network are probability functions; the upward messages are likelihood functions. At each node, we combine these two pieces of information via Bayes' Law to get a posterior probability function. (Important note: the posterior probability *is not the information we then pass on*; this would create a double-counting problem, just like two people reinforcing each other's opinions and then each taking the strength of the other's opinion as *yet more* evidence in the same direction, even though no new evidence has entered the system. Think of the posteriors as private conclusions, people have reached, while their public messages only ever convey information which the recipient of that message *hasn't* seen yet.)

Here's the posterior calculation for one node:



I suspect it's quite intuitive how the node needs to transform the P(A) message through the P(B|A) matrix to get the P(A) it needs. What might be less intuitive is how the L(C) message is transformed in the same way through the P(C|B) matrix, to get L(B). B then sends the P(B) message down to node C, and the L(B) message up to node A.

You can see how the probability messages are transformed incrementally to become the right prior information needed to apply Bayes' Rule at a given node; and similarly, the likelihood messages are transformed incrementally to produce the right evidence information needed at a given node.

OK, so a likelihood/probability distinction helps us organize Bayes Net calculations. Are there any other reasons why we should take this distinction seriously?

Why It Matters

A major problem people have in applying Bayesian reasoning (at least in the kind of artificial example problem you might see as a brain teaser / textbook question / etc) is *neglecting the givens:* failing to use one or another of the numbers given to you in the problem which are needed to compute the Bayesian update. It's understandable; there can be a lot of numbers to keep track of. At least for me, it helps a lot to organize those numbers into probability functions and likelihood functions (which lets me think in terms of 2 vectors rather than 4+ numbers right there), with the symmetric picture above, where our goal is to combine prior and likelihood together to get a posterior. As I mentioned earlier, thinking of probability/likelihood as a *type* distinction helps avoid mistakes combining the wrong numbers: "a likelihood needs to combine with a prior before it can be regarded as a proper probability".

Base-Rate Neglect

I have a pet theory that some biases can be explained as a mix-up between probability and likelihood. (I don't know if this is a *good* explanation.) For example, base-rate neglect is just straightforwardly the mistake of treating a likelihood as a probability. A simple example of base-rate neglect would be to think someone is lying to you because they deny lying when you ask, just like a liar would. L(lying|response) is high, but this doesn't mean P(lying|response) is high.

Conjunction Fallacy

Another bias which seems to fit this pattern is the conjunction fallacy. Consider the classic example:

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*
>
> Which is more probable?
>
> 1. Linda is a bank teller.
> 2. Linda is a bank teller and is active in the feminist movement.

People rate "Linda is a bank teller" as less probable than "Linda is a bank teller and is active in the feminist movement". Notice that the second is *more likely* given everything we're told about Linda: L(teller & feminist | bio) > L(feminist | bio). So we can understand this as a probability/likelihood confusion -- indeed, it's just an example of base-rate neglect.

(Keep in mind there's a long list of failed attempted explanations of this human behavior, though, so don't rush to put a bunch of credence on mine.)

An optimistic theory is that a person makes these mistakes because we have a *thing in our head* roughly corresponding to likelihood, due to Bayes-net-like cognition. Then we have an *explicit concept* of probability. But if you don't also have an explicit concept of likelihood, the thing in your head corresponding to likelihood doesn't have anywhere good to land when you do explicit reasoning; so, it gets articulated as *probability* instead.

If so, having an explicit concept of likelihood should make these mistakes easier to avoid, as you can label the thing in your head correctly: you are experiencing the feeling of likelihood, *not* probability.

Using the Words

Practically everyone uses the terms probable/likely & probability/likelihood as interchangeable terms, except in very formal situations (when talking specifically about likelihood functions for hypothesis testing). I'm proposing that we make this distinction ***all the time***, for the good of our probabilistic reasoning skills. It's a *big ask!* This is a significantly different way of speaking than we're used to.

It's fortunate that these words really are interchangeable in English; at least I'm not taking away someone's ability to mean something. It's a simple matter of saying "probable"/"probably" in every case where you would have used any of the four words, and using "likely"/"likelihood" to indicate the thing which you would have previously labelled as base-rate-negligent probability. (Or do you call it something else? Frankly, I'm not sure how Bayesian people get by without a word for likelihood... *what do you call it?* How do you refer

to a theory which fits the data well? How do you say that adding details can make a story more plausible in the sense of explaining what we know, while making it less probable overall?)

