On expected utility

Joe Carlsmith

March 24, 2022

Part I

Skyscrapers and madmen

Summary. Suppose that you're trying to do something. Maybe: get a job, or pick a restaurant, or raise money to pay medical bills.

Some people think that unless you're messing up in silly ways, you should be acting "as if" you're *maximizing expected utility*—i.e., assigning consistent, real-numbered probabilities and utilities to the possible outcomes of your actions, and picking the action with the highest expected utility (the sum, across the action's possible outcomes, of each outcome's utility multiplied by its probability—example here).

But in combination with plausible ethical views, expected utility maximization (EUM) can lead to a focus on lower-probability, higher-stakes events—a focus that can be emotionally difficult. For example, faced with a chance to save someone's live for certain, it directs you to choose a 1% chance of saving 1000 lives instead—even though this choice will probably benefit no one. And EUM says to do this even for one shot, or few shot, choices—for example, choices about your career.

Why do this? The "quick argument"—namely, that EUM maximizes *actual* utility, given repeated choices and independent trials—isn't enough for few-shot cases. And neither are appeals to "collective action" across EUM-ish people with your values.

Rather, I think the strongest arguments for EUM come from a cluster of related theorems, which say something like: your choices conform to certain attractive ideals of rationality if and only if you act like an EUM-er. But the theorems themselves often go unexplained. Informal discussions typically list some set of axioms, and then state the theorem, but they leave the proof, and the basic dynamics underlying it, as a black box.

Early on in my exposure to such theorems, I found this frustrating. If I was going to make big decisions (especially one-shot decisions) using EUM as a guiding ideal, I wanted to understand its rationale more deeply. But the full proofs are often lengthy and difficult.

This series of essays is an attempt to write the type of thing I would've loved to read, at that point in my life.

- Part 1 (*Skyscrapers and madmen*) presents a way of visualizing EUM that I call the "skyscraper model." It also discusses (1) the failure of the "quick argument," (2) reasons not to let your epistemic relationship to EUM stop at "apparently the math says blah," and (3) whether being "representable" as an EUM-er is trivial (in the relevant sense, I don't think so).
- Part 2 (*Why it can be OK to predictably lose*) focuses on why it makes sense, in cases like "save one life for certain, or 1000 with 1% probability," to choose the risky option, and hence to "predictably lose." The answer is unsurprising: sometimes the upside is worth it. I offer three arguments to make this clearer and more vivid with respect to life-

saving in particular. Ultimately, though, while EUM imposes consistency constraints on our choices, it does not provide guidance about what's "worth it" or not. You have to decide for yourself.

- Part 3 (*VNM, separability, and more*) examines three theorems that take probability assignments for granted, and derive a conclusion of the form: "Your choices satisfy XYZ conditions iff you act like an EUM-er": the von Neumann-Morgenstern theorem; a proof based on the very general connection between "separability" and "additivity"; and a related "direct" axiomatization of EUM in Peterson (2017). In each case, I aim to go beyond listing axioms, and to give some intuitive (if still basic and informal) flavor for how the reasoning underlying the proof works.
- Part 4 (*Dutch books, Cox, and Complete Class*) tries this same project on theorems that try to justify subjective probability assignments: Dutch Book theorems; Cox's Theorem (this one is still a bit of a black box to me); and the Complete Class Theorem (this one also supports EUM more broadly). I also briefly discuss Savage, Jeffrey-Bolker, and a certain very general argument for making consistent trade-offs on the margin—both across goods, and across worlds.

Exactly how much support these theorems give to EUM as a normative ideal is a further question, which I don't try to tackle comprehensively (though collectively, I find them pretty compelling). And there are lots of further issues in this vicinity I don't discuss. Regardless of whether you embrace EUM, though, I think engaging with these theorems at a level deeper than "apparently the math says blah" can result in a more visceral and clear-headed relationship with the moves and vibes that structure EUM-ish thinking. I've found such engagement useful, and I hope some readers will too.

Thanks to Katja Grace, Cate Hall, Petra Kosonen, Ketan Ramakrishnan, and especially to John Wentworth for discussion.

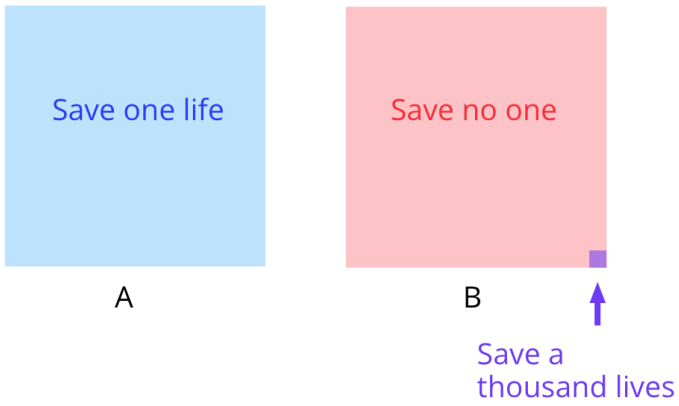
1 Maximal skyscraper

What does an expected utility maximizer (EUM-er) do? Three things:

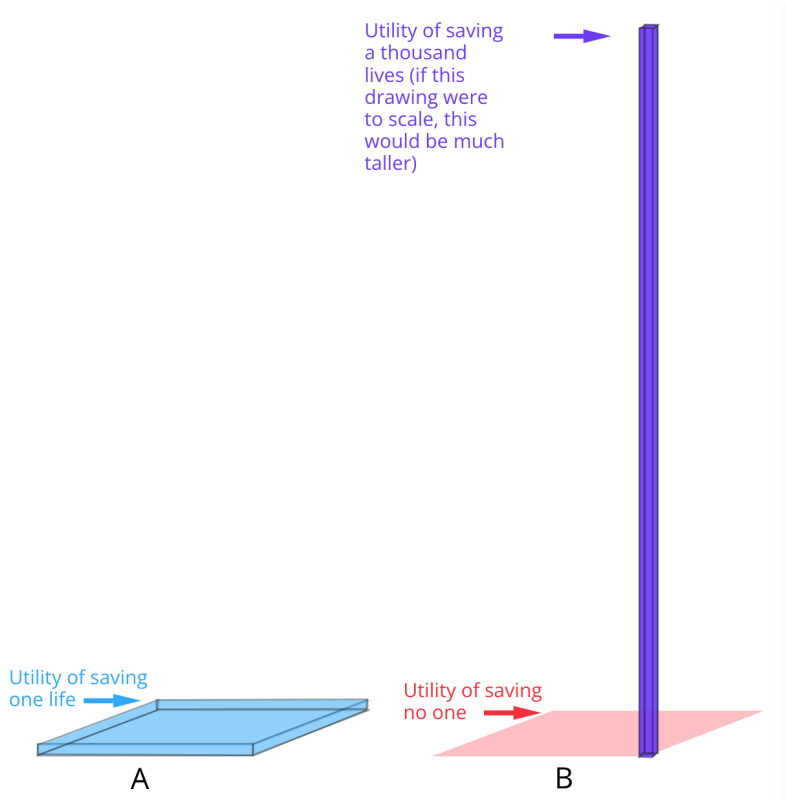
1. Assigns probabilities to the possible outcomes of their actions (call this "probabilism").
2. Assigns real-numbered "utilities" to these outcomes, representing something like their value/preferred-ness (call this "utility-ism").
3. Chooses the action that yields the highest expected utility—i.e., the sum, across the action's possible outcomes, of each outcome's utility multiplied by its probability.

Thus, suppose you can save A one life for certain, or (B) a thousand lives with 1% chance. And suppose you value each life equally, such that saving a thousand lives is a thousand times better than saving one. EUM then says to choose B, because it saves ten lives in expectation, whereas A saves only one.

We can think of EUM via what I call a "skyscraper" model. Probabilities, recall, behave like the area of a space, like a 1×1 square (see here and here for more). Thus, in the choice above:



Utilities then behave like an extra dimension.



So each action gets a "city-scape"—that is, a probability square as the ground, and "skyscrapers" sprouting up from the different regions, with bases corresponding to the probability of the outcome in that region, and heights corresponding to that outcome's

utility. And EUM says to choose the action with the largest total volume of skyscraper. It's a pro-housing vibe.

Thus, A is a very short skyscraper taking up the whole city. B is mostly an empty lot, but it's got a very tall and thin skyscraper sitting in the corner. And this skyscraper is sufficiently tall that B actually has more total housing overall (the drawing above does not capture this well). So EUM chooses B.

(At times in this series, I'll also use 2d visualizations, with probability on the horizontal axis, and utility on the vertical axis, where the expected utility is then the area of the 2d "housing." This is often simpler and better—but I like the way the 3d picture can combine with visualizing probability as 2d, which I find generally useful in other contexts.)

