Concepts in formal epistemology

1. Interpretations of "probability"
2. Correspondence visualizations for different interpretations of "probability"
3. Probability interpretations: Examples
4. Coherent decisions imply consistent utilities
5. A Semitechnical Introductory Dialogue on Solomonoff Induction

Interpretations of "probability"

(*Written for Arbital in 2016.*)

What does it *mean* to say that a flipped coin has a 50% probability of landing heads?

Historically, there are two popular types of answers to this question, the "frequentist" and "subjective" (aka "Bayesian") answers, which give rise to radically different approaches to experimental statistics. There is also a third "propensity" viewpoint which is largely discredited (assuming the coin is deterministic). Roughly, the three approaches answer the above question as follows:

- **The propensity interpretation:** Some probabilities are just out there in the world. It's a brute fact about coins that they come up heads half the time. When we flip a coin, it has a fundamental *propensity* of 0.5 for the coin to show heads. When we say the coin has a 50% probability of being heads, we're talking directly about this propensity.
- **The frequentist interpretation:** When we say the coin has a 50% probability of being heads after this flip, we mean that there's a class of events similar to this coin flip, and across that class, coins come up heads about half the time. That is, the *frequency* of the coin coming up heads is 50% inside the event class, which might be "all other times this particular coin has been tossed" or "all times that a similar coin has been tossed", and so on.
- **The subjective interpretation:** Uncertainty is in the mind, not the environment. If I flip a coin and slap it against my wrist, it's already landed either heads or tails. The fact that I don't know whether it landed heads or tails is a fact about me, not a fact about the coin. The claim "I think this coin is heads with probability 50%" is an *expression of my own ignorance,* and 50% probability means that I'd bet at 1 : 1 odds (or better) that the coin came up heads.

For a visualization of the differences between these three viewpoints, see Correspondence visualizations for different interpretations of "probability". For examples of the difference, see Probability interpretations: Examples. See also the Stanford Encyclopedia of Philosophy article on interpretations of probability.

The propensity view is perhaps the most intuitive view, as for many people, it just feels like the coin is intrinsically random. However, this view is difficult to reconcile with the idea that once we've flipped the coin, it has already landed heads or tails. If the event in question is decided deterministically, the propensity view can be seen as an instance of the mind projection fallacy: When we mentally consider the coin flip, it feels 50% likely to be heads, so we find it very easy to imagine a *world* in which the coin is *fundamentally* 50%-heads-ish. But that feeling is actually a fact about *us,* not a fact about the coin; and the coin has no physical 0.5-heads-propensity hidden in there somewhere — it's just a coin.

The other two interpretations are both self-consistent, and give rise to pragmatically different statistical techniques, and there has been much debate as to which is preferable. The subjective interpretation is more generally applicable, as it allows one to assign probabilities (interpreted as betting odds) to one-off events.

Frequentism vs subjectivism

As an example of the difference between frequentism and subjectivism, consider the question: "What is the probability that Hillary Clinton will win the 2016 US presidential election?", as analyzed in the summer of 2016.

A stereotypical (straw) frequentist would say, "The 2016 presidential election only happens once. We can't *observe* a frequency with which Clinton wins presidential elections. So we can't do any statistics or assign any probabilities here."

A stereotypical subjectivist would say: "Well, prediction markets tend to be pretty well-calibrated about this sort of thing, in the sense that when prediction markets assign 20% probability to an event, it happens around 1 time in 5. And the prediction markets are currently betting on Hillary at about 3 : 1 odds. Thus, I'm comfortable saying she has about a 75% chance of winning. If someone offered me 20 : 1 odds *against* Clinton — they get $1 if she loses, I get $20 if she wins — then I'd take the bet. I suppose you could refuse to take that bet on the grounds that you Just Can't Talk About Probabilities of One-off Events, but then you'd be pointlessly passing up a really good bet."

A stereotypical (non-straw) frequentist would reply: "I'd take that bet too, of course. But my taking that bet *is not based on rigorous epistemology,* and we shouldn't allow that sort of thinking in experimental science and other important venues. You can do subjective reasoning about probabilities when making bets, but we should exclude subjective reasoning in our scientific journals, and that's what frequentist statistics is designed for. Your paper should not conclude "and therefore, having observed thus-and-such data about carbon dioxide levels, I'd personally bet at 9 : 1 odds that anthropogenic global warming is real," because you can't build scientific consensus on opinions."

...and then it starts getting complicated. The subjectivist responds "First of all, I agree you shouldn't put posterior odds into papers, and second of all, it's not like your method is truly objective — the choice of "similar events" is arbitrary, abusable, and has given rise to p-hacking and the replication crisis." The frequentists say "well your choice of prior is even more subjective, and I'd like to see you do better in an environment where peer pressure pushes people to abuse statistics and exaggerate their results," and then down the rabbit hole we go.

The subjectivist interpretation of probability is common among artificial intelligence researchers (who often design computer systems that manipulate subjective probability distributions), Wall Street traders (who need to be able to make bets even in relatively unique situations), and common intuition (where people feel like they can say there's a 30% chance of rain tomorrow without worrying about the fact that tomorrow only happens once). Nevertheless, the frequentist interpretation is commonly taught in introductory statistics classes, and is the gold standard for most scientific journals.

A common frequentist stance is that it is virtuous to have a large toolbox of statistical tools at your disposal. Subjectivist tools have their place in that toolbox, but they don't deserve any particular primacy (and they aren't generally accepted when it comes time to publish in a scientific journal).

An aggressive subjectivist stance is that frequentists have invented some interesting tools, and many of them are useful, but that refusing to consider subjective probabilities is toxic. Frequentist statistics were invented in a (failed) attempt to keep subjectivity out of science in a time before humanity really understood the laws of probability theory. Now we have theorems about how to manage subjective probabilities correctly, and how to factor personal beliefs out from the objective evidence provided by the data, and if you ignore these theorems you'll get in trouble. The frequentist interpretation is broken, and that's why science has p-hacking and a replication crisis even as all the wall-street traders and AI scientists use the Bayesian interpretation. This "let's compromise and agree that everyone's viewpoint is valid" thing is all well and good, but how much worse do things need to get before we say "oops" and start acknowledging the subjective probability interpretation across all fields of science?

The most common stance among scientists and researchers is much more agnostic, along the lines of "use whatever statistical techniques work best at the time, and use frequentist techniques when publishing in journals because that's what everyone's been doing for decades upon decades upon decades, and that's what everyone's expecting."

See also Subjective probability and Likelihood functions, p-values, and the replication crisis.

Which interpretation is most useful?

Probably the subjective interpretation, because it subsumes the propensity and frequentist interpretations as special cases, while being more flexible than both.

When the frequentist "similar event" class is clear, the subjectivist can take those frequencies (often called base rates in this context) into account. But unlike the frequentist, she can also combine those base rates with other evidence that she's seen, and assign probabilities to one-off events, and make money in prediction markets and/or stock markets (when she knows something that the market doesn't).

When the laws of physics actually do "contain uncertainty", such as when they say that there are multiple different observations you might make next with differing likelihoods (as the Schrodinger equation often will), a subjectivist can combine her propensity-style uncertainty with her personal uncertainty in order to generate her aggregate subjective probabilities. But unlike a propensity theorist, she's not forced to think that *all* uncertainty is physical uncertainty: She can act like a propensity theorist with respect to Schrodinger-equation-induced uncertainty, while still believing that her uncertainty about a coin that has already been flipped and slapped against her wrist is in her head, rather than in the coin.

This fully general stance is consistent with the belief that frequentist tools are useful for answering frequentist questions: The fact that you can *personally* assign probabilities to one-off events (and, e.g., evaluate how good a certain trade is on a prediction market or a stock market) does not mean that tools labeled "Bayesian" are always better than tools labeled "frequentist". Whatever interpretation of "probability" you use, you're encouraged to use whatever statistical tool works best for you at any given time, regardless of what "camp" the tool comes from. Don't let the fact that you think it's possible to assign probabilities to one-off events prevent you from using useful frequentist tools!

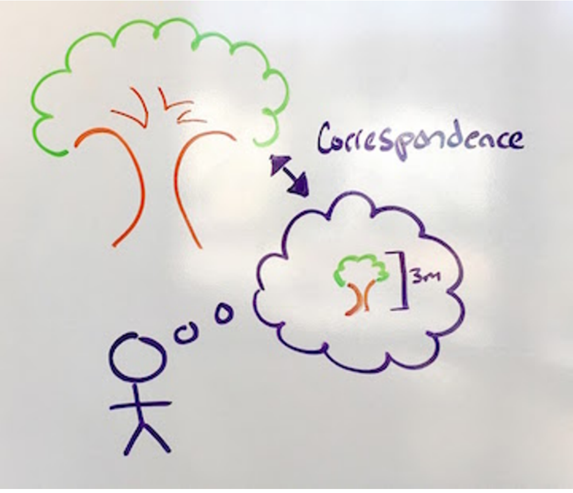
Correspondence visualizations for different interpretations of "probability"

(*Written for Arbital in 2016.*)

Recall that there are three common interpretations of what it means to say that a coin has a 50% probability of landing heads:

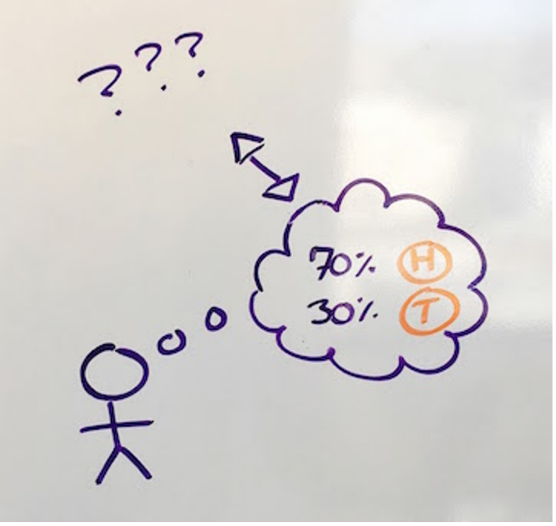
- **The propensity interpretation:** Some probabilities are just out there in the world. It's a brute fact about coins that they come up heads half the time; we'll call this the coin's physical "propensity towards heads." When we say the coin has a 50% probability of being heads, we're talking directly about this propensity.
- **The frequentist interpretation:** When we say the coin has a 50% probability of being heads after this flip, we mean that there's a class of events similar to this coin flip, and across that class, coins come up heads about half the time. That is, the *frequency* of the coin coming up heads is 50% inside the event class (which might be "all other times this particular coin has been tossed" or "all times that a similar coin has been tossed" etc).
- **The subjective interpretation:** Uncertainty is in the mind, not the environment. If I flip a coin and slap it against my wrist, it's already landed either heads or tails. The fact that I don't know whether it landed heads or tails is a fact about me, not a fact about the coin. The claim "I think this coin is heads with probability 50%" is an *expression of my own ignorance,* which means that I'd bet at 1 : 1 odds (or better) that the coin came up heads.

One way to visualize the difference between these approaches is by visualizing what they say about when a model of the world should count as a good model. If a person's model of the world is definite, then it's easy enough to tell whether or not their model is good or bad: We just check what it says against the facts. For example, if a person's model of the world says "the tree is 3m tall", then this model is correct if (and only if) the tree is 3 meters tall.

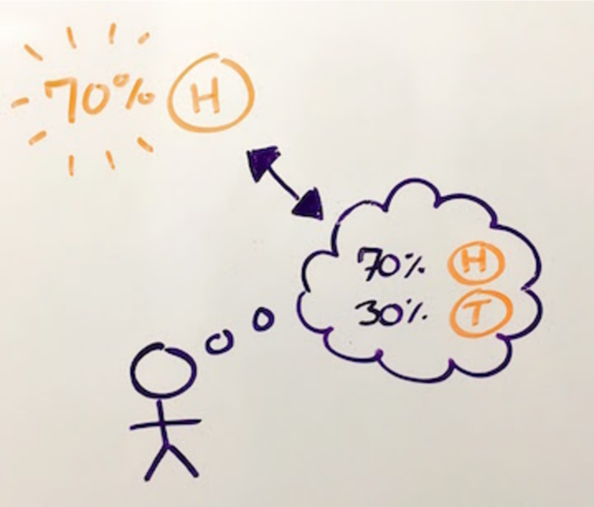


Definite claims in the model are called "true" when they correspond to reality, and "false" when they don't. If you want to navigate using a map, you had better ensure that the lines drawn on the map correspond to the territory.

But how do you draw a correspondence between a map and a territory when the map is probabilistic? If your model says that a biased coin has a 70% chance of coming up heads, what's the correspondence between your model and reality? If the coin is actually heads, was the model's claim true? 70% true? What would that mean?



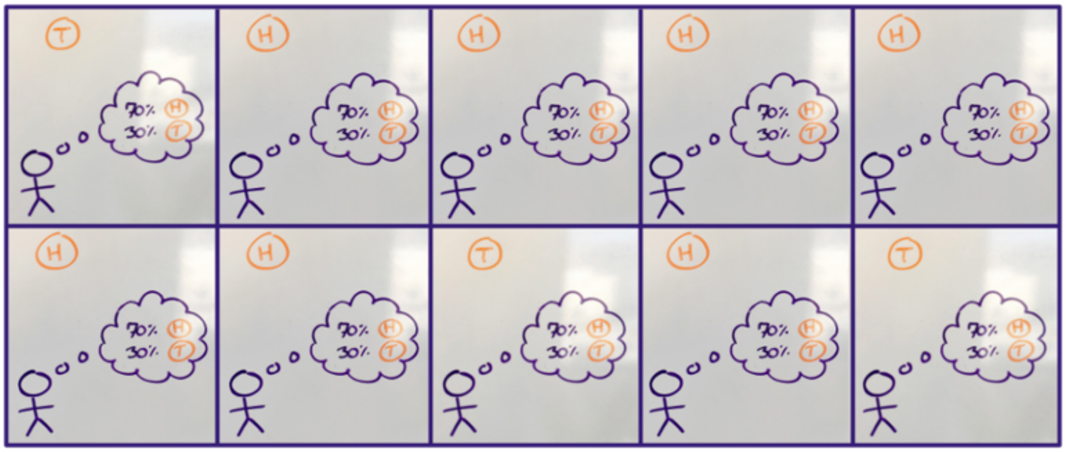
The advocate of **propensity** theory says that it's just a brute fact about the world that the world contains ontologically basic uncertainty. A model which says the coin is 70% likely to land heads is true if and only the actual physical propensity of the coin is 0.7 in favor of heads.



This interpretation is useful when the laws of physics *do* say that there are multiple different observations you may make next (with different likelihoods), as is sometimes the case (e.g., in quantum physics). However, when the event is deterministic — e.g., when it's a coin that has been tossed and slapped down and is already either heads or tails — then this view is largely regarded as foolish, and an example of the mind projection fallacy: The coin is just a coin, and has no special internal structure (nor special physical status) that makes it *fundamentally* contain a little 0.7 somewhere inside it. It's already either heads or tails, and while it may *feel* like the coin is fundamentally uncertain, that's a feature of your brain, not a feature of the coin.

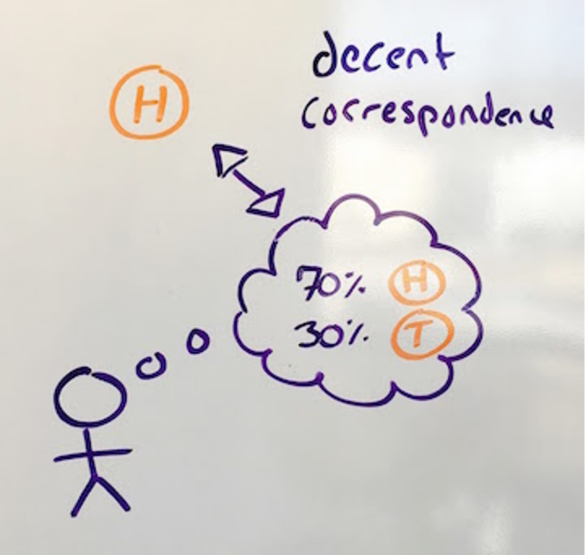
How, then, should we draw a correspondence between a probabilistic map and a deterministic territory (in which the coin is already definitely either heads or tails?)

A **frequentist** draws a correspondence between a single probability-statement in the model, and multiple events in reality. If the map says "that coin over there is 70% likely to be heads", and the actual territory contains 10 places where 10 maps say something similar, and in 7 of those 10 cases the coin is heads, then a frequentist says that the claim is true.



Thus, the frequentist preserves black-and-white correspondence: The model is either right or wrong, the 70% claim is either true or false. When the map says "That coin is 30% likely to be tails," that (according to a frequentist) means "look at all the cases similar to this case where my map says the coin is 30% likely to be tails; across all those places in the territory, 3/10ths of them have a tails-coin in them." That claim is definitive, given the set of "similar cases."

By contrast, a **subjectivist** generalizes the idea of "correctness" to allow for shades of gray. They say, "My uncertainty about the coin is a fact about *me,* not a fact about the coin; I don't need to point to other 'similar cases' in order to express uncertainty about *this* case. I know that the world right in front of me is either a heads-world or a tails-world, and I have a probability distribution that puts 70% probability on heads." They then draw a correspondence between their probability distribution and the world in front of them, and declare that the more probability their model assigns to the correct answer, the better their model is.



If the world *is* a heads-world, and the probabilistic map assigned 70% probability to "heads," then the subjectivist calls that map "70% accurate." If, across all cases where their map says something has 70% probability, the territory is actually that way 7/10ths of the time, then the Bayesian calls the map "well calibrated". They then seek methods to make their maps more accurate, and better calibrated. They don't see a need to interpret probabilistic maps as making definitive claims; they're happy to interpret them as making estimations that can be graded on a sliding scale of accuracy.

Debate

In short, the frequentist interpretation tries to find a way to say the model is definitively "true" or "false" (by identifying a collection of similar events), whereas the subjectivist interpretation extends the notion of "correctness" to allow for shades of gray.

Frequentists sometimes object to the subjectivist interpretation, saying that frequentist correspondence is the only type that has any hope of being truly objective. Under Bayesian correspondence, who can say whether the map should say 70% or 75%, given that the probabilistic claim is not objectively true or false either way? They claim that these subjective assessments of "partial accuracy" may be intuitively satisfying, but they have no place in science. Scientific reports ought to be restricted to frequentist statements, which are definitively either true or false, in order to increase the objectivity of science.