So, to help ease the transition, I'll try to come up with a number of examples.

- A scenario is **unlikely** if it doesn't explain the evidence well. For example, in a COVID epidemic, if you get sick, the **prior probability** that it's COVID might be relatively high; but if your symptoms don't match COVID, it's **unlikely**; you can therefore conclude that the **probability** is lower.
- A scenario is **likely** if it explains the data well. For example, many conspiracy theories are **very likely** because they have an answer for every question: a powerful group is conspiring to cover up the truth, meaning that the evidence we see is exactly what they'd want us to see.
- You are trying to figure out who killed the butler. You find a kitchen knife with the butler's blood on it, placed behind a bush as if hastily hidden. It has the chef's fingerprints on it. Thus it's **likely** that the chef killed the butler, **raising the probability** of that hypothesis. However, another **similarly likely** hypothesis is that someone framed the chef by using gloves to wield a knife which the chef had touched for other reasons. Because the chef regularly touches the kitchen knives, this alternate hypothesis is relatively **probable.** Furthermore, because the chef has no motive, the chef being the murderer is overall **improbable**.
- You are studying the life of the early solar system. You have a set of four scenarios mutually exclusive and jointly exhaustive (or close enough to exhaustive that you're willing to treat it that way). Each scenario implies something about the distribution of orbits in the Kuiper belt. When you measure the Kuiper belt, you see something you were't expecting under any of your hypotheses. This is an **improbable** observation, which now makes all four hypotheses **unlikely**. However, some hypotheses are **more unlikely than others**. Because you didn't favor any of the hypotheses especially to begin with, your Bayesian update now strongly favors the **least unlikely** hypothesis, making that hypothesis **quite probable**. You begin to focus on the sub-scenarios within than scenario which make the observation you saw the *most* **probable** (that is, the **most likely** sub-scenarios). Out of these sub-scenarios, you narrow things down further by considering which scenarios are actually **probable**. You publish an analysis which takes all these considerations into account in order to report the **most probable scenarios** for the early solar system.
- You have several hypotheses about life on Venus. In order to test these hypotheses, you measure the chemical composition of the atmosphere of Venus. Unfortunately, your measurements are **very probable** under all of your hypotheses, so all of your hypotheses remain **likely**. Your best guess is still your **prior probability** on each hypothesis.

I found as I was writing the above examples that I wished there were *three* words, not two, so that I could more conveniently distinguish prior probability from posterior probability. It's very understandable that there's not, though, because one man's posterior is another man's prior -- these concepts flow into one another based on context. Likelihood is more clearly distinguished, since it does not sum to 1.*

Ambiguity & Context-Sensitivity

A given conversational context has *contextual assumptions*, information which we've already updated on, which gives us the *prior;* then we have information we're treating as *observations* in the context, which gives us our *likelihoods;* then we have the information we're treating as *unknowns* (hypotheses), which are what the prior and likelihoods are about. Whenever we say "probability" or "likelihood", the reader/listener has to infer from context what the assumptions/observations/unknowns are.

What's treated as assumption in one part of a conversation may be treated as unknown in another part; observations at one point may become assumptions later on; and so on. There is no strict rule that observations or assumptions are certain: as we see with a complex Bayesian network, we can get likelihood messages which propagate up from some distance away, which acts as uncertain evidence. Similarly, prior messages act as uncertain assumptions.

(None of the things we routinely treat as evidence are certain. If I say I read the number 98 off a thermometer, what I mean is that I'm quite confident it was 98. In Radical Probabilism, I attempted to convey Richard Jeffrey's view that *none* of our observations are certain.)

Nor does there need to be a strict temporal order behind these three categories. This will often be the case, when we treat the past as given, and update on present observations, to predict an unknown future. However, we can also try to explain the past (treating some elements of the present as assumptions, and others as unknowns), or uncover the past (treating it as unknown), or many other combinations.

You can think of a Bayes Net as a tool for keeping track of what constitutes assumptions/observations in a consistent way for a bunch of possible unknowns, so that we can split up our reasoning into a bunch of individual reasoning tasks that we can understand with this trichotomy.

Exchanging Virtual Evidence

One of Pearl's analogies for Bayesian Networks is a network of people trying to communicate about something. You can think of it as a network of experts. Each expert is in charge of one special variable, and the experts only talk to other experts if their specialties are closely related. If we take this analogy seriously, a striking fact is that *these people are communicating with virtual evidence*. (I discussed virtual evidence in my post on Radical Probabilism; the discussion there is a prerequisite for this section.)