2 Only one shot

OK, but: what's supposed to be cool about this? In particular: it seems like B, here, probably does nothing. If you choose B over A, you predictably lose. At least a 1000 people die, and you could've saved one of them for certain (throughout this series, I'll use the term "losing" broadly, to mean "ending up in a dis-preferred state"). And if EUM results in predictably losing, why do it?

One argument is: EUM does well in the long run. In particular: it maximizes *actual* utility, if you make the same choice enough times. For example, if you choose B over A a zillion times (and the 1%, in B, is e.g. a new dice-roll each time), then you're ~guaranteed to save ~10x the lives, relative to choosing A over B instead.

This is the "quick argument." But it's also the "not good enough" argument. For one thing: sometimes failures are correlated. If the 1% comes from saving 1000 if the nth through $n+6$th digits of pi are all odd (really, this would be 0.8%), and they're not, but you never get to find out whether you're saving people or not, you can choose B over and over, and never save anyone.

But also: suppose that you're only choosing between A and B once. Suppose, indeed, that choosing between A and B is literally the only thing you will ever do. What then? EUM fans still say: B. But why? The quick argument is silent.

What's more, this silence matters in real life. Consider careers. I'm in favor of applying EUM-ish reasoning to career choice. For example, I think it can be worth spending your whole career trying to prevent some form of existential catastrophe, even if the overall risk of that type of catastrophe is low, and all your work will probably end up irrelevant. The stakes are that large.

But careers are especially bad candidates for "if you repeat it enough times, you'll eventually come out ahead." You can only change careers so many times, and feedback about whether you're in a "everything I'm doing is irrelevant" world doesn't always come readily. And beyond this, you've only got one life—one single chance to do something in this world. You've got to make it count. But if you choose via EUM, then sometimes, it

probably won't.

Perhaps you say: "yes, but if everyone with your values does EUM, then you get to repeat the choice across people, rather than across time." But this isn't enough, either.

- For one thing, EUM fans generally want to say that you should choose B over A, even if this is the only time *anyone* will make a choice like this.
- For another, sometimes no one shares your values. Bob, for example, might aim to eat maximum corn fritters. But he stands little chance of justifying his "expected fritters-eaten-by-Bob" maximization via reference to some kind of community effort. (Here I'm setting aside EU-maximization across quantum branches, which won't work for logical uncertainty anyway—e.g., getting one fritter with certainty, or 1000 if blah digits of pi are all odd).
- For a third, "community effort" arguments are famously slippery. (What if you know the other people won't act like you do? See also: rule consequentialism.)
- For a fourth, sometimes failures are correlated across people, too. For example, if there's a 5% chance the asteroid is headed towards earth, and all the EUM-ers join the asteroid prevention effort (let's say the numbers work out in favor of this), then the whole community effort still has a 95% chance of irrelevance.

Pretty clearly, some other argument is needed.

3 Against "apparently the scary math says so?"

The type of argument I like best appeals to a cluster of related theorems pointing at EUM's equivalence to satisfying various attractive constraints on rationality (often formulated as "axioms"). But I don't like the way this argument is often left as a black box of scary math.

In particular, when I was first learning about EUM, it felt like people would often mention the theorems, without explaining how or why they work. Sometimes, they would get as far as listing and debating some set of axioms: but then they'd just move from discussing the axioms to the stating theorem, while skipping the proof. Indeed, it remains unclear to me how many fans of EUM have ever actually engaged with the proofs in question.

Perhaps you say: "But, if the proofs are valid, do you need to understand them?" And indeed, not necessarily. But I wanted to. In particular: it felt like EUM was asking a lot of me. It was asking me, for example, to predictably lose—to predictably let people die, to predictably waste my time, for the sake of...something. "Rationality?" Rationality is not an end in itself. What, then? If I was going to make few-shot, high-stakes, predictably-losing life choices on EUM-ish grounds, I wanted to get it "in my gut." And I hoped that being able to see the flow of logic, from premises to conclusion, would help.

What's more, absent deeper understanding, I felt some worry about internal coercion. If I left the theorems as black boxes, it felt easy to round them off to: "apparently the scary

math I'll never understand says that I have do EUM, otherwise I'm silly and bad." And this didn't feel like a healthy or sustainable set up—especially if EUM was going to ask me to do lots of emotionally unrewarding things. Indeed, I wonder how many people currently relate to EUM this way—and about the costs of doing so.

Plus, there was this stuff about fanaticism (i.e., cases where EUM leads to obsession with *tiny* probabilities of very high stakes outcomes). It felt like people liked EUM until they didn't. Some probabilities were "pascalian," and some weren't. 1%, I guess, wasn't (I do think this is right). Why not? How do you tell the difference? Hand-wave, hand-wave, we don't know. (I, at least, still don't know. My current best guess is something about bounded utility functions, which you maybe want anyway, but I haven't worked it out, and I don't expect to like it.) Bit of a red flag? Maybe not a time to just sit pretty with the definitely-fine math (see also: infinite ethics). And maybe understanding it better could shed light (indeed: the proof of the VNM theorem I'll present explicitly assumes bounded utility functions, which don't lead to fanaticism—but it's easy to not know that "blah proof assumes bounded utility functions," if you've never actually looked at the proof in question).

(Also: some people think these theorems are relevant, or aren't, to whether we should expect advanced AI systems to kill us all by default—see e.g. Omohundro (2008), Yudkowsky (1, 2), Shah (2018), Ngo (2019), Grace (2021). I'm not going to delve into this topic, but I think it's another reason to actually understand how the theorems work—and it's part of what prompts my own interest.)

I'm hoping, here, to write the type of thing I would've loved to read, when I was first looking into EUM. I want to acknowledge, though, that the relevant audience might be correspondingly limited. In particular: there's going to be a bit of (basic) math, but I'm also going to work through it slowly, informally, and with various simplifying assumptions. So I worry I'll lose the math-phobic as soon as we hit the symbols, and bore/frustrate the math-fluent. So it goes. Hopefully, there are a few readers in my specific demographic, for whom it scratches the right itch.

I also want to acknowledge that I'm not speaking from a place of "I've gone through and really understood the original versions of all these proofs, including for infinite cases." Often, indeed, the original version has been too long/intense for me. Rather, I've tried to find, understand, and present *some* relatively short and comprehensible explanation of the basic thing going on in at least some finite version of the proof. For me, at this point, it's enough. Those who seek more depth, though, can consult the original sources, which I'll generally link to (and for those who want a more comprehensive but still accessible introduction to EUM, I recommend Peterson (2017)—though he doesn't include various of the theorems I'll discuss).

4 Is being an EUM-er trivial?

I want to head off one other objection before we dive in.

No one (sensible) thinks you should go around doing comprehensive, explicit EU calculations for every decision. Nor, indeed, do philosophical fans of EUM necessarily claim that *ideally rational agents* do this. Rather, they generally claim that ideally rational agents *act like* an EUM-ers. That is, if you're an ideally rational agent with respect to some set of alternatives A, it's possible to construct a probability assignment p and a utility function u, such that you make the same choices about A that an EUM-er with u and p would make.

But now perhaps we wonder: is constructing such a p and u trivial?

- Consider a rock. Is it maximizing a utility function that prioritizes "just sitting there"?
- Or consider a twitching mad-man (see e.g. Shah (2018)). Is he maximizing the utility function that gives "twitching in exactly this way, in exactly this situation" 1, and all else zero?
- Or consider that time you paid a dollar to trade an apple for an orange at t_1, another dollar to trade an orange for a pear at t_2, and another dollar to trade a pear for an apple at t_3 ("money-pumping"). Were you, maybe, maximizing a utility function on which t_1 apples $<$ t_1 oranges, t_2 oranges $<$ t_2 pears, and t_3 pears $<$ t_3 apples? Nothing intransitive about that. (See e.g. Dreier (1996) for useful discussion.)
- Or consider that time you chose B over A, above, and then patted yourself on the back for your solid EUM-ing. Were you, maybe, *failing* to maximize a utility function that values saving one person *more* than it values saving a thousand people?

The worry, then, is that representation is too cheap. I can represent as anything as an EUM-er, or as violating EUM, if I try.

I do think there are tricky issues in this vicinity. But for a given physical system S, and a given set of alternatives A, we should take care to distinguish between:

(1) Does it make sense to think of S as having preferences over A, and if so, how do we tell what they are?

(2) For any given set of preferences over A, is it trivial to represent this set of preferences as maximizing expected utility?

(1) is hard, and I won't tackle it here. But (2) is easy: the answer is no. For example, if your preferences over A are intransitive, they can't be represented as maximizing expected utility: maybe you prefer puppies to flowers to mud to puppies, but there is no function u to the real numbers such that $u(\text{puppies}) > u(\text{flowers}) > u(\text{mud}) > u(\text{puppies})$. Indeed, this is the type of thing we learn from the theorems I'll discuss.