Subjectivists reply that the frequentist approach is hardly objective, as it depends entirely on the choice of "similar cases". In practice, people can (and do!) abuse frequentist statistics by choosing the class of similar cases that makes their result look as impressive as possible (a technique known as "p-hacking"). Furthermore, the manipulation of subjective probabilities is subject to the iron laws of probability theory (which are the only way to avoid inconsistencies and pathologies when managing your uncertainty about the world), so it's not like subjective probabilities are the wild west or something. Also, science has things to say about situations even when there isn't a huge class of objective frequencies we can observe, and science should let us collect and analyze evidence even then.

For more on this debate, see Likelihood functions, p-values, and the replication crisis.

Probability interpretations: Examples

(*Written for Arbital in 2016.*)

Betting on one-time events

Consider evaluating, in June of 2016, the question: "What is the probability of Hillary Clinton winning the 2016 US presidential election?"

On the **propensity** view, Hillary has some fundamental chance of winning the election. To ask about the probability is to ask about this objective chance. If we see a prediction market in which prices move after each new poll — so that it says 60% one day, and 80% a week later — then clearly the prediction market isn't giving us very strong information about this objective chance, since it doesn't seem very likely that Clinton's *real* chance of winning is swinging so rapidly.

On the **frequentist** view, we cannot formally or rigorously say anything about the 2016 presidential election, because it only happens once. We can't *observe* a frequency with which Clinton wins presidential elections. A frequentist might concede that they would cheerfully buy for $1 a ticket that pays $20 if Clinton wins, considering this a favorable bet in an *informal* sense, while insisting that this sort of reasoning isn't sufficiently rigorous, and therefore isn't suitable for being included in science journals.

On the **subjective** view, saying that Hillary has an 80% chance of winning the election summarizes our *knowledge about* the election or our *state of uncertainty* given what we currently know. It makes sense for the prediction market prices to change in response to new polls, because our current state of knowledge is changing.

A coin with an unknown bias

Suppose we have a coin, weighted so that it lands heads somewhere between 0% and 100% of the time, but we don't know the coin's actual bias.

The coin is then flipped three times where we can see it. It comes up heads twice, and tails once: HHT.

The coin is then flipped again, where nobody can see it yet. An honest and trustworthy experimenter lets you spin a wheel-of-gambling-odds — reducing the worry that the experimenter might know more about the coin than you, and be offering you a deliberately rigged bet — and the wheel lands on (2 : 1). The experimenter asks if you'd enter into a gamble where you win $2 if the unseen coin flip is tails, and pay $1 if the unseen coin flip is heads.

On a **propensity** view, the coin has some objective probability between 0 and 1 of being heads, but we just don't know what this probability is. Seeing HHT tells us that the coin isn't all-heads or all-tails, but we're still just guessing — we don't really know the answer, and can't say whether the bet is a fair bet.

On a **frequentist** view, the coin would (if flipped repeatedly) produce some long-run frequency f of heads that is between 0 and 1. If we kept flipping the coin long enough, the actual proportion p of observed heads is guaranteed to approach f arbitrarily closely, eventually. We can't say that the *next* coin flip is guaranteed to be H or T, but we can make an objectively true statement that p will approach f to within epsilon if we continue to flip the coin long enough.

To decide whether or not to take the bet, a frequentist might try to apply an unbiased estimator to the data we have so far. An "unbiased estimator" is a rule for taking an observation and producing an estimate e of f, such that the expected value of e is f. In other words, a frequentist wants a rule such that, if the hidden bias of the coin was in fact to yield 75% heads, and we repeat many times the operation of flipping the coin a few times and then asking a new frequentist to estimate the coin's bias using this rule, the *average* value of the estimated bias will be 0.75. This is a property of the *estimation rule* which is objective. We can't hope for a rule that will always, in any particular case, yield the true f from just a few coin flips; but we can have a rule which will provably have an *average* estimate of f, if the experiment is repeated many times.

In this case, a simple unbiased estimator is to guess that the coin's bias f is equal to the observed proportion of heads, or 2/3. In other words, if we repeat this experiment many many times, and whenever we see p heads in 3 tosses we guess that the coin's bias is $\frac{p}{3}$, then this rule definitely is an unbiased estimator. This estimator says that a bet of $2 vs. $1 is fair, meaning that it doesn't yield an expected profit, so we have no reason to take the bet.

On a **subjectivist** view, we start out personally unsure of where the bias f lies within the interval [0, 1]. Unless we have any knowledge or suspicion leading us to think otherwise, the coin is just as likely to have a bias between 33% and 34%, as to have a bias between 66% and 67%; there's no reason to think it's more likely to be in one range or the other.

Each coin flip we see is then evidence about the value of f, since a flip H happens with different probabilities depending on the different values of f, and we update our beliefs about f using Bayes' rule. For example, H is twice as likely if f = $\frac{2}{3}$ than if f = $\frac{1}{3}$ so by Bayes's Rule we should now think f is twice as likely to lie near $\frac{2}{3}$ as it is to lie near $\frac{1}{3}$.

When we start with a uniform prior, observe multiple flips of a coin with an unknown bias, see M heads and N tails, and then try to estimate the odds of the next flip coming up heads, the result is Laplace's Rule of Succession which estimates $(M + 1) : (N + 1)$ for a probability of $\frac{M+1}{M+N+2}$.

In this case, after observing HHT, we estimate odds of $2 : 3$ for tails vs. heads on the next flip. This makes a gamble that wins $2 on tails and loses $1 on heads a profitable gamble in expectation, so we take the bet.

Our choice of a uniform prior over f was a little dubious — it's the obvious way to express total ignorance about the bias of the coin, but obviousness isn't everything. (For example, maybe we actually believe that a fair coin is more likely than a coin biased 50.0000023% towards heads.) However, all the reasoning after the choice of prior was rigorous according to the laws of probability theory, which is the only method of manipulating quantified uncertainty that obeys obvious-seeming rules about how subjective uncertainty should behave.

Probability that the 98,765th decimal digit of π is 0

What is the probability that the 98,765th digit in the decimal expansion of π is 0?

The **propensity** and **frequentist** views regard as nonsense the notion that we could talk about the *probability* of a mathematical fact. Either the 98,765th decimal digit of π is 0 or it's not. If we're running *repeated* experiments with a random number generator, and looking at different digits of π, then it might make sense to say that the random number generator has a 10% probability of picking numbers whose corresponding decimal digit of π is 0. But if we're just picking a non-random number like 98,765, there's no sense in which we could say that the 98,765th digit of π has a 10% propensity to be 0, or that this digit is 0 with 10% frequency in the long run.

The **subjectivist** considers probabilities to just refer to their own uncertainty. So if a subjectivist has picked the number 98,765 without yet knowing the corresponding digit of π, and hasn't made any observation that is known to them to be entangled with the 98,765th digit of π, and they're pretty sure their friend hasn't yet looked up the 98,765th digit of π either, and their friend offers a whimsical gamble that costs $1 if the digit is non-zero and pays $20 if the digit is zero, the Bayesian takes the bet.

Note that this demonstrates a difference between the subjectivist interpretation of "probability" and Bayesian probability theory. A perfect Bayesian reasoner that knows the rules of logic and the definition of π must, by the axioms of probability theory, assign probability either 0 or 1 to the claim "the 98,765th digit of π is a 0" (depending on whether or not it is). This is one of the reasons why perfect Bayesian reasoning is intractable. A subjectivist that is not a perfect Bayesian nevertheless claims that they are personally uncertain about the value of the 98,765th digit of π. Formalizing the rules of subjective probabilities about mathematical facts (in the way that probability theory formalized the rules for manipulating subjective probabilities about empirical facts, such as which way a coin came up) is an open problem; this in known as the problem of logical uncertainty.

Coherent decisions imply consistent utilities

(*Written for Arbital in 2017.*)

Introduction to the introduction: Why expected utility?

So we're talking about how to make good decisions, or the idea of 'bounded rationality', or what sufficiently advanced Artificial Intelligences might be like; and somebody starts dragging up the concepts of 'expected utility' or 'utility functions'.

And before we even ask what those are, we might first ask, *Why?*

There's a mathematical formalism, 'expected utility', that some people invented to talk about making decisions. This formalism is very academically popular, and appears in all the textbooks.

But so what? Why is that *necessarily* the best way of making decisions under every kind of circumstance? Why would an Artificial Intelligence care what's academically popular? Maybe there's some better way of thinking about rational agency? Heck, why is this formalism popular in the first place?

We can ask the same kinds of questions about probability theory:

> Okay, we have this mathematical formalism in which the chance that X happens, aka $P(X)$, plus the chance that X doesn't happen, aka $P(\neg X)$, must be represented in a way that makes the two quantities sum to unity: $P(X) + P(\neg X) = 1$.

That formalism for probability has some neat mathematical properties. But so what? Why should the best way of reasoning about a messy, uncertain world have neat properties? Why shouldn't an agent reason about 'how likely is that' using something completely unlike probabilities? How do you *know* a sufficiently advanced Artificial Intelligence would reason in probabilities? You haven't seen an AI, so what do you think you know and how do you think you know it?

That entirely reasonable question is what this introduction tries to answer. There are, indeed, excellent reasons beyond academic habit and mathematical convenience for why we would by default invoke 'expected utility' and 'probability theory' to think about good human decisions, talk about rational agency, or reason about sufficiently advanced AIs.

The broad form of the answer seems easier to show than to tell, so we'll just plunge straight in.

Why not circular preferences?

De gustibus non est disputandum, goes the proverb; matters of taste cannot be disputed. If I like onions on my pizza and you like pineapple, it's not that one of us is right and one of us is wrong. We just prefer different pizza toppings.

Well, but suppose I declare to you that I *simultaneously*:

- Prefer onions to pineapple on my pizza.
- Prefer pineapple to mushrooms on my pizza.
- Prefer mushrooms to onions on my pizza.

If we use $>_P$ to denote my pizza preferences, with $X >_P Y$ denoting that I prefer X to Y, then I am declaring:

$$\text{onions} >_P \text{pineapple} >_P \text{mushrooms} >_P \text{onions}$$

That sounds strange, to be sure. But is there anything *wrong* with that? Can we disputandum it?

We used the math symbol $>$ which denotes an ordering. If we ask whether $>_P$ can be an ordering, it naughtily violates the standard transitivity axiom $x > y,\ y > z \implies x > z$.

Okay, so then maybe we shouldn't have used the symbol $>_P$ or called it an ordering. Why is that necessarily bad?

We can try to imagine each pizza as having a numerical score denoting how much I like it. In that case, there's no way we could assign consistent numbers x, y, z to those three pizza toppings such that $x > y > z > x$.

So maybe I don't assign numbers to my pizza. Why is that so awful?

Are there any grounds besides "we like a certain mathematical formalism and your choices don't fit into our math," on which to criticize my three simultaneous preferences?

(Feel free to try to answer this yourself before continuing...)

Click here to reveal and continue:

Suppose I tell you that I prefer pineapple to mushrooms on my pizza. Suppose you're about to give me a slice of mushroom pizza; but by paying one penny ($0.01) I can instead get a slice of pineapple pizza (which is just as fresh from the oven). It seems realistic to say that most people with a pineapple pizza preference would probably pay the penny, if they happened to have a penny in their pocket.[1]

After I pay the penny, though, and just before I'm about to get the pineapple pizza, you offer me a slice of onion pizza instead—no charge for the change! If I was telling the truth about preferring onion pizza to pineapple, I should certainly accept the substitution if it's free.

And then to round out the day, you offer me a mushroom pizza instead of the onion pizza, and again, since I prefer mushrooms to onions, I accept the swap.

I end up with exactly the same slice of mushroom pizza I started with... and one penny poorer, because I previously paid $0.01 to swap mushrooms for pineapple.

This seems like a *qualitatively* bad behavior on my part. By virtue of my incoherent preferences which cannot be given a consistent ordering, I have shot myself in the foot, done something self-defeating. We haven't said *how* I ought to sort out my inconsistent preferences. But no matter how it shakes out, it seems like there must be *some* better alternative—some better way I could reason that wouldn't spend a penny to go in circles. That is, I could at least have kept my original pizza slice and not spent the penny.

In a phrase you're going to keep hearing, I have executed a 'dominated strategy': there exists some other strategy that does strictly better.[2]

Or as Steve Omohundro put it: If you prefer being in Berkeley to being in San Francisco; prefer being in San Jose to being in Berkeley; and prefer being in San Francisco to being in San Jose; then you're going to waste a lot of time on taxi rides.

None of this reasoning has told us that a non-self-defeating agent must prefer Berkeley to San Francisco or vice versa. There are at least six possible consistent orderings over pizza toppings, like mushroom $>_P$ pineapple $>_P$ onion etcetera, and *any* consistent ordering would avoid paying to go in circles.[3] We have not, in this argument, used pure logic to derive that pineapple pizza must taste better than mushroom pizza to an ideal rational agent. But we've seen that eliminating a certain kind of shoot-yourself-in-the-foot behavior, corresponds to imposing a certain *coherence* or *consistency* requirement on whatever preferences are there.

It turns out that this is just one instance of a large family of *coherence theorems* which all end up pointing at the same set of core properties. All roads lead to Rome, and all the roads say, "If you are not shooting yourself in the foot in sense X, we can view you as having coherence property Y."

There are some caveats to this general idea.

For example: In complicated problems, perfect coherence is usually impossible to compute—it's just too expensive to consider *all* the possibilities.

But there are also caveats to the caveats! For example, it may be that if there's a powerful machine intelligence that is not *visibly to us humans* shooting itself in the foot in way X, then *from our perspective* it must look like the AI has coherence property Y. If there's some sense in which the machine intelligence is going in circles, because *not* going in circles is too hard to compute, well, *we* won't see that either with our tiny human brains. In which case it may make sense, from our perspective, to think about the machine intelligence *as if* it has some coherent preference ordering.

We are not going to go through all the coherence theorems in this introduction. They form a very large family; some of them are a *lot* more mathematically intimidating; and honestly I don't know even 5% of the variants.

But we can hopefully walk through enough coherence theorems to at least start to see the reasoning behind, "Why expected utility?" And, because the two are a package deal, "Why probability?"

Human lives, mere dollars, and coherent trades

An experiment in 2000—from a paper titled "The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals"—asked subjects to consider the dilemma of a hospital administrator named Robert:

> Robert can save the life of Johnny, a five year old who needs a liver transplant, but the transplant procedure will cost the hospital $1,000,000 that could be spent in other ways, such as purchasing better equipment and enhancing salaries to recruit talented doctors to the hospital. Johnny is very ill and has been on the waiting list for a transplant but because of the shortage of local organ donors, obtaining a liver will be expensive. Robert could save Johnny's life, or he could use the $1,000,000 for other hospital needs.

The main experimental result was that most subjects got angry at Robert for even considering the question.

After all, you can't put a dollar value on a human life, right?

But better hospital equipment also saves lives, or at least one hopes so.[4] It's not like the other potential use of the money saves zero lives.

Let's say that Robert has a total budget of $100,000,000 and is faced with a long list of options such as these:

- $100,000 for a new dialysis machine, which will save 3 lives
- $1,000,000 for a liver for Johnny, which will save 1 life
- $10,000 to train the nurses on proper hygiene when inserting central lines, which will save an expected 100 lives
- ...

Now suppose—this is a supposition we'll need for our theorem—that Robert *does not care at all about money,* not even a tiny bit. Robert *only* cares about maximizing the total number of lives saved. Furthermore, we suppose for now that Robert cares about every human life equally.

If Robert does save as many lives as possible, given his bounded money, then Robert must *behave like* somebody assigning some consistent dollar value to saving a human life.

We should be able to look down the long list of options that Robert took and didn't take, and say, e.g., "Oh, Robert took all the options that saved more than 1 life per $500,000 and rejected all options that saved less than 1 life per $500,000; so Robert's behavior is *consistent* with his spending $500,000 per life."

Alternatively, if we can't view Robert's behavior as being coherent in this sense—if we cannot make up *any* dollar value of a human life, such that Robert's choices are consistent with that dollar value—then it must be possible to move around the same amount of money, in a way that saves more lives.

We start from the qualitative criterion, "Robert must save as many lives as possible; it shouldn't be possible to move around the same money to save more lives." We end up with the quantitative coherence theorem, "It must be possible to view Robert as trading dollars for lives at a consistent price."

We haven't proven that dollars have some intrinsic worth that trades off against the intrinsic worth of a human life. By hypothesis, Robert doesn't care about money at all. It's just that every dollar has an *opportunity cost* in lives it could have saved if deployed differently; and this opportunity cost is the same for every dollar because money is fungible.

An important caveat to this theorem is that there may be, e.g., an option that saves a hundred thousand lives for $200,000,000. But Robert only has $100,000,000 to spend. In this case, Robert may fail to take that option even though it saves 1 life per $2,000. It was a good option, but Robert didn't have enough money in the bank to afford it. This does mess up the elegance of being able to say, "Robert must have taken *all* the options saving at least 1 life per $500,000", and instead we can only say this with respect to options that are in some sense small enough or granular enough.

Similarly, if an option costs $5,000,000 to save 15 lives, but Robert only has $4,000,000 left over after taking all his other best opportunities, Robert's last selected option might be to save 8 lives for $4,000,000 instead. This again messes up the elegance of the reasoning, but Robert is still doing exactly what an agent *would* do if it consistently valued lives at 1 life per $500,000—it would buy all the best options *it could afford* that purchased at least that many lives per dollar. So that part of the theorem's conclusion still holds.

Another caveat is that we haven't proven that there's some specific dollar value in Robert's head, as a matter of psychology. We've only proven that Robert's outward behavior can be *viewed as if* it prices lives at *some* consistent value, assuming Robert saves as many lives as possible.

It could be that Robert accepts every option that spends less than $500,000/life and rejects every option that spends over $600,000, and there aren't any available options in the middle. Then

Robert's behavior can equally be *viewed as* consistent with a price of $510,000 or a price of $590,000. This helps show that we haven't proven anything about Robert explicitly *thinking* of some number. Maybe Robert never lets himself think of a specific threshold value, because it would be taboo to assign a dollar value to human life; and instead Robert just fiddles the choices until he can't see how to save any more lives.