Take the Bayes Net A->B->C. Expert B understands everything going on here, because B has to talk to everyone. However, A doesn't understand C at all, and vice versa; if they talked to each other, they would be mutually unintelligible. So, B has the job of translating between them. B accepts messages from C, takes just the information about B which the message implies, and passes that along to A (who understands B-talk). And similarly in the other direction.

All of the experts trust one another, but because they can't understand the ontology of distant evidence, they have to accept the virtual evidence implied by the likelihood functions they're handed by those around them: they accept "there was some evidence which induced this likelihood function", without worrying about whether the evidence is represented as a proposition.

So that's one of the core uses of virtual evidence: conveying information when there's trust of rationality, but no common language with which to describe all evidence.

(Again, one of Jeffrey's insights is that it's not important that the evidence is represented as a proposition *anywhere* in the network; we don't need to require every update to come with a crisp thing-which-we-updated-on. Virtual evidence is just a likelihood function which we don't attribute to any particular proposition.)**

Notice that for Bayesian networks, it's *much much more efficient* to pass information this way; if we instead attempted to convey all the information explicitly, and make our experts all understand each other, we would be back to calculating probabilities by taking a global exponential-sized sum over possibilities.

However, as I mentioned briefly earlier, we have to get things just right to avoid double-counting information when we use message-passing like this.

Propagating Posterior Probabilities

In Aumann's agreement theorem, Aumann famously considers the case of Bayesians passing information back and forth *only* by stating their new posterior probability after each communication they receive. Bayesians could possibly communicate this way in situations of incompatible ontologies, rather than using virtual evidence. But I'll argue it's not very efficient.

As I mentioned earlier, we want to avoid double-counting evidence. The outgoing messages in a Bayesian network are always (proportional to) the posterior probability of a node, *divided by* the *incoming* message along that same link. What this intuitively means is that when Alice talks to Bob, Alice has to "factor out" the information she got from Bob in the first place; she only wants to tell Bob her *other* information, which Bob *doesn't* know yet.

If Alice and Bob communicate with posterior probabilities, then it's always up to the *listener* to factor out *their own* message.

For example, suppose Alice tells Bob that it's a beautiful day outside. Bob nods. This tells Alice that Bob agrees: it's a beautiful day outside. But Bob is reporting honest posterior probabilities, so Bob's nod of assent *probably* just reflects his having updated on Alice. So Alice does not further update; Bob is not giving Alice any further evidence. If Bob had nodded *vigorously,* then Alice would interpret Bob's position as stronger even then hers; she would thus infer that Bob had independent evidence that it's a beautiful day outside. In *that* case, she would update.

The point is that in this situation, Alice has a somewhat difficult job interpreting Bob's message. Bob gives her his posterior, and she has to figure out what portion of that is new evidence for her.

Propagating Likelihoods

If Alice and Bob instead communicate *factoring out* each other's information, then they exchange *virtual updates* for each other, which are easy to take into account.

If Alice tells Bob "it's nice out" and Bob nods, Alice then knows that Bob has independent confirmation that it's nice out.

This is a much better case for Alice. She can form her posterior simply by multiplying in Bob's apparent likelihood function, and renormalizing.

Unfortunately, in realistic cases, we have a worst-case scenario: we don't really have one protocol. Whet someone expresses agreement or disagreement with a proposition, it's not really clear whether they're giving their all-things-considered posterior, or factoring out the information they got from us in the first place in order to give us a clean update.

So, realistically, when Bob nods, Alice doesn't have a great idea of what it means at all.

I know the examples with Alice trying to interpret Bob's nods are slightly absurd; Bob is *obviously* not giving Alice enough information there, so *of course* there are potential interpretation issues. But I'm using these examples because I think these ideas about virtual evidence and so on *are* mainly important in cases where there's not enough time to communicate fully.

Even if you're having a long, thorough discussion about some topic of interest, there will be many many small acts of communication with ambiguity and insufficient follow-up. Maybe an hour-long discussion on topic X involves 125 smaller propositions. 5 of these are major propositions, which get about 10 minutes each. Those 10-minute discussions involve about 20 smaller propositions each. These smaller propositions are given varying amounts of time, so many of them only get a few seconds.

So lots of propositions only get a brief response like a nod, or "sure", or one sentence.