The examples above get their force from not knowing how to answer (1). Presented with a given episode of real-world behavior in a physical system S, they observe that this episode is compatible, in principle, with some set of preferences over some set of alternatives, and some utility function representing those preferences, such that the episode is EU-maximizing (or not). And because they don't know how to answer (1), they assume these preferences are viable candidates for S's preferences.

Plausibly, though, (1) has an answer. A rock, for example, doesn't have preferences. A twitching madman doesn't, actually, maximally prefer a world where he twitches in exactly X way (intuition pump: if God took him to a realm "beyond the world," and offered him a chance to create a "I twitch in exactly X way" world by pressing button A, or a "I am sane and healthy" world by pressing button B, he would not, consistently, press button A. Rather, he would twitch until he hit a random button). And presumably, the answer to (1) would allow us to evaluate whether your feelings about fruit are sensitive to their temporal location in some consistent way; or whether you do, in fact, prefer one life saved to a thousand.

But the theorems I'll discuss aren't about (1). They *assume* that you are a preference-haver, and that you are trying to define preferences over some set of alternatives. And they tell you that if and only if you do this in XYZ attractive ways, *then* you're representable as an EUM-er. And doing it in XYZ ways is not trivial at all. You can't just randomize, or flail around. Rocks will have trouble. Madmen, too.

What's more, in the real world, excuses that appeal to (1)-ish ambiguities seem like cold comfort—at least when I try to apply them to myself. Suppose I choose A over B above. I can say, if I want, that actually, I am an EUM-er; I just value one life saved more than a thousand. But do I? Or suppose I get money-pumped. I can say, if I want, that I actually just like it when fruit changes hands. But do I?

Maybe, if I say these things, I don't have to label myself "irrational," or my choice a "mistake." Maybe I've blocked some sort of abstract insult. But if my values aren't actually like this, am I coming out ahead? This was supposed to be about lives, or fruit. Who cares about abstract insults? That's not the bullying to worry about. And anyway, if it's not, actually, *me* who is EUM-ish, but rather some forced re-interpretation, am I, perhaps, still insulted?

Granted, figuring out your "true values" is tough and ambiguous stuff (see here for trickiness); and figuring out the "true values" of some other physical system, even more so. But even absent some nice theory, we shouldn't mistake this trickiness for "anything goes," or confuse it with the (false) claim that all preference relations over alternatives are representable as EU-maximizing.

In the next post, I start diving into substantive arguments in favor of EUM, starting with choices like A vs. B above.

Part II

Why it can be OK to predictably lose

Previously in sequence: Skyscrapers and madmen

This is the second essay in a four-part series on expected utility maximization (EUM). This part focuses on why it can make sense, in cases like "save one life for certain, or 1000 with 1% chance," to choose the risky option, and hence to "predictably lose." The answer is unsurprising: sometimes the upside is worth it. I offer three arguments to clarify this with respect to life-saving in particular. Ultimately, though, while EUM imposes consistency constraints on our choices, it does not provide guidance about what's "worth it" or not—and in some cases, the world doesn't either. Ultimately, you have to decide.

5 Sometimes it's worth it

When I first started looking into expected utility maximization (EUM), the question of "why, exactly, should I be OK with predictably losing in one-shot cases?" was especially important to me. In particular, it felt to me like the fact that you'll "predictably lose" was the key problem with fanaticism (though some fanaticism cases, like Pascal's Mugging, introduce further elements, like an adversarial dynamic, and/or a very "made-up" tiny probability as opposed to e.g. a draw from a ludicrously large urn). But it also applies (to a far milder degree) to choices that EUM fans endorse, like passing up (A) saving one life for certain, in favor of (B) a 1% chance of saving 1000 lives.

Now, I'm not going to tackle the topic of fanaticism here. But I think the EUM fans are right about B over A—indeed, importantly so. What's more, I think it's useful to get a more direct, visceral flavor for *why* they're right—one that assists in navigating the emotional challenge of passing up "safer," more moderate prizes for the sake of riskier, larger ones. This essay draws on moves from the theorems I'll discuss in parts 3 and 4 to try to bring out such a flavor with respect to life-saving in particular.

I'll flag up front, though, that here (and throughout this series) I'm going to be focusing on highly stylized and oversimplified cases. Attempts to apply EUM-ish reasoning in the real world are rarely so straightforward (see e.g. here for some discussion), and I think that hesitations about such attempts sometimes track important considerations that naive estimates leave out (for example, considerations related to how much one expects the relevant probability estimates to change as one learns more). My aim here is to bring out the sense in which B is better than A in theory. Particular instances of practice, as ever, are further questions (though I think that EUM is generally under-used in practice, too).

Also: even if predictably losing is OK, that doesn't make losing, itself, any less of a loss. Even if you're aiming for lower-probability, higher upside wins, this doesn't make it reasonable to give up, in the back of your mind, on actually winning, or to stop trying

to drive the probability of losing down. EUM is not an excuse, after you lose, to say: "well, it was always a long shot, and we never really expected it to pay off." Having acted "reasonably" is little comfort: it's the actual outcome that matters. Indeed, that's why the predictable loss bites so hard.

In that case, though: why is predictably losing OK? My answer is flat-footed and unsurprising: sometimes, what happens if you win is worth it. B is better than A, because saving a thousand lives is *that important*. A career spent working to prevent a 1% existential risk can be worth it, because the future is *that precious*, and the cost of catastrophe *that high*.

Ultimately, that's the main thing. But sometimes, you can bring it into clearer focus. Here I'll offer a few tools for doing so.

6 Small probabilities are just bigger conditional probabilities in disguise

One approach is to remember that there's nothing special about small probabilities—they're just bigger conditional probabilities in disguise. Consistency about how you compare prizes X and Y, vs. how you compare X and Y *if* e.g. a coin comes up heads, or if many coins *all* come up heads, thus leads to tolerance of small-probability wins (this is a principle that some of the theorems I discuss draw on).

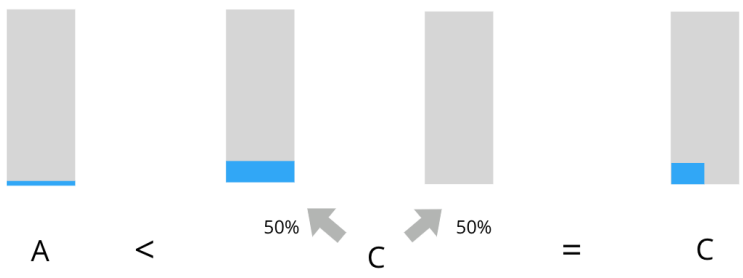
To see this, consider:

A: Certainty of saving 1 life.

C: Heads you save 5, tails you save 0.

Do you prefer C over A? I do—and I don't have some "but I'll predictably lose" reaction to C.

Here's the comparison in skyscraper terms (for simplicity, I'll just do a two-dimensional version):

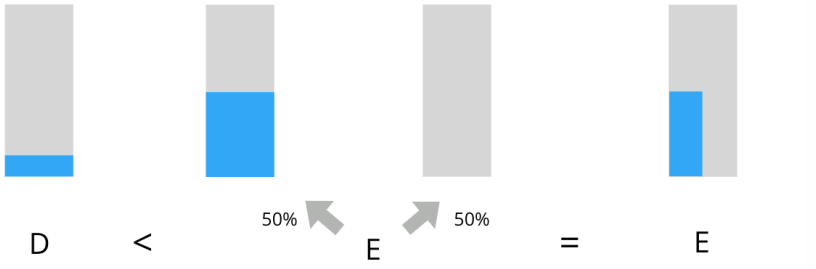


But now consider:

D: Certainty of saving 5.

E: Heads you save 15, tails you save 0.

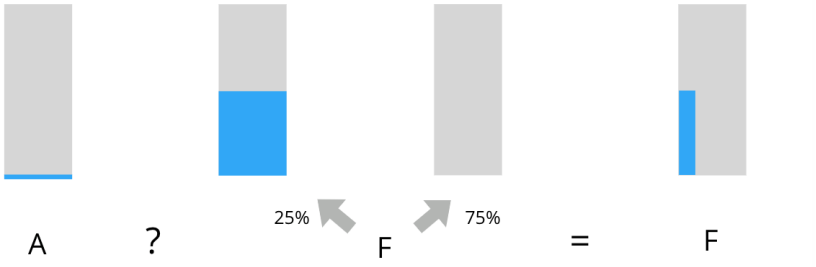
Do you prefer E over D? I do. And again: nice, hefty probabilities either way.



So what about:

A: Certainty of saving 1.