We naturally have not proved by pure logic that Robert must want, in the first place, to save as many lives as possible. Even if Robert is a good person, this doesn't follow. Maybe Robert values a 10-year-old's life at 5 times the value of a 70-year-old's life, so that Robert will sacrifice five grandparents to save one 10-year-old. A lot of people would see that as entirely consistent with valuing human life in general.

Let's consider that last idea more thoroughly. If Robert considers a preteen equally valuable with 5 grandparents, so that Robert will shift $100,000 from saving 8 old people to saving 2 children, then we can no longer say that Robert wants to save as many 'lives' as possible. That last decision would decrease by 6 the total number of 'lives' saved. So we can no longer say that there's a qualitative criterion, 'Save as many lives as possible', that produces the quantitative coherence requirement, 'trade dollars for lives at a consistent rate'.

Does this mean that coherence might as well go out the window, so far as Robert's behavior is concerned? Anything goes, now? Just spend money wherever?

"Hm," you might think. "But... if Robert trades 8 old people for 2 children *here*... and then trades 1 child for 2 old people *there*..."

To reduce distraction, let's make this problem be about apples and oranges instead. Suppose:

- Alice starts with 8 apples and 1 orange.
- Then Alice trades 8 apples for 2 oranges.
- Then Alice trades away 1 orange for 2 apples.
- Finally, Alice trades another orange for 3 apples.

Then in this example, Alice is using a strategy that's *strictly dominated* across all categories of fruit. Alice ends up with 5 apples and one orange, but could've ended with 8 apples and one orange (by not making any trades at all). Regardless of the *relative* value of apples and oranges, Alice's strategy is doing *qualitatively* worse than another possible strategy, if apples have any positive value to her at all.

So the fact that Alice can't be viewed as having any coherent relative value for apples and oranges, corresponds to her ending up with qualitatively less of some category of fruit (without any corresponding gains elsewhere).

This remains true if we introduce more kinds of fruit into the problem. Let's say the set of fruits Alice can trade includes {apples, oranges, strawberries, plums}. If we can't look at Alice's trades and make up some relative quantitative values of fruit, such that Alice could be trading consistently with respect to those values, then Alice's trading strategy must have been dominated by some other strategy that would have ended up with strictly more fruit across all categories.

In other words, we need to be able to look at Alice's trades, and say something like:

"Maybe Alice values an orange at 2 apples, a strawberry at 0.1 apples, and a plum at 0.5 apples. That would explain why Alice was willing to trade 4 strawberries for a plum, but not willing to trade 40 strawberries for an orange and an apple."

And if we *can't* say this, then there must be some way to rearrange Alice's trades and get *strictly more fruit across all categories* in the sense that, e.g., we end with the same number of plums and apples, but one more orange and two more strawberries. This is a bad thing if Alice *qualitatively* values fruit from each category—prefers having more fruit to less fruit, ceteris paribus, for each category of fruit.

Now let's shift our attention back to Robert the hospital administrator. *Either* we can view Robert as consistently assigning some *relative* value of life for 10-year-olds vs. 70-year-olds, *or* there must be a way to rearrange Robert's expenditures to save either strictly more 10-year-olds or strictly more 70-year-olds. The same logic applies if we add 50-year-olds to the mix. We must be able to say something like, "Robert is consistently behaving as if a 50-year-old is worth a third of a ten-year-old". If we *can't* say that, Robert must be behaving in a way that pointlessly discards some saveable lives in some category.

Or perhaps Robert is behaving in a way which implies that 10-year-old girls are worth more than 10-year-old boys. But then the relative values of those subclasses of 10-year-olds need to be viewable as consistent; or else Robert must be qualitatively failing to save one more 10-year-old boy than could've been saved otherwise.

If you can denominate apples in oranges, and price oranges in plums, and trade off plums for strawberries, all at consistent rates... then you might as well take it one step further, and factor out an abstract unit for ease of notation.

Let's call this unit *1 utilon,* and denote it €1. (As we'll see later, the letters 'EU' are appropriate here.)

If we say that apples are worth €1, oranges are worth €2, and plums are worth €0.5, then this tells us the relative value of apples, oranges, and plums. Conversely, if we *can* assign consistent relative values to apples, oranges, and plums, then we can factor out an abstract unit at will—for example, by arbitrarily declaring apples to be worth €100 and then calculating everything else's price in apples.

Have we proven by pure logic that all apples have the same utility? Of course not; you can prefer some particular apples to other particular apples. But when you're done saying which things you qualitatively prefer to which other things, if you go around making tradeoffs in a way that can be *viewed as* not qualitatively leaving behind some things you said you wanted, we can *view you* as assigning coherent quantitative utilities to everything you want.

And that's one coherence theorem—among others—that can be seen as motivating the concept of *utility* in decision theory.

Utility isn't a solid thing, a separate thing. We could multiply all the utilities by two, and that would correspond to the same outward behaviors. It's meaningless to ask how much utility you scored at the end of your life, because we could subtract a million or add a million to that quantity while leaving everything else conceptually the same.

You could pick anything you valued—say, the joy of watching a cat chase a laser pointer for 10 seconds—and denominate everything relative to that, without needing any concept of an extra abstract 'utility'. So (just to be extremely clear about this point) we have not proven that there is a separate thing 'utility' that you should be pursuing instead of everything else you wanted in life.

The coherence theorem says nothing about which things to value more than others, or how much to value them relative to other things. It doesn't say whether you should value your happiness more than someone else's happiness, any more than the notion of a consistent preference ordering $>_P$ tells us whether onions $>_P$ pineapple.

(The notion that we should assign equal value to all human lives, or equal value to all sentient lives, or equal value to all Quality-Adjusted Life Years, is *utilitarianism.* Which is, sorry about the confusion, a whole 'nother separate different philosophy.)

The conceptual gizmo that maps thingies to utilities—the whatchamacallit that takes in a fruit and spits out a utility—is called a 'utility function'. Again, this isn't a separate thing that's written on a stone tablet. If we multiply a utility function by 9.2, that's conceptually the same utility function because it's consistent with the same set of behaviors.

But in general: If we can sensibly view any agent as doing as well as qualitatively possible at *anything*, we must be able to view the agent's behavior as consistent with there being some coherent relative quantities of wantedness for all the thingies it's trying to optimize.

Probabilities and expected utility

We've so far made no mention of *probability.* But the way that probabilities and utilities interact, is where we start to see the full structure of *expected utility* spotlighted by all the coherence theorems.

The basic notion in expected utility is that some choices present us with uncertain outcomes.

For example, I come to you and say: "Give me 1 apple, and I'll flip a coin; if the coin lands heads, I'll give you 1 orange; if the coin comes up tails, I'll give you 3 plums." Suppose you relatively value fruits as described earlier: 2 apples / orange and 0.5 apples / plum. Then *either* possible outcome gives you something that's worth more to you than 1 apple. Turning down a so-called 'gamble' like that... why, it'd be a dominated strategy.

In general, the notion of 'expected utility' says that we assign certain quantities called *probabilities* to each possible outcome. In the example above, we might assign a 'probability' of 0.5 to the coin landing heads (1 orange), and a 'probability' of 0.5 to the coin landing tails (3 plums). Then the total value of the 'gamble' we get by trading away 1 apple is:

$$P(heads) \cdot U(\text{1 orange}) + P(tails) \cdot U(\text{3 plums})$$

$$= 0.50 \cdot \overset{\text{€}}{\quad} 2 + 0.50 \cdot \overset{\text{€}}{\quad} 1.5 = \overset{\text{€}}{\quad} 1.75$$

Conversely, if we just keep our 1 apple instead of making the trade, this has an expected utility of $1 \cdot U(\text{1 apple}) = \text{€}1$. So indeed we ought to trade (as the previous reasoning suggested).

"But wait!" you cry. "Where did these probabilities come from? Why is the 'probability' of a fair coin landing heads 0.5 and not, say, -0.2 or 3? Who says we ought to multiply utilities by probabilities in the first place?"

If you're used to approaching this problem from a Bayesian standpoint, then you may now be thinking of notions like prior probability and Occam's Razor and universal priors...

But from the standpoint of coherence theorems, that's putting the cart before the horse.

From the standpoint of coherence theorems, we don't *start with* a notion of 'probability'.

Instead we ought to prove something along the lines of: if you're not using qualitatively dominated strategies, then you must *behave as if* you are multiplying utilities by certain quantitative thingies.

We might then furthermore show that, for non-dominated strategies, these utility-multiplying thingies must be between 0 and 1 rather than say -0.3 or 27.

Having determined what coherence properties these utility-multiplying thingies need to have, we decide to call them 'probabilities'. And *then*—once we know in the first place that we need 'probabilities' in order to not be using dominated strategies—we can start to worry about exactly what the numbers ought to be.

Probabilities summing to 1

Here's a taste of the kind of reasoning we might do:

Suppose that—having already accepted some previous proof that non-dominated strategies dealing with uncertain outcomes, must multiply utilities by quantitative thingies—you then say that you are going to assign a probability of 0.6 to the coin coming up heads, and a probability of 0.7 to the coin coming up tails.

If you're already used to the standard notion of probability, you might object, "But those probabilities sum to 1.3 when they ought to sum to 1!"[5] But now we are in coherence-land; we don't ask "Did we violate the standard axioms that all the textbooks use?" but "What rules must non-dominated strategies obey?" *De gustibus non est disputandum;* can we *disputandum* somebody saying that a coin has a 60% probability of coming up heads and a 70% probability of coming up tails? (Where these are the only 2 possible outcomes of an uncertain coinflip.)

Well—assuming you've already accepted that we need utility-multiplying thingies—I might then offer you a gamble. How about you give me one apple, and if the coin lands heads, I'll give you 0.8 apples; while if the coin lands tails, I'll give you 0.8 apples.

According to you, the expected utility of this gamble is:

$$P(\text{heads}) \cdot U(\text{0.8 apples}) + P(\text{tails}) \cdot U(\text{0.8 apples})$$

$$= 0.6 \cdot \text{€}\,0.8 + 0.7 \cdot \text{€}\,0.8 = \text{€}\,1.04.$$

You've just decided to trade your apple for 0.8 apples, which sure sounds like one of 'em dominated strategies.

And that's why *the thingies you multiply probabilities by*—the thingies that you use to weight uncertain outcomes in your imagination, when you're trying to decide how much you want one branch of an uncertain choice—must sum to 1, whether you call them 'probabilities' or not.

Well... actually we just argued[6] that probabilities for mutually exclusive outcomes should sum to *no more than 1.* What would be an example showing that, for non-dominated strategies, the probabilities for exhaustive outcomes should sum to no less than 1?

Why exhaustive outcomes should sum to at least 1:

Suppose that, in exchange for 1 apple, I credibly offer:

* To pay you 1.1 apples if a coin comes up heads.
* To pay you 1.1 apples if a coin comes up tails.
* To pay you 1.1 apples if anything else happens.

If the probabilities you assign to these three outcomes sum to say 0.9, you will refuse to trade 1 apple for 1.1 apples.

(This is strictly dominated by the strategy of agreeing to trade 1 apple for 1.1 apples.)

Dutch book arguments

Another way we could have presented essentially the same argument as above, is as follows:

Suppose you are a market-maker in a prediction market for some event X. When you say that your price for event X is x, you mean that you will sell for $\$x$ a ticket which pays $\$1$ if X happens (and pays out nothing otherwise). In fact, you will sell any number of such tickets!

Since you are a market-maker (that is, you are trying to encourage trading in X for whatever reason), you are also willing to *buy* any number of tickets at the price $\$x$. That is, I can say to you (the market-maker) "I'd like to sign a contract where you give me $N \cdot \$x$ now, and in return I must pay you $\$N$ iff X happens;" and you'll agree. (We can view this as you selling me a negative number of the original kind of ticket.)

Let X and Y denote two events such that *exactly one* of them must happen; say, X is a coin landing heads and Y is the coin not landing heads.

Now suppose that you, as a market-maker, are motivated to avoid combinations of bets that lead into *certain* losses for you—not just losses that are merely probable, but combinations of bets such that *every* possibility leads to a loss.

Then if exactly one of X and Y must happen, your prices x and y must sum to exactly $\$1$. Because:

- If $x + y < \$1$, I buy both an X-ticket and a Y-ticket and get a guaranteed payout of $\$1$ minus costs of $x + y$. Since this is a guaranteed profit for me, it is a guaranteed loss for you.
- If $x + y > \$1$, I sell you both tickets and will at the end pay you $\$1$ after you have already paid me $x + y$. Again, this is a guaranteed profit for me of $x + y - \$1 > \$0$.

This is more or less exactly the same argument as in the previous section, with trading apples. Except that: (a) the scenario is more crisp, so it is easier to generalize and scale up much more complicated similar arguments; and (b) it introduces a whole lot of assumptions that people new to expected utility would probably find rather questionable.

"What?" one might cry. "What sort of crazy bookie would buy and sell bets at exactly the same price? Why ought *anyone* to buy and sell bets at exactly the same price? Who says that I must value a gain of $1 exactly the opposite of a loss of $1? Why should the price that I put on a bet represent my degree of uncertainty about the environment? What does all of this argument about gambling have to do with real life?"

So again, the key idea is not that we are assuming anything about people valuing every real-world dollar the same; nor is it in real life a good idea to offer to buy or sell bets at the same prices.[7] Rather, Dutch book arguments can stand in as shorthand for some longer story in which we only assume that you prefer more apples to less apples.

The Dutch book argument above has to be seen as one more added piece in the company of all the *other* coherence theorems—for example, the coherence theorems suggesting that you ought to be quantitatively weighing events in your mind in the first place.

Conditional probability

With more complicated Dutch book arguments, we can derive more complicated ideas such as 'conditional probability'.

Let's say that we're pricing three kinds of gambles over two events Q and R:

- A ticket that costs $x, and pays $1 if Q happens.
- A ticket that doesn't cost anything or pay anything if Q doesn't happen (the ticket price is refunded); and if Q does happen, this ticket costs $y, then pays $1 if R happens.
- A ticket that costs $z, and pays $1 if Q and R both happen.

Intuitively, the idea of conditional probability is that the probability of Q and R both happening, should be equal to the probability of Q happening, times the probability that R happens assuming that Q happens:

$$P(Q \wedge R) = P(Q) \cdot P(R \mid Q)$$

To exhibit a Dutch book argument for this rule, we want to start from the assumption of a qualitatively non-dominated strategy, and derive the quantitative rule $z = x \cdot y$.

So let's give an example that violates this equation and see if there's a way to make a guaranteed profit. Let's say somebody:

- Prices at x = $0.60 the first ticket, aka P(Q).
- Prices at y = $0.70 the second ticket, aka P(R | Q).
- Prices at z = $0.20 the third ticket, aka P(Q ∧ R), which ought to be $0.42 assuming the first two prices.

The first two tickets are priced relatively high, compared to the third ticket which is priced relatively low, suggesting that we ought to sell the first two tickets and buy the third.

Okay, let's ask what happens if we sell 10 of the first ticket, sell 10 of the second ticket, and buy 10 of the third ticket.

- If Q doesn't happen, we get $6, and pay $2. Net +$4.
- If Q happens and R doesn't happen, we get $6, pay $10, get $7, and pay $2. Net +$1.
- If Q happens and R happens, we get $6, pay $10, get $7, pay $10, pay $2, and get $10. Net: +$1.

That is: we can get a guaranteed positive profit over all three possible outcomes.

More generally, let A, B, C be the (potentially negative) amount of each ticket X, Y, Z that is being bought (buying a negative amount is selling). Then the prices x, y, z can be combined into a 'Dutch book' whenever the following three inequalities can be simultaneously true, with at least one inequality strict:

$$\begin{array}{rrrl} -Ax & +0 & -Cz & \geqq 0 \\ A(1-x) & -By & -Cz & \geqq 0 \\ A(1-x) & +B(1-y) & +C(1-z) & \geqq 0 \end{array}$$

For $x, y, z \in (0..1)$ this is impossible exactly iff $z = x \cdot y$. The proof via a bunch of algebra is left as an exercise to the reader.[8]

The Allais Paradox

By now, you'd probably like to see a glimpse of the sort of argument that shows in the first place that we need expected utility—that a non-dominated strategy for uncertain choice must behave as if multiplying utilities by some kinda utility-multiplying thingies ('probabilities').

As far as I understand it, the real argument you're looking for is Abraham Wald's complete class theorem, which I must confess I don't know how to reduce to a simple demonstration.

But we can catch a glimpse of the general idea from a famous psychology experiment that became known as the Allais Paradox (in slightly adapted form).

Suppose you ask some experimental subjects which of these gambles they would rather play:

- 1A: A certainty of $1,000,000.
- 1B: 90% chance of winning $5,000,000, 10% chance of winning nothing.

Most subjects say they'd prefer 1A to 1B.

Now ask a separate group of subjects which of these gambles they'd prefer:

- 2A: 50% chance of winning $1,000,000; 50% chance of winning $0.
- 2B: 45% chance of winning $5,000,000; 55% chance of winning $0.

In this case, most subjects say they'd prefer gamble 2B.

Note that the $ sign here denotes real dollars, not utilities! A gain of five million dollars isn't, and shouldn't be, worth exactly five times as much to you as a gain of one million dollars. We can use the € symbol to denote the expected utilities that are abstracted from how much you relatively value different outcomes; $ is just money.

So we certainly aren't claiming that the first preference is paradoxical because 1B has an expected dollar value of $4.5 million and 1A has an expected dollar value of $1 million. That would be silly. We care about expected utilities, not expected dollar values, and those two concepts aren't the same at all!

Nonetheless, the combined preferences 1A > 1B and 2A < 2B are not compatible with any coherent utility function. We cannot simultaneously have:

$$\begin{array}{rcl} U(\text{gain \$1M}) & > & 0.9 \cdot U(\text{gain \$5M}) + 0.1 \cdot U(\text{gain \$0}) \\ 0.5 \cdot U(\text{gain \$0}) + 0.5 \cdot U(\text{gain \$1M}) & < & 0.45 \cdot U(\text{gain \$5M}) + 0.55 \cdot U(\text{gain \$0}) \end{array}$$

This was one of the earliest experiments seeming to demonstrate that actual human beings were not expected utility maximizers—a very tame idea nowadays, to be sure, but the *first definite* demonstration of that was a big deal at the time. Hence the term, "Allais Paradox".