In that brief amount of time, we form an impression of what each other think of that proposition. We make some small update one way or another, based on a vague impression, dozens or hundreds of times in an hour-long conversation.

I don't think it's psychologically or socially realistic for humans to *only* exchange virtual evidence. It's too socially useful to express positive sentiment when someone tells us something (reinforcing that we like them, we think they're reasonable, etc).

But I am interested in trying to do more little things to signal what type of information I'm giving off.

Using phrases like "all things considered," signals that I'm giving my posterior.

Using phrases like "if you hadn't told me that," signals that I'm factoring out what you told me.

If I say "I think that's very likely", it can mean that viewing what you told me as a hypothesis, it fits evidence in my possession well. Whereas "I think that's very probable" is more probably a posterior (although the ambiguity of whether "probable" stands for posterior or prior probability hurts us here).***

---

*: Not that I'd complain if someone came up with a decent proposal for a 3rd word.

**: There's a complication here. Notice that the messages in a Bayesian network are of two kinds: probability messages and likelihood messages. Yet I'm referring to them all as "virtual updates" and claiming that virtual updates are a kind of likelihood. In a Bayesian network, it's more natural to view some messages as probabilities, providing the priors for the local node posterior calculations. In a network of experts, it's more natural to think that every expert already has a prior when they start out, and so, only virtual evidence is communicated between the experts.

***: But it gets even more complicated than implied by (1), because as I argued, what's natural to think of as evidence/assumption/unknown will shift around in any given conversation, and must be inferred from context. I can't use "likely" to signal that I'm trying to convey virtual evidence to you if we're in a context where we're considering some hypothesis together, in light of some evidence; "likely" will sound like I mean something makes that evidence probable, rather than saying that *my* personal evidence makes *your statement* probable.

Reflective Bayesianism

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

I've argued in several places that traditional Bayesian reasoning is unable to properly handle embeddedness, logical uncertainty, and related issues.

However, in the light of hindsight, it's possible to imagine a Bayesian pondering these things, and "escaping the trap". This is because the trap was largely made out of assumptions we didn't usually *explicitly recognize* we were making.

Therefore, in order to aid moving beyond the traditional view, I think it's instructive to paint a detailed picture of what a traditional Bayesian might believe. This can be seen as a partner to my post Radical Probabilism, which explained in considerable detail how to move beyond the traditional Bayesian view.

Simple Belief vs Reflective Belief

There is an important distinction between the explicit dogmas of a view (ie, what adherents explicitly endorse) vs what one would need to believe in order to agree with them. One can become familiar with this distinction by studying logic, especially Gödel's incompleteness theorems and Tarski's undefinability theorem. In particular:

- Axiomatic systems such as set theory cannot proclaim themselves self-consistent (unless they are, indeed, inconsistent). However, adherents believe the system to be consistent. If they seriously doubted this, they would not use the system.
- Such axiomatic systems cannot discuss their own semantics (their map-territory correspondence). However, adherents regularly discuss this.

This becomes especially interesting when we're studying rationality (rather than, say, mathematics), because a theory of rationality is supposed to characterize normatively correct thinking. Yet, due to the above facts, philosophers will *typically* end up in a strange position where they endorse principles very different from the ones they themselves are using. For example, a philosopher arguing that set theory is the ultimate foundation of rational thought would find themselves in the strange position of utilizing "irrational" tools (tools which go beyond set theory) to come to its defense, EG in discussing the semantics of set theory or arguing for its consistency.[1]

I'll use the term ***simple belief*** to indicate accepting the explicit dogmas, and ***reflective belief*** to indicate accepting the meta-dogmas which justify the dogmas.[2,3] (These are not terms intended to stand the test of time; I'm not sure about the best terminology, and probably won't reliably use these particular terms outside of this blog post. Leverage Research, iiuc, uses the term "endorsed belief" for what I'm calling "reflective belief", and "belief" for what I'm calling "simple belief".)

Reflective belief and simple belief need not go together. There's a joke from Eliezer (h/t to Ben Pace for pointing me to the source):

> The rules say we must use consequentialism, but good people are deontologists, and virtue ethics is what actually works.
>
> —Eliezer Yudkowsky, Twitter

If this were true, then *simple belief* in consequentialism would imply *reflective belief* in virtue ethics (because you evaluate moral frameworks on their effects, not whether they're morally correct!). Similarly, simple belief in virtue ethics would imply reflective belief in deontology, and simple belief in deontology would imply reflective belief in consequentialism.