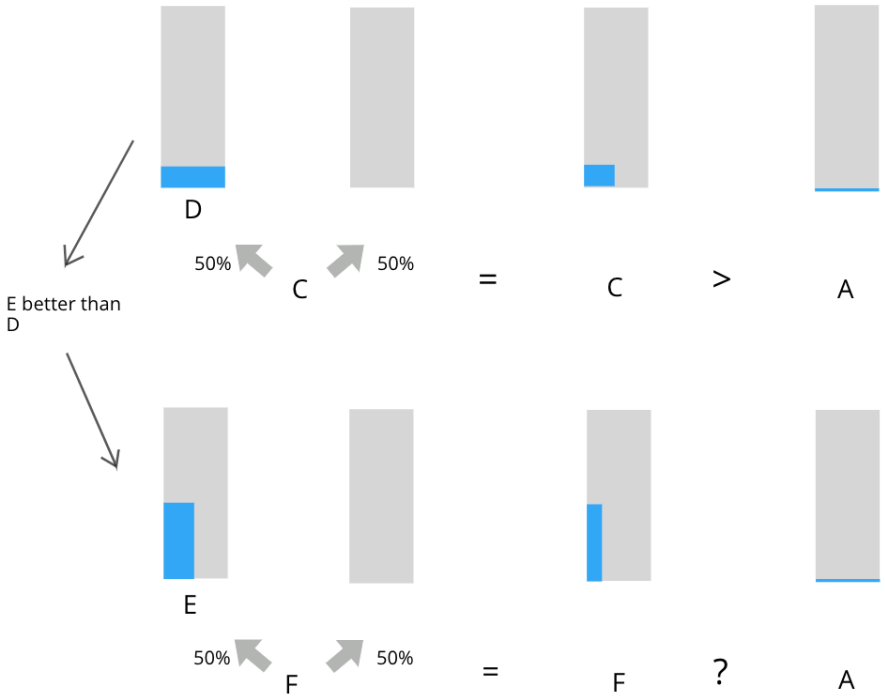
F: Two coin flips. Double heads you save 15, anything else you save 0.



Suppose that here you say: "hmm...a 75% chance of losing is starting to feel a bit too predictable." But here's an argument that if you prefer C to A, and E to D, you should prefer F to A. Suppose you start with a ticket for A. I offer you a chance to pay a dollar to switch to C, and you say yes. We flip, the coin homes up heads, and you're about to cash in your C ticket for five lives saved. But first, I offer you a chance to trade your C ticket, plus a dollar, to play another coin-flip, with pay-outs like E: if the second coin is heads, you'll save fifteen, and no one otherwise. Since, given the first heads, your C-ticket has

been converted into a "five lives with certainty" ticket, this is a choice between D and E, so you go for it. But now, in sequence, you've actually just chosen F.

That is: you like C better than A (in virtue of C's win condition), you like E better than D, and F *just* is a version of C with E as the "win condition" instead of D. Looking at the skyscrapers makes this clear:



To bring out the need for consistency here, suppose that when I first offer you C, I also tell you ahead of time that I'm going to offer you E if the first coin comes up heads—and I give you a chance to choose, prior to the coin flip, what answer to give later. It seems strange if, before flipping, you *don't* want to switch, given heads; but then, once it actually comes up heads, you do. And if that's your pattern of preference, I can start you off with an "E-given-first-coin-heads" ticket, which is actually just an F ticket. Then you'll pay to switch to a C ticket, then we'll flip, and if it comes up heads, you'll pay to switch back—thereby wasting two dollars with 50% probability.

We can run longer versions of an argument like this, to get to a preference for B over A. Suppose, for simplicity, that you're always indifferent between saving x lives, and a coin flip between saving $2x$ vs. no one (the argument also works if you require $> 2x$, as long as you don't demand too many additional lives saved at each step). Then we can string seven coin-flips together, to get indifference between A, and a lottery A' with a ~0.8% (0.5^7) chance of saving 128 (2^7) lives, and no one otherwise. But .8% is *less* than 1%, and 128 is *less* than 1000. So if you prefer saving *more lives* with *higher probability*, with no downside, B is better than A'. So, it's better than A.

(In "skyscraper" terms, what we're doing here is bunching the same volume of housing into progressively taller and thinner buildings, until we have a building, equivalent to A's flat and short situation, that is both shorter *and* thinner than the building in B.)

I think that moves in this broad vein can be helpful for breaking down "but I'll predictably lose"-type reactions into conditional strings of less risky gambles. But really (at least in the eyes of EUM), they're just reminding of you the fact that you can value one outcome *a lot* more than you value another. That is: if, in the face of a predictable loss, it's hard to remember that e.g. you value saving a thousand lives *a thousand times more* than saving one, then you can remember, via coin-flips, that you value saving two *twice as much* as saving one, saving four *twice as much* as saving two, and so on.

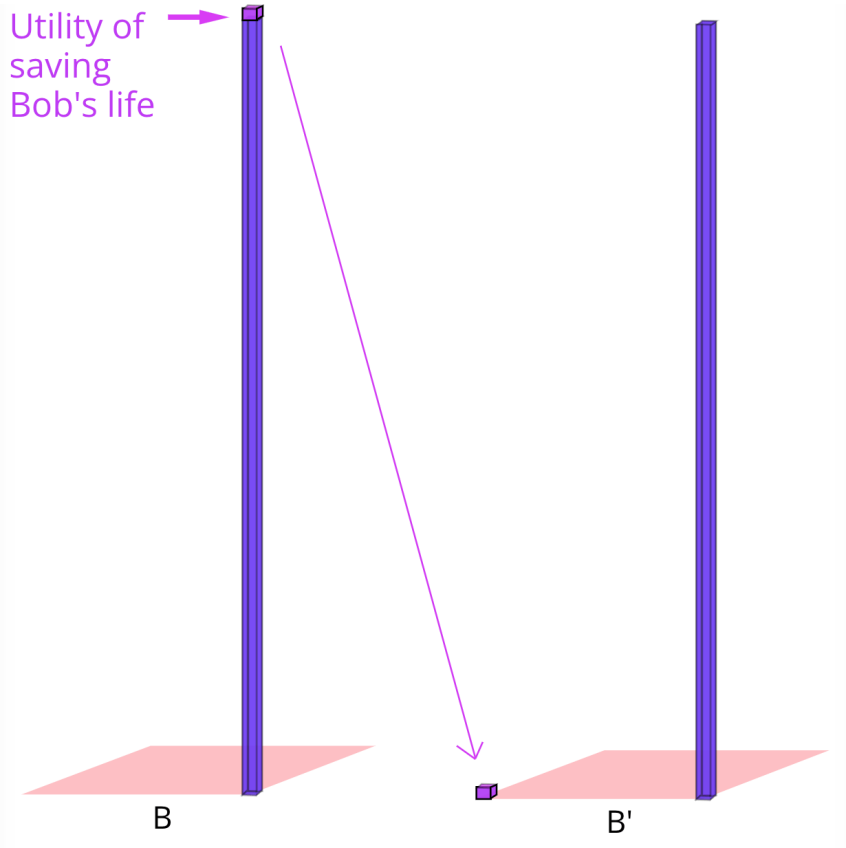
Of course, coupled with an unbounded utility function (such that, e.g., for *every* outcome O, there's always *some* outcome O', such that you'll take O' if heads, and nothing otherwise, over O with certainty) then this form of argument also leads straight to fanaticism (e.g., for any outcome O, no matter how good, and any probability p, no matter how small, there's some outcome O'', such that you'll take O'' with p, and nothing otherwise, over O with certainty). This is one of the reasons I'm interested in bounded utility functions. But more broadly: if you've got an unbounded utility function, then worries about fanaticism will pop up with *every* good argument for EUM, because fanaticism follows, straightforwardly, from EUM + an unbounded utility function. This is, indeed, a red flag. But we should still examine the arguments for EUM on the merits.

7 Nothing special about getting saved on heads vs. tails

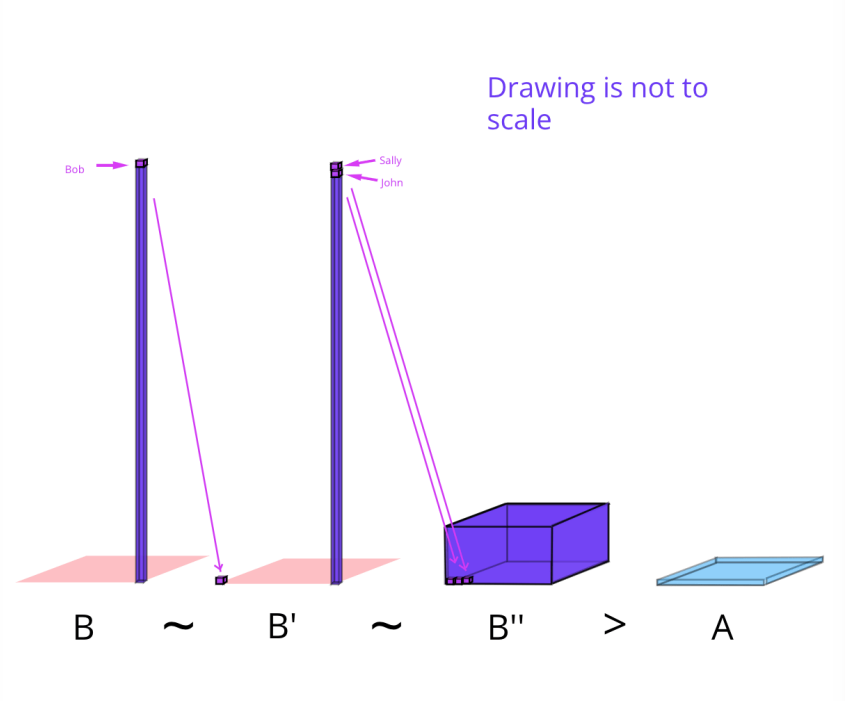
Here's a different argument for B over A—again, borrowing moves from the theorems I'll discuss.