Now, by the general idea behind coherence theorems, since we can't *view this behavior* as corresponding to expected utilities, we ought to be able to show that it corresponds to a dominated strategy somehow—derive some way in which this behavior corresponds to shooting off your own foot.

In this case, the relevant idea seems non-obvious enough that it doesn't seem reasonable to demand that you think of it on your own; but if you like, you can pause and try to think of it anyway. Otherwise, just continue reading.

Again, the gambles are as follows:

- 1A: A certainty of $1,000,000.
- 1B: 90% chance of winning $5,000,000, 10% chance of winning nothing.
- 2A: 50% chance of winning $1,000,000; 50% chance of winning $0.
- 2B: 45% chance of winning $5,000,000; 55% chance of winning $0.

Now observe that Scenario 2 corresponds to a 50% chance of playing Scenario 1, and otherwise getting $0.

This, in fact, is why the combination 1A > 1B; 2A < 2B is incompatible with expected utility. In terms of one set of axioms frequently used to describe expected utility, it violates the Independence Axiom: if a gamble L is preferred to M (that is, $L > M$), then we ought to be able to take a constant probability $p > 0$ and another gamble N and have

$p \cdot L + (1 - p) \cdot N > p \cdot M + (1 - p) \cdot N$.

To put it another way, if I flip a coin to decide whether or not to play some entirely different game N, but otherwise let you choose L or M, you ought to make the same choice as if I just ask you whether you prefer L or M. Your preference between L and M should be 'independent' of the possibility that, instead of doing anything whatsoever with L or M, we will do something else instead.

And since this is an axiom of expected utility, any violation of that axiom ought to correspond to a dominated strategy somehow.

In the case of the Allais Paradox, we do the following:

First, I show you a switch that can be set to A or B, currently set to A.

In one minute, I tell you, I will flip a coin. If the coin comes up heads, you will get nothing. If the coin comes up tails, you will play the gamble from Scenario 1.

From your current perspective, that is, we are playing Scenario 2: since the switch is set to A, you have a 50% chance of getting nothing and a 50% chance of getting $1 million.

I ask you if you'd like to pay a penny to throw the switch from A to B. Since you prefer gamble 2B to 2A, and some quite large amounts of money are at stake, you agree to pay the penny. From your perspective, you now have a 55% chance of ending up with nothing and a 45% chance of getting $5M.

I then flip the coin, and luckily for you, it comes up tails.

From your perspective, you are now in Scenario 1B. Having observed the coin and updated on its state, you now think you have a 90% chance of getting $5 million and a 10% chance of getting nothing. By hypothesis, you would prefer a certainty of $1 million.

So I offer you a chance to pay another penny to flip the switch back from B to A. And with so much money at stake, you agree.

I have taken your two cents on the subject.

That is: You paid a penny to flip a switch and then paid another penny to switch it back, and this is dominated by the strategy of just leaving the switch set to A.

And that's at least a glimpse of why, if you're not using dominated strategies, the thing you do with relative utilities is multiply them by probabilities in a consistent way, and prefer the choice that leads to a greater expectation of the variable representing utility.

From the Allais Paradox to real life

The real-life lesson about what to do when faced with Allais's dilemma might be something like this:

There's *some* amount that $1 million would improve your life compared to $0.

There's some amount that an additional $4 million would further improve your life after the first $1 million.

You ought to visualize these two improvements as best you can, and decide whether another $4 million can produce at least *one-ninth* as much improvement, as much true value to you, as the first $1 million.

If it can, you should consistently prefer 1B > 1A; 2B > 2A. And if not, you should consistently prefer 1A > 1B; 2A > 2B.

The standard 'paradoxical' preferences in Allais's experiment are standardly attributed to a certainty effect: people value the *certainty* of having $1 million, while the difference between a 50% probability and a 55% probability looms less large. (And this ties in to a number of other results about certainty, need for closure, prospect theory, and so on.)

It may sound intuitive, in an Allais-like scenario, to say that you ought to derive some value from being *certain* about the outcome. In fact this is just the reasoning the experiment shows people to be using, so of course it might sound intuitive. But that does, inescapably, correspond to a kind of thinking that produces dominated strategies.

One possible excuse might be that certainty is valuable if you need to make plans about the future; knowing the exact future lets you make better plans. This is admittedly true and a phenomenon within expected utility, though it applies in a smooth way as confidence increases rather than jumping suddenly around 100%. But in the particular dilemma as described here, you only have 1 minute before the game is played, and no time to make other major life choices dependent on the outcome.

Another possible excuse for certainty bias might be to say: "Well, I value the emotional feeling of certainty."

In real life, we do have emotions that are directly about probabilities, and those little flashes of happiness or sadness are worth something if you care about people being happy or sad. If you say that you value the emotional feeling of being *certain* of getting $1 million, the freedom from the fear of getting $0, for the minute that the dilemma lasts and you are experiencing the emotion—well, that may just be a fact about what you value, even if it exists outside the expected utility formalism.

And this genuinely does not fit into the expected utility formalism. In an expected utility agent, probabilities are just thingies-you-multiply-utilities-by. If those thingies start generating their own utilities once represented inside the mind of the person who is an object of ethical value, you really are going to get results that are incompatible with the formal decision theory.

However, *not* being viewable as an expected utility agent does always correspond to employing dominated strategies. You are giving up *something* in exchange, if you pursue that feeling of certainty. You are potentially losing all the real value you could have gained from another $4 million, if that realized future actually would have gained you more than one-ninth the value of the first $1 million. Is a fleeting emotional sense of certainty over 1 minute, worth *automatically* discarding the potential $5-million outcome? Even if the correct answer given your values is that you properly ought to take the $1 million, treasuring 1 minute of emotional gratification doesn't seem like the wise reason to do that. The wise reason would be if the first $1 million really was worth that much more than the next $4 million.

The danger of saying, "Oh, well, I attach a lot of utility to that comfortable feeling of certainty, so my choices are coherent after all" is not that it's mathematically improper to value the emotions we feel while we're deciding. Rather, by saying that the *most valuable* stakes are the emotions you feel during the minute you make the decision, what you're saying is, "I get a huge amount of value by making decisions however humans instinctively make their decisions, and that's much more important than the thing I'm making a decision *about.*" This could well be true for something like buying a stuffed animal. If millions of dollars or human lives are at stake, maybe not so much.

Conclusion

The demonstrations we've walked through here aren't the professional-grade coherence theorems as they appear in real math. Those have names like "Cox's Theorem" or "the complete class theorem"; their proofs are difficult; and they say things like "If seeing piece of information A followed by piece of information B leads you into the same epistemic state as seeing piece of information B followed by piece of information A, plus some other assumptions, I can show an isomorphism between those epistemic states and classical probabilities" or "Any decision rule for taking different actions depending on your observations either corresponds to Bayesian updating given some prior, or else is strictly dominated by some Bayesian strategy".

But hopefully you've seen enough concrete demonstrations to get a general idea of what's going on with the actual coherence theorems. We have multiple spotlights all shining on the same core mathematical structure, saying dozens of different variants on, "If you aren't running around in circles or stepping on your own feet or wantonly giving up things you say you want, we can see your behavior as corresponding to this shape. Conversely, if we can't see your behavior as corresponding to this shape, you must be visibly shooting yourself in the foot." Expected utility is the only structure that has this great big family of discovered theorems all saying that. It has a scattering of academic competitors, because academia is academia, but the competitors don't have anything like that mass of spotlights all pointing in the same direction.

So if we need to pick an interim answer for "What kind of quantitative framework should I try to put around my own decision-making, when I'm trying to check if my thoughts make sense?" or "By default and barring special cases, what properties might a sufficiently advanced machine intelligence *look to us* like it possessed, at least approximately, if we couldn't see it *visibly* running around in circles?", then there's pretty much one obvious candidate: Probabilities, utility functions, and expected utility.

Further reading

- To learn more about agents and AI: Consequentialist cognition; the orthogonality of agents' utility functions and capabilities; epistemic and instrumental efficiency; instrumental strategies sufficiently capable agents tend to converge on; properties of sufficiently advanced agents.
- To learn more about decision theory: The controversial counterfactual at the heart of the expected utility formula.

[1] It could be that somebody's pizza preference is real, but so weak that they wouldn't pay one penny to get the pizza they prefer. In this case, imagine we're talking about some stronger preference instead. Like your willingness to pay at least one penny not to have your house burned down, or something.

[2] This does assume that the agent prefers to have more money rather than less money. "Ah, but why is it bad if one person has a penny instead of another?" you ask. If we insist on pinning down every point of this sort, then you can also imagine the $0.01 as standing in for the *time* I burned in order to move the pizza slices around in circles. That time was burned, and nobody else has it now. If I'm an effective agent that goes around pursuing my preferences, I should in general be able to sometimes convert time into other things that I want. In other words, my circular preference can lead me to incur an opportunity cost denominated in the sacrifice of other things I want, and not in a way that benefits anyone else.

[3] There are more than six possibilities if you think it's possible to be absolutely indifferent between two kinds of pizza.

[4] We can omit the 'better doctors' item from consideration: The supply of doctors is mostly constrained by regulatory burdens and medical schools rather than the number of people who want to become doctors; so bidding up salaries for doctors doesn't much increase the total number of doctors; so bidding on a talented doctor at one hospital just means some other hospital doesn't get that talented doctor. It's also illegal to pay for livers, but let's ignore that particular issue with the problem setup or pretend that it all takes place in a more sensible country than the United States or Europe.

[5] Or maybe a tiny bit less than 1, in case the coin lands on its edge or something.

[6] Nothing we're walking through here is really a coherence theorem *per se*, more like intuitive arguments that a coherence theorem ought to exist. Theorems require proofs, and nothing here is what real mathematicians would consider to be a 'proof'.

[7] In real life this leads to a problem of 'adversarial selection', where somebody who knows more about the environment than you can decide whether to buy or sell from you. To put it another way, from a Bayesian standpoint, if an *intelligent* counterparty is deciding whether to buy or sell from you a bet on X, the fact that they choose to buy (or sell) should cause you to update in favor (or against) X actually happening. After all, they wouldn't be taking the bet unless they thought they knew something you didn't!

[8] The quick but advanced argument would be to say that the left-hand-side must look like a singular matrix, whose determinant must therefore be zero.

A Semitechnical Introductory Dialogue on Solomonoff Induction

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

(*Originally posted in December 2015: A dialogue between Ashley, a computer scientist who's never heard of Solomonoff's theory of inductive inference, and Blaine, who thinks it is the best thing since sliced bread.*)

---

i. Unbounded analysis

ASHLEY: Good evening, Msr. Blaine.

BLAINE: Good evening, Msr. Ashley.

ASHLEY: I've heard there's this thing called "Solomonoff's theory of inductive inference".

BLAINE: The rumors have spread, then.

ASHLEY: Yeah, so, what the heck is that about?

BLAINE: Invented in the 1960s by the mathematician Ray Solomonoff, the key idea in Solomonoff induction is to do sequence prediction by using Bayesian updating on a prior composed of a mixture of all computable probability distributions—

ASHLEY: Wait. Back up a lot. Before you try to explain what Solomonoff induction *is*, I'd like you to try to tell me what it *does*, or why people study it in the first place. I find that helps me organize my listening. Right now I don't even know why I should be interested in this.

BLAINE: Um, okay. Let me think for a second...

ASHLEY: Also, while I can imagine things that "sequence prediction" might mean, I haven't yet encountered it in a technical context, so you'd better go a bit further back and start more at the beginning. I do know what "computable" means and what a "probability distribution" is, and I remember the formula for Bayes's Rule although it's been a while.

BLAINE: Okay. So... one way of framing the usual reason why people study this general field in the first place, is that sometimes, by studying certain idealized mathematical questions, we can gain valuable intuitions about epistemology. That's, uh, the field that studies how to reason about factual questions, how to build a map of reality that reflects the territory—

ASHLEY: I have some idea what 'epistemology' is, yes. But I think you might need to start even further back, maybe with some sort of concrete example or something.

BLAINE: Okay. Um. So one anecdote that I sometimes use to frame the value of computer science to the study of epistemology is Edgar Allen Poe's argument in 1833 that chess was uncomputable.

ASHLEY: That doesn't sound like a thing that actually happened.

BLAINE: I know, but it totally *did* happen and not in a metaphorical sense either! Edgar Allen Poe wrote an essay explaining why no automaton would ever be able to play chess, and he specifically mentioned "Mr. Babbage's computing engine" as an example.

You see, in the nineteenth century, there was for a time this sensation known as the Mechanical Turk—supposedly a machine, an automaton, that could play chess. At the grandmaster level, no less.

Now today, when we're accustomed to the idea that it takes a reasonably powerful computer to do that, we can know *immediately* that the Mechanical Turk must have been a fraud and that there must have been a concealed operator inside—a person with dwarfism, as it turned out. *Today* we know that this sort of thing is *hard* to build into a machine. But in the 19th century, even that much wasn't known.

So when Edgar Allen Poe, who besides being an author was also an accomplished magician, set out to write an essay about the Mechanical Turk, he spent the *second* half of the essay dissecting what was known about the Turk's appearance to (correctly) figure out where the human operator was hiding. But Poe spent the first half of the essay arguing that no automaton—nothing like Mr. Babbage's computing engine—could possibly play chess, which was how he knew *a priori* that the Turk had a concealed human operator.

ASHLEY: And what was Poe's argument?

BLAINE: Poe observed that in an algebraical problem, each step followed from the previous step of necessity, which was why the steps in solving an algebraical problem could be represented by the deterministic motions of gears in something like Mr. Babbage's computing engine. But in a chess problem, Poe said, there are many possible chess moves, and no move follows with necessity from the position of the board; and even if you did select one move, the opponent's move would not follow with necessity, so you couldn't represent it with the determined motion of automatic gears. Therefore, Poe said, whatever was operating the Mechanical Turk must have the nature of Cartesian mind, rather than the nature of deterministic matter, and this was knowable *a priori*. And then he started figuring out where the required operator was hiding.

ASHLEY: That's some amazingly impressive reasoning for being completely wrong.

BLAINE: I know! Isn't it great?

ASHLEY: I mean, that sounds like Poe correctly identified the *hard* part of playing computer chess, the branching factor of moves and countermoves, which is the reason why no *simple* machine could do it. And he just didn't realize that a deterministic machine could deterministically check many possible moves in order to figure out the game tree. So close, and yet so far.

BLAINE: More than a century later, in 1950, Claude Shannon published the first paper ever written on computer chess. And in passing, Shannon gave the formula for playing perfect chess if you had unlimited computing power, the algorithm you'd use to extrapolate the entire game tree. We could say that Shannon gave a short program that would solve chess if you ran it on a hypercomputer, where a hypercomputer is an ideal computer that can run any finite computation immediately. And then Shannon passed on to talking about the problem of locally guessing how good a board position was, so that you could play chess using only a *small* search.

I say all this to make a point about the value of knowing how to solve problems using hypercomputers, even though hypercomputers don't exist. Yes, there's often a *huge* gap between the unbounded solution and the practical solution. It wasn't until 1997, forty-seven years after Shannon's paper giving the unbounded solution, that Deep Blue actually won the world chess championship—

ASHLEY: And that wasn't just a question of faster computing hardware running Shannon's ideal search algorithm. There were a lot of new insights along the way, most notably the alpha-beta pruning algorithm and a lot of improvements in positional evaluation.

BLAINE: Right!

But I think some people overreact to that forty-seven year gap, and act like it's *worthless* to have an unbounded understanding of a computer program, just because you might still be forty-seven years away from a practical solution. But if you don't even have a solution that would run on a hypercomputer, you're Poe in 1833, not Shannon in 1950.

The reason I tell the anecdote about Poe is to illustrate that Poe was *confused* about computer chess in a way that Shannon was not. When we don't know how to solve a problem even given infinite computing power, the very work we are trying to do is in some sense murky to us. When we can state code that would solve the problem given a hypercomputer, we have become *less* confused. Once we have the unbounded solution we understand, in some basic sense, *the kind of work we are trying to perform,* and then we can try to figure out how to do it efficiently.

ASHLEY: Which may well require new insights into the structure of the problem, or even a conceptual revolution in how we imagine the work we're trying to do.

BLAINE: Yes, but the point is that you can't even get started on that if you're arguing about how playing chess has the nature of Cartesian mind rather than matter. At that point you're not 50 years away from winning the chess championship, you're 150 years away, because it took an extra 100 years to move humanity's understanding to the point where Claude Shannon could trivially see how to play perfect chess using a large-enough computer. I'm not trying to exalt the unbounded solution by denigrating the work required to get a bounded solution. I'm not saying that when we have an unbounded solution we're practically there and the rest is a matter of mere lowly efficiency. I'm trying to compare having the unbounded solution to the horrific confusion of *not understanding what we're trying to do.*

ASHLEY: Okay. I think I understand why, on your view, it's important to know how to solve problems using infinitely fast computers, or hypercomputers as you call them. When we can say how to answer a question using infinite computing power, that means we crisply understand the question itself, in some sense; while if we can't figure out how to solve a problem using unbounded computing power, that means we're *confused* about the problem, in some sense. I mean, anyone who's ever tried to teach the more doomed sort of undergraduate to write code knows what it means to be confused about what it takes to compute something.

BLAINE: Right.

ASHLEY: So what does this have to do with "Solomonoff induction"?

BLAINE: Ah! Well, suppose I asked you how to do epistemology using infinite computing power?

ASHLEY: My good fellow, I would at once reply, "Beep. Whirr. Problem 'do epistemology' not crisply specified." At this stage of affairs, I do not think this reply indicates any fundamental confusion on my part; rather I think it is you who must be clearer.

BLAINE: Given unbounded computing power, how would you reason in order to construct an accurate map of reality?

ASHLEY: That still strikes me as rather underspecified.

BLAINE: Perhaps. But even there I would suggest that it's a mark of intellectual progress to be able to take vague and underspecified ideas like 'do good epistemology' and turn them *into* crisply specified problems. Imagine that I went up to my friend Cecil, and said, "How would you do good epistemology given unlimited computing power and a short Python program?" and Cecil at once came back with an answer—a good and reasonable answer, once it was explained. Cecil would probably know something quite interesting that you do not presently know.