So, not only does simple belief in X not imply reflective belief in X; furthermore, reflective belief in X need not imply simple belief in X! (This is, indeed, belief-in-belief.)

Hence, by "reflective belief" I do not necessarily mean "reflectively consistent belief". ***Reflective consistency*** occurs only when simple belief and reflective belief are one and the same: the reasoning system you use is also the one you endorse.

Reflective Bayesianism

Applying the naive/reflective distinction to our case-in-point, I'll define two different types of Bayesian:

Simple Bayesian: simple belief in Bayesianism. This describes an agent who reasons according to the laws of probability theory, and updates beliefs via Bayes' Law. This is the type of reasoner which Bayesians *study.*

Reflective Bayesian: reflective belief in Bayesianism. For simplicity, I'll assume this *also* involves simple belief in Bayesianism. Realistically, Bayesian philosophers can't reason in a perfectly Bayesian way; so, this is a simplified, idealized Bayesian philosopher.

(A problem with the terminology in this section is that by "Bayesian" I specifically mean "dogmatic probabilism" in the terminology from my Radical Probabilism post. I don't want to construe "Bayesianism" to *necessarily* include Bayesian updates. The central belief of Bayesianism is subjective probability theory. However, repeating "simple dogmatic probabilism vs reflective dogmatic probabilism" over and over again in this essay was not very appealing.)

Actually, there's not one canonical "reflective belief in Bayes" -- one can justify Bayesianism in many ways, so there can correspondingly be many reflective-Bayesian positions. I'm going to discuss a number of these positions.

My prior is best.

The easiest way to reflectively endorse Bayesianism is to simply believe *my prior is best*. No other distribution can have more information about the world, unless it actually observes something about the world.

I think of this as *multiverse frequentism*. You think your prior literally gives the frequency of different possible universes. Now, I'm not accusing anyone of really believing this, but I *have* heard a (particularly reflective and self-critical) Bayesian

articulate the idea. And I think a lot of people have an assumption like this in mind when they think about priors like the Solomonoff prior, which are designed to be particularly "objective". This is essentially a map/territory error.

Some group of readers may think: "Now wait a minute. Shouldn't a Bayesian necessarily believe this? The expected Bayes loss of any other prior is going to be worse, when an agent considers them! Similarly, no other prior is going to look like a better tool for making decisions. So, yeah... I expect my prior to be best!"

On the other hand, those who endorse some level of modest epistemology might be giving that first group of readers some serious side-eye. Surely it's crazy to think your own beliefs are optimal *only* because they're yours?

To the first group, I would agree that *if* you've fully articulated a probability distribution, *then* you shouldn't be in a position where you think a different one is better than yours: in that case, you should update to the other one! (Or possibly to some third, even better distribution). The multiverse-frequentist fallaciously extends this result to apply to all priors.

But this *doesn't mean* you should think your distribution is best in general. For example, you can believe that someone else knows more than you, without knowing exactly what they believe.

In particular, it's easy to believe that *some computation* knows more than you. If a task somehow involves factoring numbers, you might not know the relevant prime factorizations. However, you can justifiably trust a probability distribution whose description includes running an accurate prime factorization algorithm. You can prefer to replace your own beliefs with such a probability distribution. This lays the groundwork for justified non-Bayesian updates.

I can't gain information without observing things.

Maybe our reflective Bayesian doesn't literally think theirs is the best prior possible. However, they might be a staunch empiricist: they believe knowledge is entanglement with reality, and you can only get entanglement with reality by *looking*.

Unlike the multiverse-frequentist described in the previous section, the empiricist *can* think other people have better probability distributions. What the empiricist *doesn't* believe is that we can emulate any of their expertise merely by thinking about it. Thinking is useless (except, of course, for the computational requirements of Bayes' Law itself). Therefore, although helpful non-Bayesian updates might technically be possible (eg, if you could morph into a more knowledgeable friend of yours), it's *not* possible to come up with any which you can implement just by thinking.

I can't think of any way to "justify" this assumption except if you really do have unbounded computational resources.

The best prior is already one of my hypotheses.

This is, of course, just the usual ***assumption of realizability:*** we assume that the world is in our hypothesis-space. We just have to find which of our hypotheses is the true one.

This doesn't imply as strong a rejection of non-Bayesian updates. It *could be* that we can gain some useful information by computation alone. However, the need for this must be limited, because *we already have enough computational power to simulate the whole world*.