Suppose that Bob is dying, but you'll save him if a fair coin lands heads. Would it be better, maybe, to save him if it lands tails, instead? No. Would it be worse? No. It doesn't matter whether you save Bob if heads, or if tails: they're equally probable.

Now let's generalize a bit. Imagine that B saves 1000 lives if you draw "1" out of an urn with 100 balls, labeled 1-100. Suppose Bob's is one of those lives. Is it any better, or any worse, to save Bob if you draw ball 1, vs. ball 2? No: they're equally probable. Ok, so now we've got indifference between B and B', where B' saves 999 on ball 1, 1 person (Bob) on ball 2, and no one otherwise.



But now repeat with Sally—another ball 1 life. Is it better, or worse, to save Sally on ball 1 vs. ball 3? No. So we move saving Sally, in B', to ball 3. Then we move saving John to ball 4, and so on, until we've moved ten people to each ball. Call this "ten people for each ball" lottery B". So: we've got indifference between B" and B. But B" is just a chance to save ten people with certainty, which sounds a lot better than saving one person with certainty—i.e., A. So, B" is better than A. Thus, B is better than A.



In skyscraper terms, what we're doing here is slicing chunks of housing off of the top of B's skyscraper, and moving them down to the ground. Eventually, we get a city-scape that is perfectly flat, and everywhere taller than A.

8 What would everyone prefer?

Here's a final argument for B over A. It's less general across EUM-ish contexts, I find it quite compelling in the context of altruistic projects in particular.

Suppose the situation is: 1000 people are drowning. A is a certainty of saving one of them, chosen at random. B is a 1% chance of saving all of them.

Thus, for each person, A gives them a .1% chance of living; whereas B gives them a 1% chance. So every single person wants you to choose B. Thus: if you're not choosing B, what *are* you doing, and why are you calling it "helping people"? Are you, maybe, trying to "be someone who saved someone's life," at the cost of making everyone 10x less likely to live? F*** that.

Now, some philosophers think that it makes a difference whether you know who A will save, vs. if you don't (see the literature on "Identified vs. Statistical Lives"). To me, this *really* looks like it shouldn't matter (for example, it means that everyone should pay large amounts to prevent you from learning who A will save—and plausibly that you should pay this money, too, so that it remains permissible to do the thing everyone wants). But I won't debate this here; and regardless, it's not an objection about "predictably losing."

9 Taking responsibility

OK, so we've seen three arguments for B over A, two of which drew on moves that we'll generalize, in the next essay, into full theorems.

I want to note, though, that using "x lives saved" as the relevant type of outcome, and coin flips or urn-drawings as the "states" that give rise to these outcomes, is easy mode, for EUM. In particular: the value at stake in an outcome can be broken down fairly easily into discrete parts—namely, lives—the value of which is plausibly independent of what happens with the others (see the discussion of "separability" below), and which hence function as a natural and re-arrangeable "unit" of utility. And everyone wants to be a probabilist about coins and urns.

In many other cases, though, such luxuries aren't available. Suppose, for example, that you have grown a beautiful garden. This afternoon, you were planning to take a walk—something you dearly enjoy. But you've heard at the village pub (not the most reliable source of information) that a family of deer is passing through the area and trampling on/eating people's gardens. If they stop by while you're out, they'll do this to your garden, too. If you stay to guard the garden, though, you can shoo them away (let's say, for simplicity, that the interests of the deer aren't affected either way).

How many walks is your garden worth? What are the chances that the deer stop by? What, exactly, are the city-scapes here?

EUM won't tell you these things. You need to decide for yourself. That is: ultimately, you have to give *weights* to the garden, the walk, and the worlds where the deer stop by vs. don't. Some things matter, to you, more than others. Some states of the world are more plausible than others, even if you're not certain either way. And sometimes, a given action (i.e. taking a walk) affects what matters to you differently, depending on the state of the world (i.e., the deer stop by, or they don't). Somehow, you have to weigh this all up, and output a decision.

EUM just says to respond to this predicament in a way that satisfies certain constraints. And these constraints impose useful discipline, especially when coupled with certain sorts of intuition pumps. To get more quantitative about the plausibility of deer, for example, we can ask questions like: "if I no longer had any stake in the deer situation, would I rather win $1M if they stop by, or if a red ball gets pulled from an urn with p% red balls in it?"—and call the p where you're indifferent your probability on deer (see Critch (2017), and more in part 4). To get a better grip on the value of walks vs. gardens, we can ask questions like "how many walks would you give up to prevent the certain destruction of your garden?", or we try to construct other intermediate trades in terms of a common "currency" (e.g., how much time or money will you pay to take a walk, vs. to restore your garden after it gets destroyed). And so on.

Still, most of the work—including the work involved in generating these sorts of intuitions—is on you. If your answers are inconsistent, EUM tells you to revise them, but it doesn't tell you how. And there's no secret sauce—and in many cases, no "answer" available in

the world (no "true probability," or "true utility," of blah; and no easy unit, like lives saved or balls-in-the-urn, that you can tally up). You just have to choose what sort of force you want to be in the world, under what circumstances, and you have to face the inevitable trade-offs that this choice implies.

Sometimes, people complain about this. They complain, for example, that there's "no way to estimate" the probability of deer, or to quantify the ratio of the value differences between different combinations of gardens and walks. Or they complain that various of the theorems I'll discuss aren't "action-guiding," because such theorems construct your utility function and probability assignment out of your choices (assuming your choices satisfy EUM-ish constraints), rather than telling you what probabilities and utilities to assign, which, like a good EUM-er, you can then go and multiply/add.

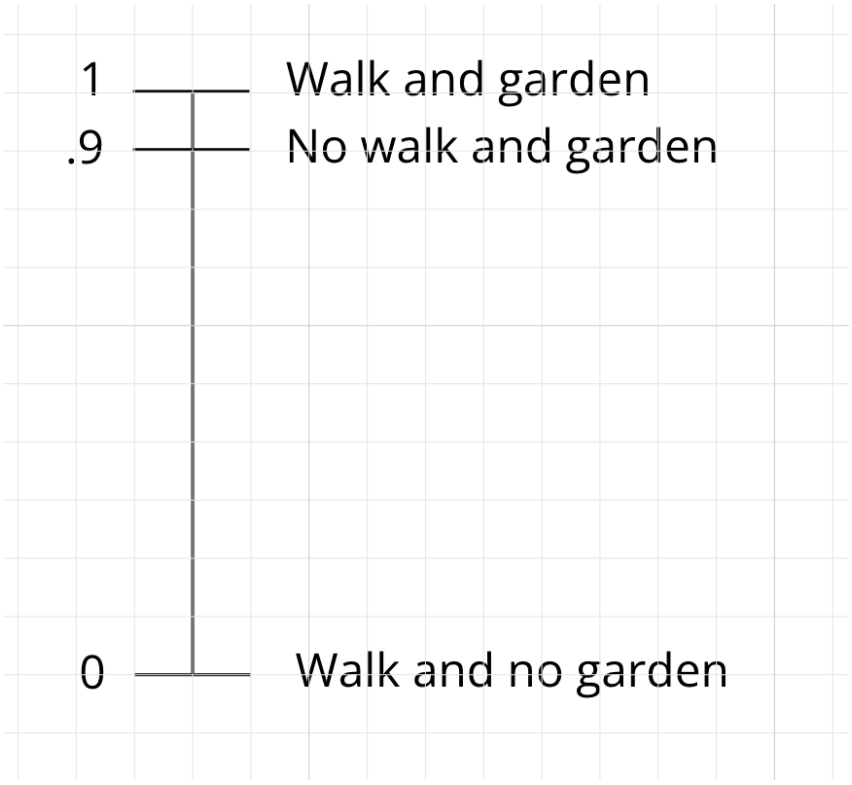
Perhaps, sometimes, it makes sense to seek such guidance. The world does in fact make it possible to be better or worse at assigning probabilities to claims like "the deer will stop by" (as the existence of superforecasters makes clear); the meta-ethical realists think that the world provides them with a "true utility function"; and anti-realists tend to look for various forms of guidance as well—for example, from their idealized selves.