ASHLEY: I confess to being rather skeptical of this hypothetical. But if that actually happened—if I agreed, to my own satisfaction, that someone had stated a short Python program that would 'do good epistemology' if run on an unboundedly fast computer—then I agree that I'd probably have learned something *quite interesting* about epistemology.

BLAINE: What Cecil knows about, in this hypothetical, is Solomonoff induction. In the same way that Claude Shannon answered "Given infinite computing power, how would you play perfect chess?", Ray Solomonoff answered "Given infinite computing power, how would you perfectly find the best hypothesis that fits the facts?"

ASHLEY: Suddenly, I find myself strongly suspicious of whatever you are about to say to me.

BLAINE: That's understandable.

ASHLEY: In particular, I'll ask at once whether "Solomonoff induction" assumes that our hypotheses are being given to us on a silver platter along with the exact data we're supposed to explain, or whether the algorithm is organizing its own data from a big messy situation and inventing good hypotheses from scratch.

BLAINE: Great question! It's the second one.

ASHLEY: Really? Okay, now I have to ask whether Solomonoff induction is a recognized concept in good standing in the field of academic computer science, because that does not sound like something modern-day computer science knows how to do.

BLAINE: I wouldn't say it's a widely known concept, but it's one that's in good academic standing. The method isn't used in modern machine learning because it requires an infinitely fast computer and isn't easily approximated the way that chess is.

ASHLEY: This really sounds very suspicious. Last time I checked, we hadn't *begun* to formalize the creation of good new hypotheses from scratch. I've heard about claims to have 'automated' the work that, say, Newton did in inventing classical mechanics, and I've found them all to be incredibly dubious. Which is to say, they were rigged demos and lies.

BLAINE: I know, but—

ASHLEY: And then I'm even more suspicious of a claim that someone's algorithm would solve this problem if only they had infinite computing power. Having some researcher claim that their Good-Old-Fashioned AI semantic network *would* be intelligent if run on a computer so large that, conveniently, nobody can ever test their theory, is not going to persuade me.

BLAINE: Do I really strike you as that much of a charlatan? What have I ever done to you, that you would expect me to try pulling a scam like that?

ASHLEY: That's fair. I shouldn't accuse you of planning that scam when I haven't seen you say it. But I'm pretty sure the problem of "coming up with good new hypotheses in a world full of messy data" is AI-complete. And even Mentif-

BLAINE: Do not say the name, or he will appear!

ASHLEY: Sorry. Even the legendary first and greatest of all AI crackpots, He-Who-Googles-His-Name, could assert that his algorithms would be all-powerful on a computer large enough to make his claim unfalsifiable. So what?

BLAINE: That's a very sensible reply and this, again, is exactly the kind of mental state that reflects a problem that is *confusing* rather than just hard to implement. It's the sort of confusion Poe might feel in 1833, or close to it. In other words, it's just the sort of conceptual issue we *would* have solved at the point where we could state a short program that could run on a hypercomputer. Which Ray Solomonoff did in 1964.

ASHLEY: Okay, let's hear about this supposed general solution to epistemology.

ii. Sequences

BLAINE: First, try to solve the following puzzle. 1, 3, 4, 7, 11, 18, 29...?

ASHLEY: Let me look at those for a moment... 47.

BLAINE: Congratulations on engaging in, as we snooty types would call it, 'sequence prediction'.

ASHLEY: I'm following you so far.

BLAINE: The smarter you are, the more easily you can find the hidden patterns in sequences and predict them successfully. You had to notice the resemblance to the Fibonacci rule to guess the next number. Someone who didn't already know about Fibonacci, or who was worse at mathematical thinking, would have taken longer to understand the sequence or maybe never learned to predict it at all.

ASHLEY: Still with you.

BLAINE: It's not a sequence of *numbers* per se... but can you see how the question, "The sun has risen on the last million days. What is the probability that it rises tomorrow?" could be viewed as a kind of sequence prediction problem?

ASHLEY: Only if some programmer neatly parses up the world into a series of "Did the Sun rise on day X starting in 4.5 billion BCE, 0 means no and 1 means yes? 1, 1, 1, 1, 1..." and so on. Which is exactly the sort of shenanigan that I see as cheating. In the real world, you go outside and see a brilliant ball of gold touching the horizon, not a giant "1".

BLAINE: Suppose I have a robot running around with a webcam showing it a 1920 × 1080 pixel field that refreshes 60 times a second with 32-bit colors. I could view that as a giant sequence and ask the robot to predict what it will see happen when it rolls out to watch a sunrise the next day.

ASHLEY: I can't help but notice that the 'sequence' of webcam frames is absolutely enormous, like, the sequence is made up of 66-megabit 'numbers' appearing 3600 times per minute... oh, right, computers much bigger than the universe. And now you're smiling evilly, so I guess that's the point. I also notice that the sequence is no longer deterministically predictable, that it is no longer a purely mathematical object, and that the sequence of webcam frames observed will depend on the robot's choices. This makes me feel a bit shaky about the analogy to predicting the mathematical sequence 1, 1, 2, 3, 5.

BLAINE: I'll try to address those points in order. First, Solomonoff induction is about assigning *probabilities* to the next item in the sequence. I mean, if I showed you a box that said 1, 1, 2, 3, 5, 8 you would not be absolutely certain that the next item would be 13. There could be some more complicated rule that just looked Fibonacci-ish but then diverged. You might guess with 90% probability but not 100% probability, or something like that.

ASHLEY: This has stopped feeling to me like math.

BLAINE: There is a *large* branch of math, to say nothing of computer science, that deals in probabilities and statistical prediction. We are going to be describing absolutely lawful and deterministic ways of assigning probabilities after seeing 1, 3, 4, 7, 11, 18.

ASHLEY: Okay, but if you're later going to tell me that this lawful probabilistic prediction rule underlies a generally intelligent reasoner, I'm already skeptical.

No matter how large a computer it's run on, I find it hard to imagine that some simple set of rules for assigning probabilities is going to encompass truly and generally intelligent answers about sequence prediction, like Terence Tao would give after looking at the sequence for a while. We just have no idea how Terence Tao works, so we can't duplicate his abilities in a formal rule, no matter how much computing power that rule gets... you're smiling evilly again. I'll be *quite* interested if that evil smile turns out to be justified.

BLAINE: Indeed.

ASHLEY: I also find it hard to imagine that this deterministic mathematical rule for assigning probabilities would notice if a box was outputting an encoded version of "To be or not to be" from Shakespeare by mapping A to Z onto 1 to 26, which I would notice eventually though not immediately upon seeing 20, 15, 2, 5, 15, 18... And you're *still* smiling evilly.

BLAINE: Indeed. That is *exactly* what Solomonoff induction does. Furthermore, we have theorems establishing that Solomonoff induction can do it way better than you or Terence Tao.

ASHLEY: A *theorem* proves this. As in a necessary mathematical truth. Even though we have no idea how Terence Tao works empirically... and there's evil smile number four. Okay. I am very skeptical, but willing to be convinced.

BLAINE: So if you actually did have a hypercomputer, you could cheat, right? And Solomonoff induction is the most ridiculously cheating cheat in the history of cheating.

ASHLEY: Go on.

BLAINE: We just run all possible computer programs to see which are the simplest computer programs that best predict the data seen so far, and use those programs to predict what comes next. This mixture contains, among other things, an exact copy of Terence Tao, thereby allowing us to prove theorems about their relative performance.

ASHLEY: Is this an actual reputable math thing? I mean really?

BLAINE: I'll deliver the formalization later, but you did ask me to first state the point of it all. The point of Solomonoff induction is that it gives us a gold-standard ideal for sequence prediction, and this gold-standard prediction only errs by a bounded amount, over infinite time, relative to the best computable sequence predictor. We can also see it as formalizing the intuitive idea that was expressed by William Ockham a few centuries earlier that simpler theories are more likely to be correct, and as telling us that 'simplicity' should be measured in algorithmic complexity, which is the size of a computer program required to output a hypothesis's predictions.

ASHLEY: I think I would have to read more on this subject to actually follow that. What I'm hearing is that Solomonoff induction is a reputable idea that is important because it gives us a kind of ideal for sequence prediction. This ideal also has something to do with Occam's Razor, and stakes a claim that the simplest theory is the one that can be represented by the shortest computer program. You identify this with "doing good epistemology".

BLAINE: Yes, those are legitimate takeaways. Another way of looking at it is that Solomonoff induction is an ideal but uncomputable answer to the question "What should our priors be?", which is left open by understanding Bayesian updating.

ASHLEY: Can you say how Solomonoff induction answers the question of, say, the prior probability that Canada is planning to invade the United States? I once saw a crackpot website that tried to invoke Bayesian probability about it, but only after setting the prior at 10% or something like that, I don't recall exactly. Does Solomonoff induction let me tell him that he's making a math error, instead of just calling him silly in an informal fashion?

BLAINE: If you're expecting to sit down with Leibniz and say, "Gentlemen, let us calculate" then you're setting your expectations too high. Solomonoff gives us an idea of how we *should* compute that quantity given unlimited computing power. It doesn't give us a firm recipe for how we can best approximate that ideal in real life using bounded computing power, or human brains. That's like expecting to play perfect chess after you read Shannon's 1950 paper. But knowing the ideal, we can extract some intuitive advice that might help our online crackpot if only he'd listen.

ASHLEY: But according to you, Solomonoff induction does say in principle what is the prior probability that Canada will invade the United States.

BLAINE: Yes, up to a choice of universal Turing machine.

ASHLEY: *(looking highly skeptical)* So I plug a universal Turing machine into the formalism, and in principle, I get out a uniquely determined probability that Canada invades the USA.

BLAINE: Exactly!

ASHLEY: Uh huh. Well, go on.

BLAINE: So, first, we have to transform this into a sequence prediction problem.

ASHLEY: Like a sequence of years in which Canada has and hasn't invaded the US, mostly zero except around 1812—

BLAINE: *No!* To get a good prediction about Canada we need much more data than that, and I don't mean a graph of Canadian GDP either. Imagine a sequence that contains all the sensory data you have ever received over your lifetime. Not just the hospital room that you saw when you opened your eyes right after your birth, but the darkness your brain received as input while you were still in your mother's womb. Every word you've ever heard. Every letter you've ever seen on a computer screen, not as ASCII letters but as the raw pattern of neural impulses that gets sent down from your retina.

ASHLEY: That seems like a lot of data and some of it is redundant, like there'll be lots of similar pixels for blue sky—

BLAINE: That data is what *you* got as an agent. If we want to translate the question of the prediction problem Ashley faces into theoretical terms, we should give the sequence predictor *all* the data that you had available, including all those repeating blue pixels of the sky. Who knows? Maybe there was a Canadian warplane somewhere in there, and you didn't notice.

ASHLEY: But it's impossible for my brain to remember all that data. If we neglect for the moment how the retina actually works and suppose that I'm seeing the same 1920×1080 @60Hz feed the robot would, that's far more data than my brain can realistically learn per second.

BLAINE: So then Solomonoff induction can do better than you can, using its unlimited computing power and memory. That's fine.

ASHLEY: But what if you can do better by forgetting more?

BLAINE: If you have limited computing power, that makes sense. With unlimited computing power, that really shouldn't happen and that indeed is one of the lessons of Solomonoff induction. An unbounded Bayesian never expects to do worse by updating on another item of evidence—for one thing, you can always just do the same policy you would have used if you hadn't seen that evidence. That kind of lesson *is* one of the lessons that might not be intuitively obvious, but which you can feel more deeply by walking through the math of probability theory. With unlimited computing power, nothing goes wrong as a result of trying to process 4 gigabits per second; every extra bit just produces a better expected future prediction.

ASHLEY: Okay, so we start with literally all the data I have available. That's 4 gigabits per second if we imagine 1920×1080 frames of 32-bit pixels repeating 60 times per second. Though I remember hearing 100 megabits per second would be a better estimate of what the retina sends out, and that it's pared down to 1 megabit per second very quickly by further processing.

BLAINE: Right. We start with all of that data, going back to when you were born. Or maybe when your brain formed in the womb, though it shouldn't make much difference.

ASHLEY: I note that there are some things I know that don't come from my sensory inputs at all. Chimpanzees learn to be afraid of skulls and snakes much faster than they learn to be afraid of other arbitrary shapes. I was probably better at learning to walk in Earth gravity than I would have been at navigating in zero G. Those are heuristics I'm born with, based on how my brain was wired, which ultimately stems from my DNA specifying the way that proteins should fold to form neurons—not from any photons that entered my eyes later.

BLAINE: So, for purposes of following along with the argument, let's say that your DNA is analogous to the code of a computer program that makes predictions. What you're observing here is that humans have 750 megabytes of DNA, and even if most of that is junk and not all of what's left is specifying brain behavior, it still leaves a pretty large computer program that could have a lot of prior information programmed into it.

Let's say that your brain, or rather, your infant pre-brain wiring algorithm, was effectively a 7.5 megabyte program—if it's actually 75 megabytes, that makes little difference to the argument. By exposing that 7.5 megabyte program to all the information coming in from your eyes, ears, nose, proprioceptive sensors telling you where your limbs were, and so on, your brain updated itself into forming the modern Ashley, whose hundred trillion synapses might be encoded by, say, one petabyte of information.

ASHLEY: The thought does occur to me that some environmental phenomena have effects on me that can't be interpreted as "sensory information" in any simple way, like the direct effect that alcohol has on my neurons, and how that feels to me from the inside. But it would be perverse to claim that this prevents you from trying to summarize all the information that the Ashley-agent receives into a single sequence, so I won't press the point.

(**ELIEZER:** (*whispering*) *More on this topic later.*)

ASHLEY: Oh, and for completeness's sake, wouldn't there also be further information embedded in the laws of physics themselves? Like, the way my brain executes implicitly says something about the laws of physics in the universe I'm in.

BLAINE: Metaphorically speaking, our laws of physics would play the role of a particular choice of Universal Turing Machine, which has some effect on which computations count as "simple" inside the Solomonoff formula. But normally, the UTM should be very simple compared to the amount of data in the sequence we're trying to predict, just like the laws of physics are very simple compared to a human brain. In terms of algorithmic complexity, the laws of physics are very simple compared to watching a 1920×1080 @60Hz visual field for a day.

ASHLEY: Part of my mind feels like the laws of physics are quite complicated compared to going outside and watching a sunset. Like, I realize that's false, but I'm not sure how to say out loud exactly why it's false...

BLAINE: Because the algorithmic complexity of a system isn't measured by how long a human has to go to college to understand it, it's measured by the size of the computer program required to generate it. The language of physics is differential equations, and it turns out that this is something difficult to beat into some human brains, but differential equations are simple to program into a simple Turing Machine.

ASHLEY: Right, like, the laws of physics actually have much fewer details to them than, say, human nature. At least on the Standard Model of Physics. I mean, in principle there could be another decillion undiscovered particle families out there.

BLAINE: The concept of "algorithmic complexity" isn't about seeing something with lots of gears and details, it's about the size of computer program required to compress all those details. The Mandelbrot set looks very complicated visually, you can keep zooming in using more and more detail, but there's a very simple rule that generates it, so we say the algorithmic complexity is very low.

ASHLEY: All the visual information I've seen is something that happens *within* the physical universe, so how can it be more complicated than the universe? I mean, I have a sense on some level that this shouldn't be a problem, but I don't know why it's not a problem.

BLAINE: That's because particular parts of the universe can have much higher algorithmic complexity than the entire universe!

Consider a library that contains all possible books. It's very easy to write a computer program that generates all possible books. So any *particular* book in the library contains much more algorithmic information than the *entire* library; it contains the information required to say 'look at this particular book here'.

If pi is normal, then somewhere in its digits is a copy of Shakespeare's *Hamlet*—but the number saying which particular digit of pi to start looking at, will be just about exactly as large as *Hamlet* itself. The copy of Shakespeare's *Hamlet* that exists in the decimal expansion of pi is more complex than pi itself.

If you zoomed way in and restricted your vision to a particular part of the Mandelbrot set, what you saw might be much more *algorithmically* complex than the entire Mandelbrot set, because the specification has to say where in the Mandelbrot set you are.

Similarly, the world Earth is much more algorithmically complex than the laws of physics. Likewise, the visual field you see over the course of a second can easily be far more algorithmically complex than the laws of physics.

ASHLEY: Okay, I think I get that. And similarly, even though the ways that proteins fold up are very complicated, in principle we could get all that info using just the simple fundamental laws of physics plus the relatively simple DNA code for the protein. There are all sorts of obvious caveats about epigenetics and so on, but those caveats aren't likely to change the numbers by a whole order of magnitude.

BLAINE: Right!

ASHLEY: So the laws of physics are, like, a few kilobytes, and my brain has say 75 megabytes of innate wiring instructions. And then I get to see a lot more information than that over my lifetime, like a megabit per second after my initial visual system finishes preprocessing it, and then most of that is forgotten. Uh... what does that have to do with Solomonoff induction again?

BLAINE: Solomonoff induction quickly catches up to any single computer program at sequence prediction, even if the original program is very large and contains a lot of prior information about the environment. If a program is 75 megabytes long, it can only predict 75 megabytes worth of data better than the Solomonoff inductor before the Solomonoff inductor catches up to it.

That doesn't mean that a Solomonoff inductor knows everything a baby does after the first second of exposure to a webcam feed, but it does mean that after the first second, the Solomonoff inductor is already no more surprised than a baby by the vast majority of pixels in the next frame.

Every time the Solomonoff inductor assigns half as much probability as the baby to the next pixel it sees, that's one bit spent permanently out of the 75 megabytes of error that can happen before the Solomonoff inductor catches up to the baby.

That your brain is written in the laws of physics also has some implicit correlation with the environment, but that's like saying that a program is written in the same programming language as the environment. The language can contribute something to the power of the program, and the environment being written in the same programming language can be a kind of prior knowledge. But if Solomonoff induction starts from a standard Universal Turing Machine as its language, that doesn't contribute any more bits of lifetime error than the complexity of that programming language in the UTM.

ASHLEY: Let me jump back a couple of steps and return to the notion of my brain wiring itself up in response to environmental information. I'd expect an important part of that process was my brain learning to *control* the environment, not just passively observing it. Like, it mattered to my brain's wiring algorithm that my brain saw the room shift in a certain way when it sent out signals telling my eyes to move.