What the assumption *does* gain you is a guarantee that you will make decisions well, eventually. If the correct hypothesis is in your hypothesis space, then once you learn it with sufficiently high confidence (which can usually happen pretty fast), you'll be making optimal decisions. This is a much stronger guarantee than the simple Bayesian has. So, the assumption does buy our Reflective Bayesian a lot of power in terms of justifying Bayesian reasoning.

My steel-man of this perspective is *the belief that the universe is intelligible*. I'm not sure what to call this belief. Here are a few versions of it:

- ***Computationalism.*** Everything is computable. It's absurd to imagine that anything physically realized would not be computable. So, it's sufficient to assume that the universe might be any computer program.
- ***Set-theory-ism.*** Anything real must have a description in ZFC. Physics might include some uncomputable aspects, but they'd be things like halting oracles, which can be captured within ZFC.
- ***Mathematicalism.*** Anything real must be mathematically describable. Not necessarily in any one axiom system such as ZFC -- we know (from Tarski's undefinability theorem) that there are mathematically describable things which fall outside any fixed axiom set. However, the universe must be mathematically describable in *some* sense.

I used to believe the third theory here. After all, what could it *possibly mean* to suppose the universe is *not* mathematically describable? Failing to assume this just seems like *giving up*.

But there is no *necessary law* saying the universe must be mathematical, any more than there's a necessary law saying the universe has to be computational. It *does* seem like we have *strong evidence* that the universe is mathematical in nature; mathematics has been *surprisingly helpful* for describing the universe we observe. However, philosophically, it makes more sense for this to be a contingent fact, not a necessary one.

There is a best hypothesis, out of those I can articulate.

The ***weak realizability assumption*** doesn't say that the universe *is* one of my hypotheses; instead, it postulates that *out of* my hypotheses, one of them is best. This is much more plausible, and gets us most of the theoretical implications.

For example, if you use the Solomonoff prior, strong realizability says that the universe is computable. Weak realizability just says that there's one computer program that's best for predicting the universe.

It makes a lot more sense to think of Solomonoff induction as *searching for the best computational way to predict.* The universe isn't necessarily computable, but computers *are*. If we're building AGI on computers, they can only use computable methods of predicting the world around them.

However, the assumption that one of your hypotheses is best is more questionable than you might realize. It's easy to set up circumstances in which no prior is best. My favorite example is a Bayesian who is observing coin-flips, and who has two hypotheses: that the coin is biased with 1/3rd probability of heads, and symmetrically, that it's biased with 1/3rd on tails. In truth, the coin is fair. We can show that the Bayesian will alternate between the two hypotheses forever: sometimes favoring one, sometimes the other.

The simple Bayesian believes that such non-convergence is possible. The reflective Bayesian thinks it is not possible -- one of the hypotheses has to be best, so beliefs cannot go back and forth forever.

The simple Bayesian therefore can reflectively prefer non-Bayesian updates -- for example, in the case of the fair coin, you'd be better off to converge to an even posterior over the two hypotheses, rather than continue updating via Bayes. (Or, even better, make an update which adds "fair coin" to the hypothesis set.)

I am calibrated, or can easily become calibrated.

Calibration is the property that of cases where you estimate, say, and 80% probability, the long-run frequency of those things happening is actually 80%. (Formally: for any number ϵ greater than zero, for any probability p, considering the sequence of all cases where you assign probability within ϵ of p, the *actual limiting frequency* of those things turning out to be true is within ϵ of p.)

Calibration is a lesser substitute for saying that a probabilistic hypothesis is "true", much like "best hypothesis out of the space" is. Or, flipping it around: being uncalibrated is a particularly egregious way for a hypothesis to be false.

To illustrate: if your sequence is 010101010101010101..., a fair coin is a *calibrated* model, even though there's a much better model. On the other hand, a biased coin is *not* a calibrated model. If we think the probability of "1" is 1/3, we will keep reporting a probability of 1/3, but the limiting frequency of the events will actually be 50%.

So, clearly, believing your probabilities to be calibrated is *a way* to reflectively endorse them, although not an extremely strong way.

I don't know that calibration implies any really *strong* defense of classical Bayesianism. However, it does provide somewhat stronger decision-theoretic guarantees. Namely, a calibrated estimate of the risks means that your strategy can't be outperformed by *really simple* adjustments. For example, if you're using a fair-coin model to make bets on the 010101010101... sequence, you will balance risks and rewards correctly (we can't make you do better by simply making you more/less risk-averse). The same cannot be said if you're using a biased-coin model.