But the case for acting consistently with EUM does not rest on the availability of such guidance—rather, it rests on the plausibility of certain very general, structural constraints on rational behavior (e.g., you shouldn't prefer A to B to C to A). And if you accept this case, then talking about whether you "can" assign utilities and probabilities to X or Y misses the point. To the extent you want your choices to be consistent with EUM, you *are*, always, assigning probabilities and utilities (or at least, some overall probability $*$ utility weight) to things. It's not optional. It's not a thing you can pick up when easy and convenient, and put down when frustrating and amorphous. You're doing it already, inevitably, at every moment. The question is how much responsibility you are taking for doing so.

To see this, let's suppose that you want your choices to be consistent with EUM, and let's look at the deer/garden situation in more detail. The situation here is:

	Deer stop by	**Deer don't stop by**
Walk	Walk and no garden	Walk and garden
Stay to guard	No walk and garden	No walk and garden

The inevitability of assigning both probabilities and utilities is especially clear if you assign *one* of them, but refuse to the assign the other—so let's start with cases like that. Suppose, for example, that you've decided that saving the garden from otherwise-certain destruction would be worth giving up your next ten walks; that each of these walks adds an equal amount of value to your life; and that the value of walks does not change depending on what's going on with the garden (and vice versa). Thus, you set a utility scale where the gap between "walk and no garden" and "no walk and garden" is 9x the size of the gap between "no walk and garden" and "walk and garden."



Let's use a utility function with 1 at the top and 0 at the bottom, such that

$$u(\text{walk and garden}) = 1,$$
$$u(\text{walk and no garden}) = 0,$$
$$u(\text{no walk and garden}) = 0.9.$$

(We can also scale this function by any positive constant c, and add any constant d—i.e. we can perform any "positive affine transformation"—while leaving EUM's overall verdicts the same. What matters is the ratios of the size of the gaps between the different outcomes, which stay constant across such transformations. In skyscraper terms, we can think of the c as stretching each building by the same factor, and d as moving the whole city up and down, vertically, in the air—a move that doesn't change the volume of housing).

Thus, if p is the probability that the deer stop by (and we assume that whether you walk, or not, doesn't affect this probability), then the expected utility of walking is $1 - p$, and the expected utility of not walking is 0.9 (i.e., a guarantee of "no walk and garden").

But now suppose that when you try to do this EUM calc, you start to feel like there's "no way to estimate" the probability p that the deer will stop by. Maybe you think that humans can't reliably make estimates like this; or that somehow, applying the idea of probability here is confused—either the deer will stop by, or they won't.

Fine. Use whatever philosophy of probability you like. Refuse to make probability estimates if you like. But: are you going to go on the walk, or not? You still have to choose. And if you want to act consistently with EUM, then choosing *amounts* to a probability estimate. In particular, on EUM, you should go on this walk if and only if your probability

on the deer showing up is a less than 10% (the solution of $1 - p = 0.9$, and hence the point where EUM is indifferent between walking and not walking). So if you go, you're treating the probability as less than 10%; and if you stay, you're treating it as more—whatever the words you're saying in your head.

And now suppose we start with probabilities instead. Maybe you estimate a 1% probability that the deer show up, but you say: "I have no idea how to estimate my utility on walks vs. gardens—these things are just too different; life is not a spreadsheet." Ok, then: but the EU of walking is $.01 \cdot u(\text{walk and no garden}) + .99 \cdot (\text{walk and garden})$, and the EU of staying is $u(\text{no walk and garden})$. So you should go on this walk if and only if the value difference between "no walk and garden" vs "walk and garden" is >1% of the value difference between "walk and no garden" vs. "walk and garden" (if $u(\text{walk and no garden}) = 0$ and $u(\text{walk and garden}) = 1$, the EU of walking is .99, and the interval between "walk and no garden" vs. "walk and garden" is 1; so the indifference point is $u(\text{no walk and garden}) = .99$). Thus, by going on the walk, or not, you are implicitly assigning a quantitative ratio to the gaps between the value of these different outcome—even if you're not thinking about it.

OK, but suppose you don't want to start with a probability assignment *or* utility assignment. Both are too fuzzy! Life is *that much* not a spreadsheet. In that case, going or not going on the walk will be compatible with multiple EUM-approved probability-utility combinations. For example, if you stay to guard the garden, an EUM-er version of you could think that probability of deer is 1%, but that the garden is worth 150 walks; or, alternatively, that the probability is 10%, and the garden is worth 15 walks (see e.g. here for some discussion).

But each of these combinations will then have implications for your other choices. For example, if you think that the probability of deer is 1%, then (assuming that your marginal utility of money isn't sensitive to what happens with your garden, and that you want to say normal things about the probabilities of balls getting drawn from urn—see part 4), you should prefer to win $1M conditional on pulling balling number 1 out of a 20-ball urn, vs. conditional on the deer coming. Whereas if your probability on deer is 10%, you should prefer to win the million conditional on the deer coming instead. So if you have to make choices about how you feel about winning stuff you want conditional on balls being drawn from urns, vs. some "can't estimate the probability of it" event happening, we can pin down the probability/utility combination required to make your behavior consistent with EUM (if there is one) even further.

Similarly, if your garden is worth less than ten walks to you (let's again assume that each walk adds equal value, and that walk-value and garden-value are independent), then if e.g. you actually *see* the deer coming for your house—such that you're no longer in some kind of "probabilities are too hard to estimate mode," and instead are treating "the deer will eat the garden unless I stay to guard it" as certain or "known" or whatever—but you'll have to give up 20 walks to protect the garden (maybe the deer are going to stick around for a while, and you can't shoo them away?), then you shouldn't do it; you should let the deer eat the garden instead. So if that's *not* what you do, then we can't think of you as an EUM-er who stayed to guard the house due to some probability on losing a garden

worth less than ten walks.

Of course, it may be that you never, in your actual life, have to make such choices. But plausibly, there are facts about how you *would* make them, in different circumstances. And if those "woulds" are compatible with being an EUM-er at all (as I noted in my previous post, this is a very non-trivial challenge), they'll imply specific a probability assignment and a specific (unique-up-to-positive-affine-transformations) utility function (or at least, a constrained set of probability/utility pairings).

(Granted, some stuff in this vicinity gets complicated. Notably, if your choices can't be represented as doing EUM, it gets harder to say exactly what probability/utility assignment it makes sense to ascribe to you. This is related to the problem of "how do we tell what preferences a given physical system has?" that I mentioned last post—and I expect that they warrant similar responses.)

There's a vibe in this vicinity that's fairly core to my own relationship with EUM: namely, something about understanding your choices as always "taking a stance," such that having values and beliefs is not some sort of optional thing you can do sometimes, when the world makes it convenient, but rather a thing that you are always doing, with every movement of your mind and body. And with this vibe in mind, I think, it's easier to get past a conception of EUM as some sort of "tool" you can use to make decisions, when you're lucky enough to have a probability assignment and a utility function lying around—but which loses relevance otherwise. EUM is not about "probabilities and utilities first, decisions second"; nor, even, need it be about "decisions first, probabilities and utilities second," as the "but it's not action-guiding!" objectors sometimes assume. Rather, it's about a certain kind of harmony in your overall pattern of decisions—one that can be achieved by getting your probabilities and utilities together first, and then figuring out your decisions, but which can also be achieved by making sure your decision-making satisfies certain attractive conditions, and letting the probabilities and utilities flow from there. And in this latter mode, faced with a choice between e.g. X with certainty, vs. Y if heads (and nothing otherwise), one need not look for some independently specifiable unit of value to tally up and check whether Y has at least twice as much of it as X. Rather, to choose Y-if-heads, here, *just* is to decide that Y, to you, is at least twice as valuable as X.

I emphasize this partly because if—as I did—you turn towards the theorems I'll discuss hoping to answer questions like "would blah resources be better devoted to existential risk reduction or anti-malarial bednets?", it's important to be clear about what sort of answers to expect. There is, in fact, greater clarity to be had, here. But it won't live your life for you (and certainly, it won't tell you to accept some particular ethic—e.g., utilitarianism). Ultimately, you need to look directly at the stakes—at the malaria, at the size and value of the future—and at the rest of the situation, however shrouded in uncertainty. Are the stakes high enough? Is success plausible enough? In some brute and basic sense, you just have to decide.

In the next post, I'll start diving into the theorems themselves.

Part III

VNM, separability, and more

Previously in sequence: Skyscrapers and madmen; Why it can be OK to predictably lose

This is the third essay in a four-part series on expected utility maximization (EUM). This part examines three theorems/arguments that take probability assignments for granted, and derive a conclusion of the form: "If your choices satisfy XYZ conditions, then you act like an EUM-er": the von Neumann-Morgenstern theorem; an argument based on a very general connection between "separability" and "additivity"; and a related "direct" axiomatization of EUM in Peterson (2017). In each case, I aim to go beyond listing axioms, and to give some intuitive (if still basic and informal) flavor for how the underlying reasoning works.