BLAINE: Indeed. But talking about the sequential *control* problem is more complicated math. AIXI is the ideal agent that uses Solomonoff induction as its epistemology and expected reward as its decision theory. That introduces extra complexity, so it makes sense to talk about just Solomonoff induction first. We can talk about AIXI later. So imagine for the moment that we were *just* looking at your sensory data, and trying to predict what would come next in that.

ASHLEY: Wouldn't it make more sense to look at the brain's inputs *and* outputs, if we wanted to predict the next input? Not just look at the series of previous inputs?

BLAINE: It'd make the problem easier for a Solomonoff inductor to solve, sure; but it also makes the problem more complicated. Let's talk instead about what would happen if you took the complete sensory record of your life, gave it to an ideally smart agent, and asked the agent to predict what you would see next. Maybe the agent could do an even better job of prediction if we also told it about your brain's outputs, but I don't think that subtracting the outputs would leave it helpless to see patterns in the inputs.

ASHLEY: It sounds like a pretty hard problem to me, maybe even an unsolvable one. I'm thinking of the distinction in computer science between needing to learn from non-chosen data, versus learning when you can choose particular queries. Learning can be much faster in the second case.

BLAINE: In terms of what can be predicted *in principle* given the data, what facts are *actually reflected in it* that Solomonoff induction might uncover, we shouldn't imagine a human trying to analyze the data. We should imagine an entire advanced civilization pondering it for years. If you look at it from that angle, then the alien civilization isn't going to balk at the fact that it's looking at the answers to the queries that Ashley's brain chose, instead of the answers to the queries it chose itself.

Like, if the Ashley had already read Shakespeare's *Hamlet*—if the image of those pages had already crossed the sensory stream—and then the Ashley saw a mysterious box outputting 20, 15, 2, 5, 15, 18, I think somebody eavesdropping on that sensory data would be equally able to guess that this was encoding 'tobeor' and guess that the next thing the Ashley saw might be the box outputting 14. You wouldn't even need an entire alien civilization of superintelligent cryptographers to guess that. And it definitely wouldn't be a killer problem that Ashley was controlling the eyeball's saccades, even if you could learn even faster by controlling the eyeball yourself.

So far as the computer-science distinction goes, Ashley's eyeball *is* being controlled to make intelligent queries and seek out useful information; it's just Ashley controlling the eyeball instead of you—that eyeball is not a query-oracle answering *random* questions.

ASHLEY: Okay, I think this example is helping my understanding of what we're doing here. In the case above, the next item in the Ashley-sequence wouldn't actually be 14. It would be this huge 1920 × 1080 visual field that showed the box flashing a little picture of '14'.

BLAINE: Sure. Otherwise it would be a rigged demo, as you say.

ASHLEY: I think I'm confused about the idea of *predicting* the visual field. It seems to me that what with all the dust specks in my visual field, and maybe my deciding to tilt my head using motor instructions that won't appear in the sequence, there's no way to *exactly* predict the 66-megabit integer representing the next visual frame. So it must be doing something other than the equivalent of guessing "14" in a simpler sequence, but I'm not sure what.

BLAINE: Indeed, there'd be some element of thermodynamic and quantum randomness preventing that exact prediction even in principle. So instead of predicting one particular next frame, we put a probability distribution on it.

ASHLEY: A probability distribution over possible 66-megabit frames? Like, a table with $2^{66,000,000}$ entries, summing to 1?

BLAINE: Sure. $2^{32 \times 1920 \times 1080}$ isn't a large number when you have unlimited computing power. As Martin Gardner once observed, "Most finite numbers are very much larger." Like I said, Solomonoff induction is an epistemic ideal that requires an unreasonably large amount of computing power.

ASHLEY: I don't deny that big computations can sometimes help us understand little ones. But at the point when we're talking about probability distributions that large, I have some trouble holding onto what the probability distribution is supposed to *mean*.

BLAINE: Really? Just imagine a probability distribution over N possibilities, then let N go to $2^{66,000,000}$. If we were talking about a letter ranging from A to Z, then putting 100 times as much probability mass on (X, Y, Z) as on the rest of the alphabet, would say that although you didn't know *exactly* what letter would happen, you expected it would be toward the end of the alphabet. You would have used 26 probabilities, summing to 1, to precisely state that prediction.

In Solomonoff induction, since we have unlimited computing power, we express our uncertainty about a 1920×1080 video frame the same way. All the various pixel fields you could see if your eye jumped to a plausible place, saw a plausible number of dust specks, and saw the box flash something that visually encoded '14', would have high probability. Pixel fields where the box vanished and was replaced with a glow-in-the-dark unicorn would have very low, though not zero, probability.

ASHLEY: Can we really get away with viewing things that way?

BLAINE: If we could not make identifications like these *in principle*, there would be no principled way in which we could say that you had ever *expected to see something happen*—no way to say that one visual field your eyes saw had higher probability than any other sensory experience. We couldn't justify science; we couldn't say that, having performed Galileo's experiment by rolling an inclined cylinder down a plane, Galileo's theory was thereby to some degree supported by having assigned *a high relative probability* to the only actual observations our eyes ever report.

ASHLEY: I feel a little unsure of that jump, but I suppose I can go along with that for now. Then the question of "What probability does Solomonoff induction assign to Canada invading?" is to be identified, in principle, with the question "Given my past life experiences and all the visual information that's entered my eyes, what is the relative probability of seeing visual information that encodes Google News with the headline 'CANADA INVADES USA' at some point during the next 300 million seconds?"

BLAINE: Right!

ASHLEY: And Solomonoff induction has an in-principle way of assigning this a relatively low probability, which that online crackpot could do well to learn from as a matter of principle, even if he couldn't *begin* to carry out the exact calculations that involve assigning probabilities to exponentially vast tables.

BLAINE: Precisely!

ASHLEY: Fairness requires that I congratulate you on having come further in formalizing 'do good epistemology' as a sequence prediction problem than I previously thought you might.

I mean, you haven't satisfied me yet, but I wasn't expecting you to get even this far.

iii. Hypotheses

BLAINE: Next, we consider how to represent a *hypothesis* inside this formalism.

ASHLEY: Hmm. You said something earlier about updating on a probabilistic mixture of computer programs, which leads me to suspect that in this formalism, a hypothesis or *way the world can be* is a computer program that outputs a sequence of integers.

BLAINE: There's indeed a version of Solomonoff induction that works like that. But I prefer the version where a hypothesis assigns *probabilities* to sequences. Like, if the hypothesis is that the world is a fair coin, then we shouldn't try to make that hypothesis predict "heads—tails—tails—tails—heads" but should let it just assign a 1/32 prior probability to the sequence **HTTTH**.

ASHLEY: I can see that for coins, but I feel a bit iffier on what this means as a statement *about the real world*.

BLAINE: A single hypothesis inside the Solomonoff mixture would be a computer program that took in a series of video frames, and assigned a probability to each possible next video frame. Or for greater simplicity and elegance, imagine a program that took in a sequence of bits, ones and zeroes, and output a rational number for the probability of the next bit being '1'. We can readily go back and forth between a program like that, and a probability distribution over sequences.

Like, if you can answer all of the questions, "What's the probability that the coin comes up heads on the first flip?", "What's the probability of the coin coming up heads on the second flip, if it came up heads on the first flip?", and "What's the probability that the coin comes up heads on the second flip, if it came up tails on the first flip?" then we can turn that into a probability distribution over sequences of two coinflips. Analogously, if we have a program that outputs the probability of the next bit, conditioned on a finite number of previous bits taken as input, that program corresponds to a probability distribution over infinite sequences of bits.

$$P_{prog}(bits_{1 \dots N}) = \prod_{i=1}^{N} InterpretProb(prog(bits_{1 \dots i-1}), bits_i)$$

$$InterpretProb(prog(x), y) = \begin{cases} InterpretFrac(prog(x)) & \text{if } y = 1 \\ 1 - InterpretFrac(prog(x)) & \text{if } y = 0 \\ 0 & \text{if } prog(x) \text{ does not halt} \end{cases}$$

ASHLEY: I think I followed along with that in theory, though it's not a type of math I'm used to (yet). So then in what sense is a program that assigns probabilities to sequences, a way the world could be—a hypothesis about the world?

BLAINE: Well, I mean, for one thing, we can see the infant Ashley as a program with 75 megabytes of information about how to wire up its brain in response to sense data, that sees a bunch of sense data, and then experiences some degree of relative surprise. Like in the baby-looking-paradigm experiments where you show a baby an object disappearing behind a screen, and the baby looks longer at those cases, and so we suspect that babies have a concept of object permanence.

ASHLEY: That sounds like a program that's a way Ashley could be, not a program that's a way the world could be.

BLAINE: Those indeed are dual perspectives on the meaning of Solomonoff induction. Maybe we can shed some light on this by considering a simpler induction rule, Laplace's Rule of Succession, invented by the Reverend Thomas Bayes in the 1750s, and named after Pierre-Simon Laplace, the inventor of Bayesian reasoning.

ASHLEY: Pardon me?

BLAINE: Suppose you have a biased coin with an unknown bias, and every possible bias between 0 and 1 is equally probable.

ASHLEY: Okay. Though in the real world, it's quite likely that an unknown frequency is exactly 0, 1, or 1/2. If you assign equal probability density to every part of the real number field between 0 and 1, the probability of 1 is 0. Indeed, the probability of all rational numbers put together is zero.

BLAINE: The original problem considered by Thomas Bayes was about an ideal billiard ball bouncing back and forth on an ideal billiard table many times and eventually slowing to a halt; and then bouncing other billiards to see if they halted to the left or the right of the first billiard. You can see why, in first considering the simplest form of this problem without any complications, we might consider every position of the first billiard to be equally probable.

ASHLEY: Sure. Though I note with pointless pedantry that if the billiard was really an ideal rolling sphere and the walls were perfectly reflective, it'd never halt in the first place.

BLAINE: Suppose we're told that, after rolling the original billiard ball and then 5 more billiard balls, one billiard ball was to the right of the original, an **R**. The other four were to the left of the original, or **L**s. Again, that's 1 **R** and 4 **L**s. Given only this data, what is the probability that the next billiard ball rolled will be on the left of the original, another **L**?

ASHLEY: Five sevenths.

BLAINE: Ah, you've heard this problem before?

ASHLEY: No, but it's obvious.

BLAINE: Uh... really?

ASHLEY: Combinatorics. Consider just the orderings of the balls, instead of their exact positions. Designate the original ball with the symbol **❚**, the next five balls as **LLLLR**, and the next ball to be rolled as **✚**. Given that the current ordering of these six balls is **LLLL❚R** and that all positions and spacings of the underlying balls are equally likely, after rolling the **✚**, there will be seven equally likely orderings **✚LLLL❚R**, **L✚LLL❚R**, **LL✚LL❚R**, and so on up to **LLLL❚✚R** and **LLLL❚R✚**. In five of those seven orderings, the **✚** is on the left of the **❚**. In general, if we see M of **L** and N of **R**, the probability of the next item being an **L** is $(M + 1)/(M + N + 2)$.

BLAINE: Gosh... Well, the much more complicated proof originally devised by Thomas Bayes starts by considering every position of the original ball to be equally likely *a priori*, the additional balls as providing evidence about that position, and then integrating over the posterior probabilities of the original ball's possible positions to arrive at the probability that the next ball lands on the left or right.

ASHLEY: Heh. And is all that extra work useful if you also happen to know a little combinatorics?

BLAINE: Well, it tells me exactly how my beliefs about the original ball change with each new piece of evidence—the new posterior probability function on the ball's position. Suppose I instead asked you something along the lines of, "Given 4 **L** and 1 **R**, where do you think the original ball **✚** is most likely to be on the number line? How likely is it to be within 0.1 distance of there?"

ASHLEY: That's fair; I don't see a combinatoric answer for the later part. You'd have to actually integrate over the density function $f^M(1 - f)^N \, df$.

BLAINE: Anyway, let's just take at face value that Laplace's Rule of Succession says that, after observing M 1s and N 0s, the probability of getting a 1 next is $(M + 1)/(M + N + 2)$.

ASHLEY: But of course.

BLAINE: We can consider Laplace's Rule as a short Python program that takes in a sequence of 1s and 0s, and spits out the probability that the next bit in the sequence will be 1. We can also consider it as a probability distribution over infinite sequences, like this:

- **0** : 1/2
- **1** : 1/2
- **00** : $1/2 * 2/3 = 1/3$
- **01** : $1/2 * 1/3 = 1/6$
- **000** : $1/2 * 2/3 * 3/4 = 1/4$
- **001** : $1/2 * 2/3 * 1/4 = 1/12$
- **010** : $1/2 * 1/3 * 1/2 = 1/12$

... and so on.

Now, we can view this as a rule someone might espouse for *predicting* coinflips, but also view it as corresponding to a particular class of possible worlds containing randomness.

I mean, Laplace's Rule isn't the only rule you could use. Suppose I had a barrel containing ten white balls and ten green balls. If you already knew this about the barrel, then after seeing M white balls and N green balls, you'd predict the next ball being white with probability $(10 - M)/(20 - M - N)$.

If you use Laplace's Rule, that's like believing the world was like a billiards table with an original ball rolling to a stop at a random point and new balls ending up on the left or right. If you use $(10 - M)/(20 - M - N)$, that's like the hypothesis that there are ten green balls and ten white balls in a barrel. There isn't really a sharp border between rules we can use to predict the world, and rules for how the world behaves—

ASHLEY: Well, that sounds just plain wrong. The map is not the territory, don'cha know? If Solomonoff induction can't tell the difference between maps and territories, maybe it doesn't contain all epistemological goodness after all.

BLAINE: Maybe it'd be better to say that there's a dualism between good ways of computing predictions and being in actual worlds where that kind of predicting works well? Like, you could also see Laplace's Rule as implementing the rules for a world with randomness where the original billiard ball ends up in a random place, so that the first thing you see is equally likely to be 1 or 0. Then to ask what probably happens on round 2, we tell the world what happened on round 1 so that it can update what the background random events were.

ASHLEY: Mmmaybe.

BLAINE: If you go with the version where Solomonoff induction is over programs that just spit out a determined string of ones and zeroes, we could see those programs as corresponding to particular environments—ways the world *could be* that would produce our sensory input, the sequence.

We could jump ahead and consider the more sophisticated decision-problem that appears in AIXI: an environment is a program that takes your motor outputs as its input, and then returns your sensory inputs as its output. Then we can see a program that produces Bayesian-updated predictions as corresponding to a hypothetical probabilistic environment that implies those updates, although they'll be conjugate systems rather than mirror images.

ASHLEY: Did you say something earlier about the deterministic and probabilistic versions of Solomonoff induction giving the same answers? Like, is it a distinction without a difference whether we ask about simple programs that reproduce the observed data versus simple programs that assign high probability to the data? I can't see why that should be true, especially since Turing machines don't include a randomness source.

BLAINE: I'm *told* the answers are the same but I confess I can't quite see why, unless there's some added assumption I'm missing. So let's talk about programs that assign probabilities for now, because I think that case is clearer.

iv. Simplicity

BLAINE: The next key idea is to prefer *simple* programs that assign high probability to our observations so far.

ASHLEY: It seems like an obvious step, especially considering that you were already talking about "simple programs" and Occam's Razor a while back. Solomonoff induction is part of the Bayesian program of inference, right?

BLAINE: Indeed. Very much so.

ASHLEY: Okay, so let's talk about the program, or hypothesis, for "This barrel has an unknown frequency of white and green balls", versus the hypothesis "This barrel has 10 white and 10 green balls", versus the hypothesis, "This barrel always puts out a green ball after a white ball and vice versa."

Let's say we see a green ball, then a white ball, the sequence **GW**. The first hypothesis assigns this probability $1/2 * 1/3 = 1/6$, the second hypothesis assigns this probability $10/20 * 9/19$ or roughly $1/4$, and the third hypothesis assigns probability $1/2 * 1$.

Now it seems to me that there's some important sense in which, even though Laplace's Rule assigned a lower probability to the data, it's significantly simpler than the second and third hypotheses and is the wiser answer. Does Solomonoff induction agree?

BLAINE: I think you might be taking into account some prior knowledge that isn't in the sequence itself, there. Like, things that alternate either **101010...** or **010101...** are *objectively* simple in the sense that a short computer program simulates them or assigns probabilities to them. It's just unlikely to be true about an actual barrel of white and green balls.

If **10** is literally the first sense data that you ever see, when you are a fresh new intelligence with only two bits to rub together, then "The universe consists of alternating bits" is no less reasonable than "The universe produces bits with an unknown random frequency anywhere between 0 and 1."

ASHLEY: Conceded. But as I was going to say, we have three hypotheses that assigned $1/6$, $\sim 1/4$, and $1/2$ to the observed data; but to know the posterior probabilities of these hypotheses we need to actually say how relatively likely they were *a priori*, so we can multiply by the odds ratio. Like, if the prior odds were $3:2:1$, the posterior odds would be $3:2:1 * (2/12 : 3/12 : 6/12) = 3:2:1 * 2:3:6 = 6:6:6 = 1:1:1$. Now, how would Solomonoff induction assign prior probabilities to those computer programs? Because I remember you saying, way back when, that you thought Solomonoff was the answer to "How should Bayesians assign priors?"

BLAINE: Well, how would you do it?

ASHLEY: I mean... yes, the simpler rules should be favored, but it seems to me that there's some deep questions as to the exact relative 'simplicity' of the rules $(M+1)/(M+N+2)$, or the rule $(10-M)/(20-M-N)$, or the rule "alternate the bits"...

BLAINE: Suppose I ask you to just make up some simple rule.

ASHLEY: Okay, if I just say the rule I think you're looking for, the rule would be, "The complexity of a computer program is the number of bits needed to specify it to some arbitrary but reasonable choice of compiler or Universal Turing Machine, and the prior probability is 1/2 to the power of the number of bits. Since, e.g., there's 32 possible 5-bit programs, so each such program has probability 1/32. So if it takes 16 bits to specify Laplace's Rule of Succession, which seems a tad optimistic, then the prior probability would be 1/65536, which seems a tad pessimistic.