I recently (in private correspondence) dealt with an example where calibration provided a stronger justification for Bayesian approaches over frequentist ones, but I feel the details would be a distraction here. In general I expect a calibration assumption helps justify Bayes in a lot of contexts.

A naive argument in favor of calibration might be "if I thought I weren't calibrated, I would adjust my beliefs to become more calibrated. Therefore, I must be calibrated." This makes two mistakes:

1. It's perfectly possible to believe "I'm not calibrated" without thinking your miscalibration is in a particular direction.
2. Even if that weren't the case, the argument that you'd want to correct your probabilities is by no means decisive. Correct away! This might be a legitimate non-Bayesian update.

My steelman of the calibration assumption is this: in general, it doesn't seem too hard to watch your calibration graph and adjust your reported probabilities in response. If you're an alien intelligence watching the 01010101... sequence, it might be hard to invent the "every other" hypothesis from scratch. However, it's easy to see that your P(1)=1/3 model is too low and should be adjusted upwards.

(OK, it's not hard at all to invent the "every other" pattern. But in more complicated cases, it's difficult to come up with a really new hypothesis, but it's relatively easy to improve the calibration on hypotheses.)

Note that the formal definition of "calibration" I gave at the beginning of this subsection doesn't really distinguish between "already calibrated" vs "will become calibrated at some point"; it's all asymptotic. So if we *can calibrate* by looking at a calibration chart and compensating for our over-/under- confidence, then we "are already calibrated" from a technical standpoint. (Nonetheless, I think the intuitive distinction is meaningful and useful.)

A counterargument to my steelman: it's actually computationally quite difficult to be calibrated. Sure, it doesn't seem so hard for humans to improve their calibration in practice, but the computational difficulty should give you pause. It might not make sense to suppose that humans are even approximately calibrated in general.

Conclusion

I think Bayesian philosophy before Radical Probabilism[4] over-estimated its self-consistency, underestimating the difference between simple Bayesianism and reflective Bayesianism (effectively making a map-territory error). It did so by implicitly making the mistakes above, as well as others. Sophisticated authors added technical assumptions such as calibration and realizability. These assumptions were then progressively forgotten through iterated summarization/popularization -- EG,

1. Original author includes technical assumptions as lynchpin of argumets.
2. Popular summary is careful to mention that the result is proved "under some technical assumptions", but doesn't include them.
3. The readers take away that the result is true, forgetting that technical assumptions were mentioned.

I think this happens all the time, with even the original authors possibly forgetting their own technical assumptions when they're not thinking hard about it.

Note that I'm not accusing anyone of literally believing the Reflective Bayesian positions I've outlined. (Actually, in particular, I want to avoid accusing *you...* some of my *other* readers, perhaps...) What I'm actually saying is that it was a belief operating in the background, heuristically influencing how people thought about things.[5]

For example:

A naive argument for the reflective consistency of UDT: "UDT just takes the actions which are optimal according to its prior. It can evaluate the expected utility of alternate policies by forward-sampling interactions from its prior. The actions which it indeed selects are going to be optimal, by definition. So, other policies look at best equally good. Therefore, it should never want to self-modify to become anything else."

I think most of the people who thought about UDT probably believed something like this at some point.

There are several important mistakes in this line of reasoning.

- It plays fast and loose with the distinction between what UDT does and what UDT expects itself to do. If UDT isn't logically omniscient, it may not know exactly what its own future actions are. Thus, it might prefer to make precommitments *even if* it worsens its strategy in doing so, because it might prefer to know with certainty that it will make a halfway decent selection, rather than stay in the dark about what it will do.
- It mistakes the "outer" expected value (the average value of sampling) with the "inner" expected value (the actual subjective value of an action, according to the agent). This is usually not a distinction we need to make, but in the face of logical uncertainty, subjective expectations can certainly be different from the expectations which can be calculated by averaging according to the prior.
- It implicitly assumes that alternate policies have no additional information about the environment, IE, their actions cannot be correlated with what happens. This reflects the "I can't get information without observing things" assumption. But in fact, it's possible to get better information by thinking longer, so UDT can prefer to self-modify into something which thinks longer and ends up with a totally different policy.