10 God's lotteries

OK, let's look at some theorems. And let's start with von-Neumann Morgenstern (vNM)—one of the simplest and best-known.

A key disadvantage of vNM is that it takes the probability part of EUM for granted. Indeed, this is a disadvantage of all the theorems/arguments I discuss in this essay. However, as I discuss in the next essay, we can argue for the probability part on independent grounds. And sometimes, working with sources of randomness that everyone wants to be a probabilist about—i.e., coin flips, urns, etc—is enough to get a very EUM-ish game going.

Here's how I tend to imagine the vNM set-up. Suppose that you're hanging out in heaven with God, who is deciding what sort of world to create. And suppose, per impossible, that you and God aren't, in any sense, "part of the world." God's creation of the world isn't *adding* something to a pre-world history that included you and God hanging out; rather, the world is everything, you and God are deciding what kind of "everything" there will be, and once you decide, neither of you will ever have existed.

(This specification sounds silly, but I think it helps avoid various types of confusion—in particular, confusions involved in imagining that the process of ranking or choosing between worlds is in some sense part of the world itself (more below). On the set-up I'm imagining, it isn't.)

God's got a big menu of worlds he's considering. And let's say, for simplicity, that this menu is finite. Also, God has an arbitrarily fine-grained source of randomness, like one of those spinning wheels where you can mark some fraction as "A," and some fraction as "B." And God is going to ask you to choose between different lotteries over worlds, where a lottery XpY is a lottery with a p chance of world X, and a $1 - p$ chance of world Y (a certainty of X—i.e., XpX, or X(1)Y—counts as a lottery, too). If you say that you're indifferent between the two lotteries, then God flips a coin between them. And if you

"refuse to choose," then God tosses the lotteries to his dog Fido, and then picks the one that Fido drools on more.

Before making your choice, though, you're allowed to send out a ghost version of yourself to inspect the worlds in question at arbitrary degrees of depth; you're allowed to think as long as you want; to build ghost civilizations to help you make your choice, and so on.

And let's use "$\succ$" to mean "better than" (or, "preferred to," or "chosen over," or whatever), "$\succeq$" to mean "at least as good as," and "$\sim$" to mean "indifferent" (the curvy-ness of these symbols differentiates them from e.g. "$>$").

11 The four vNM axioms

Now suppose you want your choices to meet the following four constraints, for all lotteries A, B, etc.

1. Completeness: $A \succ B$, or $A \prec B$, or $A \sim B$.

Completeness says that either you choose A over B, or you choose B over A, or you say that you're indifferent. Some people don't like this, but I do. One reason I like it is: if your preferences are incomplete in a way that makes you OK with swapping any two lotteries you view as "incomparable," then you're OK being "money-pumped." E.g., if $A+ \succ A$, but both are incomparable to B, then you're OK swapping $A+$ for B, then B for A, then paying a dollar to trade back to $A+$.

Maybe you say: "I'll be the type of 'can't compare them' person who doesn't trade incomparable things. Thus, if I start with $A+$, and am offered B, I'll just stick with $A+$—and the same if I start with A" (see here for some discussion). But now, absent more complicated constraints about foresight and pre-commitment, you'll pass up free opportunities to trade from A to $A+$, via first trading for B.

Beyond this, though, refusing to compare stuff just looks, to me, unnecessarily "passive." It looks like it's giving up an opportunity for agency, for free. Recall that if you refuse to choose between two lotteries, God tosses them to his dog, Fido, and chooses the one that Fido drools on more. If you're *indifferent* between the two, then OK, fine, let Fido choose. Incomparability, though, is not supposed to be indifference. But then: why are you letting Fido make the call? Why so passive? Why not choose for yourself?

There's a lot more to say, here, and I don't expect fans of incompleteness to be convinced. But I'll leave it for now, and turn to the second vNM constraint:

Transitivity: If $A \succ B$, and $B \succ C$, then $A \succ C$.

Transitivity requires that you don't prefer things in a circle. Again, some people don't like it: but I do. And again, one reason I like it is money-pumps: if you prefer A to B to C to A, then you'll pay to trade A for C for B for A, which looks pretty silly.

Some people try to block money-pumps with objections like: "I'll foresee all my future options, make a plan for the overall pattern of choices to make, and stick with it." But we can create more complicated money-pumps in response. I haven't tried to get to the bottom of this dialectic, partly because I feel pretty uninterested in justifying transitivity violations, but see this book-length treatment by Gustafsson if you want to dig in on the topic.

Another argument against intransitivity is: suppose I offer you a choice between $\{A, B, C\}$, all at once. Which do you want? If you prefer A to B to C to A, then possibly, you're just stuck. Throw the choice to Fido? Or maybe you just pick something to choose when larger option sets are available. But even in that case, no matter what you choose, there was an alternative you liked better in a two-option comparison. Isn't that silly? (See Gustaffson (2013)).

Maybe you say: "what I like best depends on the choice set it's a part of. If you ask me: 'chocolate or vanilla?', I might say: 'vanilla.' But if you add: 'oh, strawberry is also available,' I might say: 'ah, in that case I'll have chocolate instead'" (see the literature on "expansion" and "contraction" consistency, which I'm borrowing this example from).

Sure sounds like a strange way to order ice cream, but we can point to more intuitive examples. Suppose that you prefer mountaineering to staying home, because if you stay home instead of mountaineering, you're a coward. But you prefer Rome to mountaineering, because that's "cultured" rather than cowardly. But you prefer staying home to Rome, because you actually don't like vacations at all and just want to watch TV. And let's say that faced with all three, you choose Rome, because not going on vacation at all, if mountains are on the menu, is cowardly (see Broome (1995), and Dreier (1996) for discussion). Isn't this an understandable set of preferences?

It's partly because of examples like these that I've made entire worlds the "outcomes" chosen, and made the process of choice take place "outside the world." Suppose, for example, that *God* offers you a choice between (A) a world where a travel agent offers you home vs. mountains, and you choose home, vs. (B) a world where a travel agent offers you home vs. Rome, and you choose home. *These are different worlds*, and it's fine to treat the choice of "home," in each of them, differently, and to view A as involving a type of cowardice that B does not. Your interaction with God, though, isn't like your interaction with the travel agent. It's not happening "inside the world." Indeed, it's not happening at all. So it doesn't make you a coward, or whatever you're worried about seeming like in the eyes of the universe.

What's more, if your choices between *worlds* start being sensitive to the option set, then I'm left with some sense that you're caring about something *other* than the world itself (thanks to Katja Grace for discussion). That is, somehow the value of the world ceases to be intrinsic to the world, and becomes dependent on other factors. Maybe some people don't mind this, but to me, it looks unattractive.

And as with incompleteness, option-set sensitivity also seems to me somehow compromising of your agency. Apparently, the type of force you want to be in the world depends a lot on how much of the menu God uncovers at a given time, and on the order of the list.

Thus, if he drafts the menu, or shows it to you, based on something about Fido's drool, the type of influence you want to exert is in Fido's hands. Why give Fido such power?

Maybe you say: "But if talking about whole worlds is allowed, then can't I excuse myself for that time I paid to trade apples for oranges at t_1, oranges for bananas at t_2, and bananas for apples at t_3, on the grounds that 'apples at t_1' worlds are different from 'apples at t_3' worlds?" Sure, you can do that if you'd like. As I said in the first essay: if your goal is to make up a set of preferences that rationalize some concrete episode of real-world behavior, you can. *But are those actually your preferences?* Do you actually like apples at t_3 better than apples at t_1? If not, you're not doing yourself any favors by trying so hard to avoid that dreaded word, "mistake." Some other person might've rationally made those trades. But not you.

Still, maybe you don't like the whole "choosing outside of the world" thing. Maybe you observe, for example, that all actual choice-making and preference-having occurs *in the world*. And fair enough. Indeed, there are lots of other arguments and examples in the literature on intransitivity, which I'm not engaging with. And at a certain point, if you insist on having intransitive preferences, I'm just going to bow out and say: "not my bag."

I'll add, though, one final vibe: namely, intransitive preferences leave me with some feeling of: "what are you even trying to do in this world?" That is, steering the world around a circular ranking feels like it isn't, as it were, *going anywhere*. It's not a force moving things in a *direction*. Rather, it's caught in a loop. It's eating its own tail.

Again: there's much more to say, and I don't expect fans of intransitivity to be convinced. But let's move on to our third condition:

3. INDEPENDENCE: $A \succ B$ if and only if $ApC \succ BpC$.

This is the type of principle that the "small probabilities are really larger conditional probabilities" argument I gave in part II relied on. And it does a lot of the work in the vNM proof itself.