BLAINE: Now just apply that rule to the infinity of possible computer programs that assign probabilities to the observed data, update their posterior probabilities based on the probability they've assigned to the evidence so far, sum over all of them to get your next prediction, and we're done. And yes, that requires a hypercomputer that can solve the halting problem, but we're talking ideals here. Let P be the set of all programs and $s_1 s_2 \ldots s_n$ also written $s_{\leq n}$ be the sense data so far, then

$$\mathrm{Sol}(s_{\leq n}) := \sum_{\mathrm{prog} \in P} 2^{-\mathrm{length}(\mathrm{prog})} \cdot \prod_{j=1}^{n} \mathrm{InterpretProb}(\mathrm{prog}(s_{\leq j-1}), s_j)$$

$$P(s_{n+1} = 1 \mid s_{\leq n}) = \frac{\mathrm{Sol}(s_1 s_2 \ldots s_n 1)}{\mathrm{Sol}(s_1 s_2 \ldots s_n 1) + \mathrm{Sol}(s_1 s_2 \ldots s_n 0)}$$

ASHLEY: Uh.

BLAINE: Yes?

ASHLEY: Um...

BLAINE: What is it?

ASHLEY: You invoked a countably infinite set, so I'm trying to figure out if my predicted probability for the next bit must necessarily converge to a limit as I consider increasingly large finite subsets in any order.

BLAINE: *(sighs)* Of course you are.

ASHLEY: I think you might have left out some important caveats. Like, if I take the rule literally, then the program "**0**" has probability 1/2, the program "**1**" has probability 1/2, the program "**01**" has probability 1/4 and now the total probability is 1.25 which is *too much.* So I can't actually normalize it because the series sums to infinity. Now, this just means we need to, say, decide that the probability of a program having length 1 is 1/2, the probability of it having length 2 is 1/4, and so on out to infinity, but it's an added postulate.

BLAINE: The conventional method is to require a prefix-free code. If "**0111**" is a valid program then "**01110**" cannot be a valid program. With that constraint, assigning "1/2 to the power of the length of the code", to all valid codes, will sum to less than 1; and we can normalize their relative probabilities to get the actual prior.

ASHLEY: Okay. And you're sure that it doesn't matter in what order we consider more and more programs as we approach the limit, because... no, I see it. Every program has positive probability mass, with the total set summing to 1, and Bayesian updating doesn't change that. So as I consider more and more programs, in any order, there are only so many large contributions that can be made from the mix—there's only so often that the final probability can change.

Like, let's say there are at most 99 programs with probability 1% that assign probability 0 to the next bit being a 1; that's only 99 times the final answer can go down by as much as 0.01, as the limit is approached.

BLAINE: This idea generalizes, and is important. List all possible computer programs, in any order you like. Use any definition of *simplicity* that you like, so long as for any given amount of simplicity, there are only a finite number of computer programs that simple. As you go on carving off chunks of prior probability mass and assigning them to programs, it *must* be the case that as programs get more and complicated, their prior probability approaches zero!—though it's still positive for every finite program, because of Cromwell's Rule.

You can't have more than 99 programs assigned 1% prior probability and still obey Cromwell's Rule, which means there must be some *most complex* program that is assigned 1% probability, which means every more complicated program must have less than 1% probability out to the end of the infinite list.

ASHLEY: Huh. I don't think I've ever heard that justification for Occam's Razor before. I think I like it. I mean, I've heard a lot of appeals to the empirical simplicity of the world, and so on, but this is the first time I've seen a *logical* proof that, in the limit, more complicated hypotheses *must* be less likely than simple ones.

BLAINE: Behold the awesomeness that is Solomonoff Induction!

ASHLEY: Uh, but you didn't actually use the notion of *computational* simplicity to get that conclusion; you just required that the supply of probability mass is finite and the supply of potential complications is infinite. Any way of counting discrete complications would imply that conclusion, even if it went by surface wheels and gears.

BLAINE: Well, maybe. But it so happens that Yudkowsky did invent or reinvent that argument after pondering Solomonoff induction, and if it predates him (or Solomonoff) then Yudkowsky doesn't know the source. Concrete inspiration for simplified arguments is also a credit to a theory, especially if the simplified argument didn't exist before that.

ASHLEY: Fair enough.

v. Choice of Universal Turing Machine

ASHLEY: My next question is about the choice of Universal Turing Machine—the choice of compiler for our program codes. There's an infinite number of possibilities there, and in principle, the right choice of compiler can make our probability for the next thing we'll see be anything we like. At least I'd expect this to be the case, based on how the "problem of induction" usually goes. So with the right choice of Universal Turing Machine, our online crackpot can still make it be the case that Solomonoff induction predicts Canada invading the USA.

BLAINE: One way of looking at the problem of good epistemology, I'd say, is that the job of a good epistemology is not to make it *impossible* to err. You can still blow off your foot if you really insist on pointing the shotgun at your foot and pulling the trigger.

The job of good epistemology is to make it *more obvious* when you're about to blow your own foot off with a shotgun. On this dimension, Solomonoff induction excels. If you claim that we ought to pick an enormously complicated compiler to encode our hypotheses, in order to make the 'simplest hypothesis that fits the evidence' be one that predicts Canada invading the USA, then it should be obvious to everyone except you that you are in the process of screwing up.

ASHLEY: Ah, but of course they'll say that their code is just the simple and natural choice of Universal Turing Machine, because they'll exhibit a meta-UTM which outputs that UTM given only a short code. And if you say the meta-UTM is complicated—

BLAINE: Flon's Law says, "There is not now, nor has there ever been, nor will there ever be, any programming language in which it is the least bit difficult to write bad code." You can't make it impossible for people to screw up, but you can make it *more obvious.* And Solomonoff induction would make it even more obvious than might at first be obvious, because—

ASHLEY: Your Honor, I move to have the previous sentence taken out and shot.

BLAINE: Let's say that the whole of your sensory information is the string **10101010...** Consider the stupid hypothesis, "This program has a 99% probability of producing a **1** on every turn", which you jumped to after seeing the first bit. What would you need to claim your priors were like—what Universal Turing Machine would you need to endorse—in order to maintain blind faith in that hypothesis in the face of ever-mounting evidence?

ASHLEY: You'd need a Universal Turing Machine **blind-utm** that assigned a very high probability to the **blind** program "def ProbNextElementIsOne(previous_sequence): return 0.99". Like, if **blind-utm** sees the code **0**, it executes the **blind** program "return 0.99".

And to defend yourself against charges that your UTM **blind-utm** was not itself simple, you'd need a meta-UTM, **blind-meta**, which, when it sees the code **10**, executes **blind-utm**.

And to really wrap it up, you'd need to take a fixed point through all towers of meta and use diagonalization to create the UTM **blind-diag** that, when it sees the program code **0**, executes "return 0.99", and when it sees the program code **10**, executes **blind-diag**.

I guess I can see some sense in which, even if that doesn't resolve Hume's problem of induction, anyone *actually advocating that* would be committing blatant shenanigans on a commonsense level, arguably more blatant than it would have been if we hadn't made them present the UTM.

BLAINE: Actually, the shenanigans have to be much worse than that in order to fool Solomonoff induction. Like, Solomonoff induction using your **blind-diag** isn't fooled for a minute, even taking **blind-diag** entirely on its own terms.

ASHLEY: Really?

BLAINE: Assuming 60 sequence items per second? Yes, absolutely, Solomonoff induction shrugs off the delusion in the first minute, unless there are further and even more blatant shenanigans.

We did require that your **blind-diag** be a *Universal* Turing Machine, meaning that it can reproduce every computable probability distribution over sequences, given some particular code to compile. Let's say there's a 200-bit code **laplace** for Laplace's Rule of Succession, "lambda sequence: return (sequence.count('1') + 1) / (len(sequence) + 2)", so that its prior probability relative to the 1-bit code for **blind** is 2^{-200}. Let's say that the sense data is around 50/50 1s and 0s. Every time we see a 1, **blind** gains a factor of 2 over **laplace** (99% vs. 50% probability), and every time we see a 0, **blind** loses a factor of 50 over **laplace** (1% vs. 50% probability).

On average, every 2 bits of the sequence, **blind** is losing a factor of 25 or, say, a bit more than 4 bits, i.e., on average **blind** is losing two bits of probability per element of the sequence observed.

So it's only going to take 100 bits, or a little less than two seconds, for **laplace** to win out over **blind**.

ASHLEY: I see. I was focusing on a UTM that assigned lots of prior probability to **blind**, but what I really needed was a compiler that, *while still being universal* and encoding every possibility somewhere, still assigned a really tiny probability to **laplace**, **faircoin** that encodes "return 0.5", and every other hypothesis that does better, round by round, than **blind**. So what I really need to carry off the delusion is **obstinate-diag** that is universal, assigns high probability to **blind**, requires billions of bits to specify **laplace**, and also requires billions of bits to specify any UTM that can execute **laplace** as a shorter code than billions of bits. Because otherwise we will say, "Ah, but given the evidence, this other UTM would have done better." I agree that those are even more blatant shenanigans than I thought.

BLAINE: Yes. And even *then*, even if your UTM takes two billion bits to specify **faircoin**, Solomonoff induction will lose its faith in **blind** after seeing a billion bits.

Which will happen before the first year is out, if we're getting 60 bits per second.

And if you turn around and say, "Oh, well, I didn't mean *that* was my UTM, I really meant *this* was my UTM, this thing over here where it takes a *trillion* bits to encode **faircoin**", then that's probability-theory-violating shenanigans where you're changing your priors as you go.

ASHLEY: That's actually a very interesting point—that what's needed for a Bayesian to maintain a delusion in the face of mounting evidence is not so much a blindly high prior for the delusory hypothesis, as a blind skepticism of all its alternatives.

But what if their UTM requires a googol bits to specify **faircoin**? What if **blind** and **blind-diag**, or programs pretty much isomorphic to them, are the only programs that can be specified in less than a googol bits?

BLAINE: Then your desire to shoot your own foot off has been made very, very visible to anyone who understands Solomonoff induction. We're not going to get absolutely objective prior probabilities as a matter of logical deduction, not without principles that are unknown to me and beyond the scope of Solomonoff induction. But we can make the stupidity really *blatant* and force you to construct a downright embarrassing Universal Turing Machine.

ASHLEY: I guess I can see that. I mean, I guess that if you're presenting a ludicrously complicated Universal Turing Machine that just refuses to encode the program that would predict Canada not invading, that's more *visibly* silly than a verbal appeal that says, "But you must just have faith that Canada will invade." I guess part of me is still hoping for a more objective sense of "complicated".

BLAINE: We could say that reasonable UTMs should contain a small number of wheels and gears in a material instantiation under our universe's laws of physics, which might in some ultimate sense provide a prior over priors. Like, the human brain evolved from DNA-based specifications, and the things you can construct out of relatively small numbers of physical objects are 'simple' under the 'prior' implicitly searched by natural selection.

ASHLEY: Ah, but what if I think it's likely that our physical universe or the search space of DNA won't give us a good idea of what's complicated?

BLAINE: For your alternative notion of what's complicated to go on being believed even as other hypotheses are racking up better experimental predictions, you need to assign a *ludicrously low probability* that our universe's space of physical systems buildable using a small number of objects, could *possibly* provide better predictions of that universe than your complicated alternative notion of prior probability.

We don't need to appeal that it's *a priori* more likely than not that "a universe can be predicted well by low-object-number machines built using that universe's physics." Instead, we appeal that it would violate Cromwell's Rule, and would constitute exceedingly special pleading, to assign the possibility of a physically learnable universe a probability of *less* than $2^{-1,000,000}$. It then takes only a megabit of exposure to notice that the universe seems to be regular.

ASHLEY: In other words, so long as you don't start with an absolute and blind prejudice against the universe being predictable by simple machines encoded in our universe's physics—so long as, on this planet of seven billion people, you don't assign probabilities less than $2^{-1,000,000}$ to the other person being right about what is a good Universal Turing Machine—then the pure logic of Bayesian updating will rapidly force you to the conclusion that induction works.

vi. Why algorithmic complexity?

ASHLEY: Hm. I don't know that good *pragmatic* answers to the problem of induction were ever in short supply. Still, on the margins, it's a more forceful pragmatic answer than the last one I remember hearing.

BLAINE: Yay! *Now* isn't Solomonoff induction wonderful?

ASHLEY: Maybe?

You didn't really use the principle of *computational* simplicity to derive that lesson. You just used that *some inductive principle* ought to have a prior probability of more than $2^{-1,000,000}$.

BLAINE: ...

ASHLEY: Can you give me an example of a problem where the *computational* definition of simplicity matters and can't be factored back out of an argument?

BLAINE: As it happens, yes I can. I can give you *three* examples of how it matters.

ASHLEY: Vun... two... three! Three examples! Ah-ah-ah!

BLAINE: Must you do that every—oh, never mind. Example one is that galaxies are not so improbable that no one could ever believe in them, example two is that the limits of possibility include Terrence Tao, and example three is that diffraction is a simpler explanation of rainbows than divine intervention.

ASHLEY: These statements are all so obvious that no further explanation of any of them is required.

BLAINE: On the contrary! And I'll start with example one. Back when the Andromeda Galaxy was a hazy mist seen through a telescope, and someone first suggested that maybe that hazy mist was an incredibly large number of distant stars—that many "nebulae" were actually *distant galaxies*, and our own Milky Way was only one of them—there was a time when Occam's Razor was invoked against that hypothesis.

ASHLEY: What? Why?

BLAINE: They invoked Occam's Razor against the galactic hypothesis, because if that were the case, then there would be *a much huger number of stars* in the universe, and the stars would be entities, and Occam's Razor said "Entities are not to be multiplied beyond necessity."

ASHLEY: That's not how Occam's Razor works. The "entities" of a theory are its types, not its objects. If you say that the hazy mists are distant galaxies of stars, then you've reduced the number of laws because you're just postulating a previously seen type, namely stars organized into galaxies, instead of a new type of hazy astronomical mist.

BLAINE: Okay, but imagine that it's the nineteenth century and somebody replies to you, "Well, I disagree! William of Ockham said not to multiply entities, this galactic hypothesis obviously creates a huge number of entities, and that's the way I see it!"

ASHLEY: I think I'd give them your spiel about there being no human epistemology that can stop you from shooting off your own foot.

BLAINE: I *don't* think you'd be justified in giving them that lecture.

I'll parenthesize at this point that you ought to be very careful when you say "I can't stop you from shooting off your own foot", lest it become a Fully General Scornful Rejoinder. Like, if you say that to someone, you'd better be able to explain exactly why Occam's Razor counts types as entities but not objects. In fact, you'd better explain that to someone *before* you go advising them not to shoot off their own foot. And once you've told them what you think is foolish and why, you might as well stop there. Except in really weird cases of people presenting us with enormously complicated and jury-rigged Universal Turing Machines, and then we say the shotgun thing.

ASHLEY: That's fair. So, I'm not sure what I'd have answered before starting this conversation, which is much to your credit, friend Blaine. But now that I've had this conversation, it's obvious that it's new types and not new objects that use up the probability mass we need to distribute over all hypotheses. Like, I need to distribute my probability mass over "Hypothesis 1: there are stars" and "Hypothesis 2: there are stars plus huge distant hazy mists". I don't need to distribute my probability mass over all the actual stars in the galaxy!

BLAINE: In terms of Solomonoff induction, we penalize a program's *lines of code* rather than its *runtime* or *RAM used*, because we need to distribute our probability mass over possible alternatives each time we add a line of code. There's no corresponding *choice between mutually exclusive alternatives* when a program uses more runtime or RAM.

(**ELIEZER:** (*whispering*) *Unless we need a leverage prior to consider the hypothesis of being a particular agent inside all that RAM or runtime.*)

ASHLEY: Or to put it another way: any fully detailed model of the universe would require some particular arrangement of stars, and the more stars there are, the more possible arrangements there are. But when we look through the telescope and see a hazy mist, we get to sum over all arrangements of stars that would produce that hazy mist. If some galactic hypothesis required a hundred billion stars to *all* be in *particular exact places* without further explanation or cause, then that would indeed be a grave improbability.

BLAINE: Precisely. And if you needed all the hundred billion stars to be in particular exact places, that's just the kind of hypothesis that would take a huge computer program to specify.

ASHLEY: But does it really require learning Solomonoff induction to understand that point? Maybe the bad argument against galaxies was just a motivated error somebody made in the nineteenth century, because they didn't want to live in a big universe for emotional reasons.

BLAINE: The same debate is playing out today over no-collapse versions of quantum mechanics, also somewhat unfortunately known as "many-worlds interpretations". Now, regardless of what anyone thinks of all the other parts of that debate, there's a *particular* sub-argument where somebody says, "It's simpler to have a collapse interpretation because all those extra quantum 'worlds' are extra entities that are unnecessary under Occam's Razor since we can't see them." And Solomonoff induction tells us that this invocation of Occam's Razor is flatly misguided because Occam's Razor does not work like that.

Basically, they're trying to cut down the RAM and runtime of the universe, at the expensive of adding an extra line of code, namely the code for the collapse postulate that prunes off parts of the wavefunction that are in undetectably weak causal contact with us.

ASHLEY: Hmm. Now that you put it that way, it's not so obvious to me that it makes sense to have *no* prejudice against *sufficiently* enormous universes. I mean, the universe we see around us is exponentially vast but not superexponentially vast—the visible atoms are 10^{80} in number or so, not $10^{10^{80}}$ or "bigger than Graham's Number". Maybe there's some fundamental limit on how much gets computed.

BLAINE: You, um, know that on the Standard Model, the universe doesn't just cut out and stop existing at the point where our telescopes stop seeing it? There isn't a giant void surrounding a little bubble of matter centered perfectly on Earth? It calls for a literally infinite amount of matter? I mean, I guess if you don't like living in a universe with more than 10^{80} entities, a universe where *too much gets computed,* you could try to specify *extra laws of physics* that create an abrupt spatial boundary with no further matter beyond them, somewhere out past where our telescopes can see—

ASHLEY: All right, point taken.

(**ELIEZER:** (*whispering*) *Though I personally suspect that the spatial multiverse and the quantum multiverse are the same multiverse, and that what lies beyond the reach of our telescopes is not entangled with us—meaning that the universe is as finitely large as the superposition of all possible quantum branches, rather than being literally infinite in space.*)

BLAINE: I mean, there is in fact an alternative formalism to Solomonoff induction, namely Levin search, which says that program complexities are further penalized by the logarithm of their runtime. In other words, it would say that 'explanations' or 'universes' that require a long time to run are inherently less probable.

Some people like Levin search more than Solomonoff induction because it's more computable. I dislike Levin search because (a) it has no fundamental epistemic justification and (b) it assigns probability zero to quantum mechanics.