My overall point, here, is just that we should be careful about these things. Simple belief and reflective belief are not identical. A Bayesian reasoner does not necessarily prefer to keep being a Bayesian reasoner. And a Bayesian reasoner can prefer a non-Bayesian update to become a different Bayesian reasoner.

1. It's very possible to prefer a different probability distribution to your own. In particular, you'd usually like to update to use priors which are better-informed. This can be characterized as "thinking longer" in the cases where the better-informed prior is expressed as a computation which you can run.
2. It's similarly possible to prefer a different utility function to your own. There is no law of Bayesian reasoning which says that your utility function is best. The Gandhi murder pill thought experiment does illustrate an important fact, that agents will tend to protect themselves from arbitrary value-shifts. However, viewing some value shifts positively is totally allowed.

The goal of a Radical Probabilist should be to understand these non-Bayesian updates, trimming the notion of "rationality" to include only that which is essential.

Footnotes

1:

Truth and Paradox by Tim Maudlin is an extreme example of this; by the end of the book, Maudlin admits that what he is writing cannot be considered true on his own account. He proceeds to develop a theory of *permissible* assertions, which may not be true, but are normatively assertible. To top it off, he shows that no theory of permissibility can be satisfactory! He even refers to this as "defeat". Yet, he sees no better alternative, and so continues to justify his work as (mostly) permissible, though untrue.

2:

Note that although the simple/reflective distinction is inspired by rigorous formal ideas in logic, I'm not in fact taking a super formal approach here. Note the absence of a formal definition of "reflective belief". I think there are several different formal definitions one could give. I mean *any* of those. I consider my definition to include any reason why someone might argue for a position, perhaps even dishonestly (although dishonesty isn't relevant to the current discussion, and should probably be viewed as a borderline case).

3:

Aside: it's difficult to reliably maintain this distinction! When asserting things, are you asserting them simply or reflectively? Suppose I read Tim Maudlin's book (see footnote #1). What is "Tim Maudlin's position"? I can see good reasons to take it as (a) the explicit assertions, (b) the belief system which would endorse those explicit assertions, or (c) the belief system which the explicit assertions would themselves endorse.

In many circumstances, you'd say that what an author *reflectively* believes is their explicit assertions, and what they *simply* believe is the implicit belief system which leads them to make those assertions. Note what this implies: if you claim X, then your *simple* belief is the *reflective* belief in X, and your *reflective* position is *simple* belief in X! Headache-inducing, right?

But this often gets *more* confusing, not less, if (as in Tim Maudlin's case) the author starts explicitly dealing with these level distinctions. What should you think if I tell you I *simply* believe in X? I think it depends on how much you trust my introspective ability. If you don't trust it, then you'll conclude that I have belief-in-belief; I *endorse* simple belief in X (which is *probably* the same as endorsing X, ie, reflectively believing X). On the other hand, if you *do* trust my introspective ability, then you might take it to mean "I believe X, but I don't know why / I don't know whether I endorse my reasons for that belief". This is like the Leverage Research concept of "belief report".

This means you can take my assertion at face value: I've given you one of my simple beliefs.

But what if someone makes a *habit* of giving you their simple beliefs, rather than their reflective beliefs? This might be an honesty thing, or possibly an unreflective habit. Philosophers, academics, and smart people generally might be stuck in a rut of only giving reflective positions, because they're expecting to have to defend their assertions (and they like making defensible assertions). This calls into question whether/when we should assume that someone is giving us their reflective beliefs rather than their simple beliefs.

And what if I tell you I *reflectively* believe in X? Do you take *that* at face value? Or do you think I *reflectively reflectively* believe X (so my simple belief is Z, a position which reflectively endorses the position Y -- where Y is a position which reflectively endorses X).

... You can see where things get difficult.

4:

By "Bayesianism before radical probabilism" I don't mean a temporal/historic thing, EG, Bayesianism before the 1950s (when Jeffrey first began inventing Radical Probabilism). Rather, I mean "the version of Bayesianism which strongly weds itself to Bayesian updates." Most centrally, I'm referring to LessWrong before Logical Induction.

5:

Simply put, early LessWrong *reflectively believed* in (classical) Bayesianism, and thus *simply believed* the justifying assumptions associated with Bayesianism. But few, if any *reflectively* believed those assumptions -- indeed, those assumptions have little justification when examined, and life gets more interesting when assuming their negation.

The only general advice I can think of to avoid this mistake is "don't lose track of your assumptions".