ASHLEY: Can you unpack that last part?

BLAINE: If, as is currently suspected, there's no way to simulate quantum computers using classical computers without an exponential slowdown, then even in principle, this universe requires exponentially vast amounts of classical computing power to simulate.

Let's say that with sufficiently advanced technology, you can build a quantum computer with a million qubits. On Levin's definition of complexity, for the universe to be like that is as improbable *a priori* as *any particular* set of laws of physics that must specify on the order of one million equations.

Can you imagine how improbable it would be to see a list of one hundred thousand differential equations, without any justification or evidence attached, and be told that they were the laws of physics? That's the kind of penalty that Levin search or Schmidhuber's Speed Prior would attach to any laws of physics that could run a quantum computation of a million qubits, or, heck, any physics that claimed that a protein was being folded in a way that ultimately went through considering millions of quarks interacting.

If you're *not* absolutely certain *a priori* that the universe *isn't* like that, you don't believe in Schmidhuber's Speed Prior. Even with a collapse postulate, the amount of computation that goes on before a collapse would be prohibited by the

Speed Prior.

ASHLEY: Okay, yeah. If you're phrasing it that way—that the Speed Prior assigns probability nearly zero to quantum mechanics, so we shouldn't believe in the Speed Prior—then I can't easily see a way to extract out the same point without making reference to ideas like penalizing algorithmic complexity but not penalizing runtime. I mean, maybe I could extract the lesson back out but it's easier to say, or more obvious, by pointing to the idea that Occam's Razor should penalize algorithmic complexity but not runtime.

BLAINE: And that isn't just *implied by* Solomonoff induction, it's pretty much the whole idea of Solomonoff induction, right?

ASHLEY: Maaaybe.

BLAINE: For example two, that Solomonoff induction outperforms even Terence Tao, we want to have a theorem that says Solomonoff induction catches up to every computable way of reasoning in the limit. Since we iterated through all possible computer programs, we know that somewhere in there is a simulated copy of Terence Tao in a simulated room, and if this requires a petabyte to specify, then we shouldn't have to make more than a quadrillion bits of error relative to Terence Tao before zeroing in on the Terence Tao hypothesis.

I mean, in practice, I'd expect far less than a quadrillion bits of error before the system was behaving like it was vastly smarter than Terence Tao. It'd take a lot less than a quadrillion bits to give you some specification of a universe with simple physics that gave rise to a civilization of vastly greater than intergalactic extent. Like, Graham's Number is a very simple number, so it's easy to specify a universe that runs for that long before it returns an answer. It's not obvious how you'd extract Solomonoff predictions from that civilization and incentivize them to make good ones, but I'd be surprised if there were no Turing machine of fewer than one thousand states which did that somehow.

ASHLEY: ...

BLAINE: And for all I know there might be even better ways than that of getting exceptionally good predictions, somewhere in the list of the first decillion computer programs. That is, somewhere in the first 100 bits.

ASHLEY: So your basic argument is, "Never mind Terence Tao, Solomonoff induction dominates *God*."

BLAINE: Solomonoff induction isn't the epistemic prediction capability of a superintelligence. It's the epistemic prediction capability of something that eats superintelligences like potato chips.

ASHLEY: Is there any point to contemplating an epistemology so powerful that it will never begin to fit inside the universe?

BLAINE: Maybe? I mean, a lot of times, you just find people *failing to respect* the notion of ordinary superintelligence, doing the equivalent of supposing that a superintelligence behaves like a bad Hollywood genius and misses obvious-seeming moves. And a lot of times you find them insisting that "there's a limit to how much information you can get from the data" or something along those lines. "That Alien Message" is intended to convey the counterpoint, that smarter entities can extract more info than is immediately apparent on the surface of things.

Similarly, thinking about Solomonoff induction might also cause someone to realize that if, say, you simulated zillions of possible simple universes, you could look at which agents were seeing exact data like the data you got, and figure out where you were inside that range of possibilities, so long as there was literally *any* correlation to use.

And if you say that an agent *can't* extract that data, you're making a claim about which shortcuts to Solomonoff induction are and aren't computable. In fact, you're probably pointing at some *particular* shortcut and claiming nobody can ever figure that out using a reasonable amount of computing power *even though the info is there in principle.* Contemplating Solomonoff induction might help people realize that, yes, the data *is* there in principle. Like, until I ask you to imagine a civilization running for Graham's Number of years inside a Graham-sized memory space, you might not imagine them trying all the methods of analysis that *you personally* can imagine being possible.

ASHLEY: If somebody is making that mistake in the first place, I'm not sure you can beat it out of them by telling them the definition of Solomonoff induction.

BLAINE: Maybe not. But to brute-force somebody into imagining that sufficiently advanced agents have Level 1 protagonist intelligence, that they are epistemically efficient rather than missing factual questions that are visible even to us, you might need to ask them to imagine an agent that can see *literally anything seeable in the computational limit* just so that their mental simulation of the ideal answer isn't running up against stupidity assertions.

Like, I think there are a lot of people who could benefit from looking over the evidence they already personally have, and asking what a Solomonoff inductor could deduce from it, so that they wouldn't be running up against stupidity assertions *about themselves.* It's the same trick as asking yourself what God, Richard Feynman, or a "perfect rationalist" would believe in your shoes. You just have to pick a real or imaginary person that you respect enough for your model of that person to lack the same stupidity assertions that you believe about yourself.

ASHLEY: Well, let's once again try to factor out the part about Solomonoff induction in particular. If we're trying to imagine something epistemically smarter than ourselves, is there anything we get from imagining a complexity-weighted prior over programs in particular? That we don't get from, say, trying to imagine the reasoning of one particular Graham-Number-sized civilization?

BLAINE: We get the surety that even anything we imagine *Terence Tao himself* as being able to figure out, is something that is allowed to be known after some bounded number of errors versus Terence Tao, because Terence Tao is inside the list of all computer programs and gets promoted further each time the dominant paradigm makes a prediction error relative to him.

We can't get that dominance property without invoking "all possible ways of computing" or something like it—we can't incorporate the power of all reasonable processes, unless we have a set such that all the reasonable processes are in it. The enumeration of all possible computer programs is one such set.

ASHLEY: Hm.

BLAINE: Example three, diffraction is a simpler explanation of rainbows than divine intervention.

I don't think I need to belabor this point very much, even though in one way it might be the most central one. It sounds like "Jehovah placed rainbows in the sky as a sign that the Great Flood would never come again" is a 'simple' explanation; you can explain it to a child in nothing flat. Just the diagram of diffraction through a raindrop, to say nothing of the Principle of Least Action underlying diffraction, is something that humans don't usually learn until undergraduate physics, and it *sounds* more alien and less intuitive than Jehovah. In what sense is this intuitive sense of simplicity wrong? What gold standard are we comparing it to, that could be a better sense of simplicity than just 'how hard is it for me to understand'?

The answer is Solomonoff induction and the rule which says that simplicity is measured by the size of the computer program, not by how hard things are for human beings to understand. Diffraction is a small computer program; any programmer who understands diffraction can simulate it without too much trouble. Jehovah would be a much huger program—a complete mind that implements anger, vengeance, belief, memory, consequentialism, etcetera. Solomonoff induction is what tells us to retrain our intuitions so that differential equations feel like less burdensome explanations than heroic mythology.

ASHLEY: Now hold on just a second, if that's actually how Solomonoff induction works then it's not working very well. I mean, Abraham Lincoln was a great big complicated mechanism from an algorithmic standpoint—he had a hundred trillion synapses in his brain—but that doesn't mean I should look at the historical role supposedly filled by Abraham Lincoln, and look for simple mechanical rules that would account for the things Lincoln is said to have done. If you've already seen humans and you've already learned to model human minds, it shouldn't cost a vast amount to say there's one *more* human, like Lincoln, or one more entity that is *cognitively humanoid*, like the Old Testament jealous-god version of Jehovah. It may be *wrong* but it shouldn't be vastly improbable *a priori*.

If you've already been forced to acknowledge the existence of some humanlike minds, why not others? Shouldn't you get to reuse the complexity that you postulated to explain humans, in postulating Jehovah?

In fact, shouldn't that be what Solomonoff induction *does?* If you have a computer program that can model and predict humans, it should only be a slight modification of that program—only slightly longer in length and added code—to predict the modified-human entity that is Jehovah.

BLAINE: Hm. That's fair. I may have to retreat from that example somewhat.

In fact, that's yet another point to the credit of Solomonoff induction! The ability of programs to reuse code, incorporates our intuitive sense that if you've already postulated one kind of thing, it shouldn't cost as much to postulate a similar kind of thing elsewhere!

ASHLEY: Uh huh.

BLAINE: Well, but even if I was wrong that Solomonoff induction should make Jehovah seem very improbable, it's still Solomonoff induction that says that the alternative hypothesis of 'diffraction' shouldn't itself be seen as burdensome—even though diffraction might require a longer time to explain to a human, it's still at heart a simple program.

ASHLEY: Hmm.

I'm trying to think if there's some notion of 'simplicity' that I can abstract away from 'simple program' as the nice property that diffraction has as an explanation for rainbows, but I guess anything I try to say is going to come down to some way of counting the wheels and gears inside the explanation, and justify the complexity penalty on probability by the increased space of possible configurations each time we add a new gear. And I can't make it be about surface details because that will make whole humans seem way too improbable.

If I have to use simply specified systems and I can't use surface details or runtime, that's probably going to end up basically equivalent to Solomonoff induction. So in that case we might as well use Solomonoff induction, which is probably simpler than whatever I'll think up and will give us the same advice. Okay, you've mostly convinced me.

BLAINE: *Mostly?* What's left?

vii. Limitations

ASHLEY: Well, several things. Most of all, I think of how the 'language of thought' or 'language of epistemology' seems to be different in some sense from the 'language of computer programs'.

Like, when I think about the laws of Newtonian gravity, or when I think about my Mom, it's not just one more line of code tacked onto a big black-box computer program. It's more like I'm crafting an explanation with modular parts—if it contains a part that looks like Newtonian mechanics, I step back and reason that it might contain other parts with differential equations. If it has a line of code for a Mom, it might have a line of code for a Dad.

I'm worried that if I understood how humans think like that, maybe I'd look at Solomonoff induction and see how it doesn't incorporate some further key insight that's needed to do good epistemology.

BLAINE: Solomonoff induction literally incorporates a copy of you thinking about whatever you're thinking right now.

ASHLEY: Okay, great, but that's *inside* the system. If Solomonoff learns to promote computer programs containing good epistemology, but is not itself good epistemology, then it's not the best possible answer to "How do you compute epistemology?"

Like, natural selection produced humans but population genetics is not an answer to "How does intelligence work?" because the intelligence is in the inner content rather than the outer system. In that sense, it seems like a reasonable worry that Solomonoff induction might incorporate only *some* principles of good epistemology rather than *all* the principles, even if the *internal content* rather than the *outer system* might bootstrap the rest of the way.

BLAINE: Hm. If you put it *that* way...

(*long pause*)

... then I guess I have to agree. I mean, Solomonoff induction doesn't explicitly say anything about, say, the distinction between analytic propositions and empirical propositions, and knowing that is part of good epistemology on my view. So if you want to say that Solomonoff induction is something that bootstraps to good epistemology rather than being all of good epistemology by itself, I guess I have no choice but to agree.

I do think the outer system already contains a *lot* of good epistemology and inspires a lot of good advice all on its own. Especially if you give it credit for formally reproducing principles that are "common sense", because correctly formalizing common sense is no small feat.

ASHLEY: Got a list of the good advice you think is derivable?

BLAINE: Um. Not really, but off the top of my head:

1. The best explanation is the one with the best mixture of simplicity and matching the evidence.
2. "Simplicity" and "matching the evidence" can both be measured in bits, so they're commensurable.
3. The simplicity of a hypothesis is the number of bits required to formally specify it—for example, as a computer program.
4. When a hypothesis assigns twice as much probability to the exact observations seen so far as some other hypothesis, that's one bit's worth of relatively better matching the evidence.
5. You should actually be making your predictions using all the explanations, not just the single best one, but explanations that poorly match the evidence will drop down to tiny contributions very quickly.
6. Good explanations let you compress lots of data into compact reasons which strongly predict seeing just that data and no other data.
7. Logic can't dictate prior probabilities absolutely, but if you assign probability less than $2^{-1,000,000}$ to the prior that mechanisms constructed using a small number of objects from your universe might be able to well predict that universe, you're being unreasonable.
8. So long as you don't assign infinitesimal prior probability to hypotheses that let you do induction, they will very rapidly overtake hypotheses that don't.
9. It is a logical truth, not a contingent one, that more complex hypotheses must in the limit be less probable than simple ones.
10. Epistemic rationality is a precise art with no user-controlled degrees of freedom in how much probability you ideally ought to assign to a belief. If you think you can tweak the probability depending on what you want the answer to be, you're doing something wrong.
11. Things that you've seen in one place might reappear somewhere else.
12. Once you've learned a new language for your explanations, like differential equations, you can use it to describe other things, because your best hypotheses will now already encode that language.
13. We can learn meta-reasoning procedures as well as object-level facts by looking at which meta-reasoning rules are simple and have done well on the evidence so far.
14. So far, we seem to have no *a priori* reason to believe that universes which are more expensive to compute are less probable.
15. People were wrong about galaxies being *a priori* improbable because that's not how Occam's Razor works. Today, other people are equally wrong about other parts of a continuous wavefunction counting as extra entities for the purpose of evaluating hypotheses' complexity.
16. If something seems "weird" to you but would be a consequence of simple rules that fit the evidence so far, well, there's nothing in these explicit laws of epistemology that adds an extra penalty term for weirdness.
17. Your epistemology shouldn't have extra rules in it that aren't needed to do Solomonoff induction or something like it, including rules like "science is not allowed to examine this particular part of reality"—

ASHLEY: This list isn't finite, is it.

BLAINE: Well, there's a *lot* of outstanding debate about epistemology where you can view that debate through the lens of Solomonoff induction and see what Solomonoff suggests.

ASHLEY: But if you don't mind my stopping to look at your last item, #17 above—again, it's attempts to add *completeness* clauses to Solomonoff induction that make me the most nervous.

I guess you could say that a good rule of epistemology ought to be one that's promoted by Solomonoff induction—that it should arise, in some sense, from the simple ways of reasoning that are good at predicting observations. But that doesn't mean a good rule of epistemology ought to explicitly be in Solomonoff induction or it's out.

BLAINE: Can you think of good epistemology that doesn't seem to be contained in Solomonoff induction? Besides the example I already gave of distinguishing logical propositions from empirical ones.

ASHLEY: I've been trying to. First, it seems to me that when I reason about laws of physics and how those laws of physics might give rise to higher levels of organization like molecules, cells, human beings, the Earth, and so on, I'm not constructing in my mind a great big chunk of code that reproduces my observations. I feel like this difference might be important and it might have something to do with 'good epistemology'.

BLAINE: I guess it could be? I think if you're saying that there might be this unknown other thing and therefore Solomonoff induction is terrible, then that would be the nirvana fallacy. Solomonoff induction is the best formalized epistemology we have *right now*—

ASHLEY: I'm not saying that Solomonoff induction is terrible. I'm trying to look in the direction of things that might point to some future formalism that's better than Solomonoff induction. Here's another thing: I feel like I didn't have to learn how to model the human beings around me from scratch based on environmental observations. I got a jump-start on modeling other humans by observing *myself*, and by recruiting my brain areas to run in a sandbox mode that models other people's brain areas—empathy, in a word.

I guess I feel like Solomonoff induction doesn't incorporate that idea. Like, maybe *inside* the mixture there are programs which do that, but there's no explicit support in the outer formalism.

BLAINE: This doesn't feel to me like much of a disadvantage of Solomonoff induction—

ASHLEY: I'm not *saying* it would be a disadvantage if we actually had a hypercomputer to run Solomonoff induction. I'm saying it might point in the direction of "good epistemology" that isn't explicitly included in Solomonoff induction.

I mean, now that I think about it, a generalization of what I just said is that Solomonoff induction assumes I'm separated from the environment by a hard, Cartesian wall that occasionally hands me observations. Shouldn't a more realistic view of the universe be about a simple program that *contains me somewhere inside it,* rather than a simple program that hands observations to some other program?

BLAINE: Hm. Maybe. How would you formalize *that?* It seems to open up a big can of worms—

ASHLEY: But that's what my actual epistemology actually says. My world-model is not about a big computer program that provides inputs to my soul, it's about an enormous mathematically simple physical universe that instantiates Ashley as one piece of it. And I think it's good and important to have epistemology that works that way. It wasn't *obvious* that we needed to think about a simple universe that embeds us. Descartes *did* think in terms of an impervious soul that had the universe projecting sensory information onto its screen, and we had to get *away* from that kind of epistemology.

BLAINE: You understand that Solomonoff induction makes only a bounded number of errors relative to any computer program which does reason the way you prefer, right? If thinking of yourself as a contiguous piece of the universe lets you make better experimental predictions, programs which reason that way will rapidly be promoted.

ASHLEY: It's still unnerving to see a formalism that seems, in its own structure, to harken back to the Cartesian days of a separate soul watching a separate universe projecting sensory information on a screen. Who knows, maybe that would somehow come back to bite you?

BLAINE: Well, it wouldn't bite you in the form of repeatedly making wrong experimental predictions.

ASHLEY: But it might bite you in the form of having no way to represent the observation of, "I drank this 'wine' liquid and then my emotions changed; could my emotions themselves be instantiated in stuff that can interact with some component of this liquid? Can alcohol touch neurons and influence them, meaning that I'm not a separate soul?" If we interrogated the Solomonoff inductor, would it be able to understand that reasoning?

Which brings up that dangling question from before about modeling the effect that my actions and choices have on the environment, and whether, say, an agent that used Solomonoff induction would be able to correctly predict "If I drop an anvil on my head, my sequence of sensory observations will *end*."

ELIEZER: And that's my cue to step in!

The natural next place for this dialogue to go, if I ever write a continuation, is the question of actions and choices, and the agent that uses Solomonoff induction for beliefs and expected reward maximization for selecting actions—the perfect rolling sphere of advanced agent theory, AIXI.

Meanwhile: For more about the issues Ashley raised with agents being a contiguous part of the universe, see "Embedded Agency."