To illustrate it: suppose that you prefer a world of puppies to a world of mud, and you have some kind of feeling (it doesn't matter what) about a world of flowers. Then, on INDEPENDENCE, you also prefer {p chance of puppies, $1-p$ chance of flowers} to a {p chance of mud, $1-p$ chance of flowers}. The intuitive idea here is that you've got the same probability of flowers either way—so regardless of how you feel about flowers, the "flowers" bit of the lottery isn't relevant to evaluating the choice. Rather, the thing that matters here is how you feel about puppies vs. mud.

We can also think about this as a consistency constraint across choices, akin to the one I discussed in my last essay. Suppose, for example, that God tells you that he's going to flip a coin. If it comes up heads, he's going to create a world of flowers. If it comes up tails, he's going to offer you a choice between puppies and mud. However, he's also offering you a chance to choose now, before the coin is flipped, whether he'll create puppies vs. mud *if* the coin comes up tails. Should you make a different choice now, about what God should do in the "tails" outcome, then you would make if you actually end up in the tails

outcome? No.

And I discussed in my last essay: if you say yes, you can get money-pumped (one notices a theme)—at least with some probability. I.e., God starts you with a puppies(.5)flowers lottery, you pay to trade for mud(.5)flowers, then if the choice comes up heads and you're about to get your mud, God offers you puppies instead, and you pay to trade back.

People often violate Independence in practice (see e.g. the Allais Paradox), but I don't find this very worrying from a normative perspective, especially once you specify that you're choosing "beyond the world," and that any feelings of fear/nervousness/regret need to be included in the prizes. To me, with this in mind, Independence looks pretty attractive.

> CONTINUITY: If $A \succ B \succ C$, there must exist some p and q (not equal to 0 or 1) such that $ApC \succ B \succ AqC$.

This is a bit more of a technical condition, required for the proof to go through. I'm told that if you weaken it, weaker versions of the proof can still work, but I haven't followed up on this.

To illustrate the idea, though: suppose that A is puppies, B is flowers, and C is mud, such that puppies $\succ$ flowers $\succ$ mud. Continuity says that is that if you start with puppies, there's some sufficiently small probability p on getting mud instead, such that you prefer taking that chance to switching to flowers with certainty. And similarly, if you start with mud, there's some sufficiently small probability of getting puppies instead such that you'll take a guarantee of flowers over that probability of trading for puppies.

This can feel a bit counterintuitive: is there really some sufficiently small risk of dying, such that I'd take that risk in order to switch from a mediocre restaurant to a good one? But I think the right answer is yes. Probabilities, recall, are just real numbers—and real numbers can get arbitrarily small. A one-in-a-graham's-number chance of dying is *ludicrously* lower than the chance of dying involved in walking a few blocks to the better restaurant, or picking up a teddy bear with padded gloves, or hiding in a fortified bunker, or whatever. By any sort of everyday standard, it's *really* not something to worry about. (Once we start talking about infinities, the discussion gets more complicated. In particular, if you value e.g. heaven infinitely, but cookies and mud only finitely, and heaven $\succ$ cookies $\succ$ mud, then you can end up violating continuity by saying that any lottery heaven(p)mud is better than cookies (thanks to Petra Kosonen for discussion). But "things get more complicated if we include infinities" is a general caveat (read: understatement) about everything in these essays—and in ethics more broadly).

We make money-pump-ish arguments for CONTINUITY, too—though they're not quite as good (see Gustaffson's book manuscript, p. 77). Suppose, for example, that our outcomes are: Gourmet $\succ$ McDonalds $\succ$ Torture, but for all values of p, you prefer McDonalds to Gourmet(p)Torture. And let's assume that you'd pay more than \$5 for Gourmet over McDonalds. Thus, Gourmet $\succ$ Gourmet $-$ \$5 $\succ$ McDonalds $\succ$ Gourmet(p)Torture, for any p. So if we start you off with Gourmet(p)Torture, you'll be willing to pay at least \$5 switch to Gourmet instead, even if we make p arbitrarily close to 1, and thus Gourmet(p)Torture arbitrarily

similar to Gourmet. This isn't quite paying $5 to switch to A from A. But it's paying $5 to switch to A from something arbitrarily similar to A—which is (arbitrarily) close.

Maybe you say: "yeah, I told you, I *really* don't like torture." And fair enough: that's why I don't think this argument is as strong as the others. But if you're *that* averse to torture, then it starts to look like torture, for you, is infinitely bad, in which case you may end up paying *arbitrarily* large finite costs to avoid *arbitrarily* small probabilities of it (this is a point from Gustaffson). Maybe that's where you're at (though: are you focusing your life solely on minimizing your risk of torture?); but it's an extreme lifestyle.

12 The vNM theorem

Maybe you like these axioms; or maybe you're not convinced. For now, though, let's move on to the theorem:

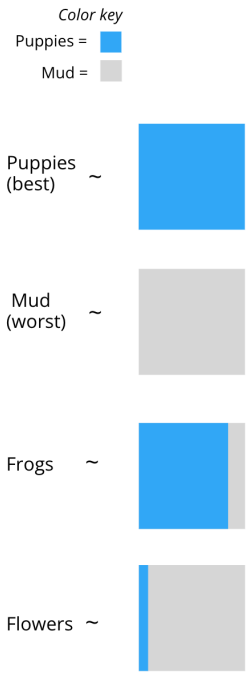
> vNM Theorem: Your choices between lotteries satisfy Completeness, Transitivity, Independence, and Continuity if and only if there exists some function u that assigns real numbers between 0 and 1 to lotteries, such that:
>
> (I) $A \succ B$ iff $u(A) \succ u(B)$.
> That is: the utility function mirrors the preference relation, such that you're always choose the lottery with the higher utility.
>
> (II) $u(ApB) = p \cdot u(A) + (1 - p) \cdot u(B)$.
> That is: the utility of a lottery is its expected utility.
>
> (III) For every other function u' satisfying (I) and (II), there are numbers $c > 0$ and d such that $u'(A) = c \cdot u(A) + d$.
> That is, the utility function u is unique up to multiplying it by a positive factor and adding any constant (a transformation that preserves the ratios between the differences between outcomes).

How does the proof work? Here I'll focus on showing that conditional on the axioms, and assuming a finite number of worlds, we can get to a u that satisfies properties (I) and (II). (Here I'm going off of the proof presented in Peterson (2009, appendix B), which may differ from the original in ways I'm not tracking. For the original discussion, see here.)

The basic thought is this. Because we've got a finite number of worlds, we can order them from best to worst, and we can think of each world as a lottery that returns that world with certainty. Now pick a world at the very top of your ordering—i.e., a world such that you don't like any other worlds better. Call this a "best world," O, and give it (and all worlds A such that $A \sim O$) utility 1. Also, pick a world at the very bottom of your list, call it a "worst world," W, and give it (and all worlds B such that $B \sim W$) utility 0.

The central vNM schtick is to construct the utility function by finding, for each world A, a lottery OpW such that $A \sim OpW$—and then we're going to think of p as the *utility for world A*. That is, your utility for A is just the probability p such that you're indifferent between A, and a lottery with p chance of a best world, and $1 - p$ chance of a worst world.

Thus, suppose we've got puppies $\succ$ frogs $\succ$ flowers $\succ$ mud. Puppies is best here, and mud is worst, so let's give puppies utility 1, and mud utility 0. Now: what's the utility of frogs? Well, for what probability p are you indifferent between frogs, on the one hand, and {p chance of puppies, $1-p$ chance of mud} on the other? Let's say that this probability is .8 (we can show that there must be one unique probability, here, but I'm going to skip this bit of the proof). And let's say your probability for flowers is .1. Thus, $u(\text{frogs}) = .8$, and $u(\text{flowers}) = .1$. Here it is in (two dimensional) skyscrapers:



From an "amount of housing" perspective (see part 1), this makes total sense. The best city has packed the entire 1×1 space with housing. The worst city is empty. Any other lotteries will have some amount of housing in between. And if we slide our probability slider such that it makes some fraction of the best city empty, we can get to any amount of housing between 1 and 0.

This sort of utility function, combined with some basic calculations to do with compound lotteries, gets us to (I) and (II) above. Let's walk through this. (This bit is going to be a bit dense, so feel free to skip to section 13 if you're persuaded, or if you start glazing over. The key move—namely, defining the utility of A as the p such that $A \sim OpW$—has already been made, and in my opinion, it's the main thing to remember about vNM.)

Note, first, that because of the Independence axiom, faced with OpW-style lotteries, you always want the higher probability of the best outcome O. After all, if the probability of O is different between the two, there will be some bit of probability space where you get O in one case, and W otherwise (some bit of the city where it's empty in one case, and full in the other)—and the rest of probability space will be equivalent to the same lottery in both cases. So because you prefer O to W, you've got to prefer the lottery with the higher probability of O.

For example, suppose that A is {90% puppies, 10% mud} and B is {50% puppies, 50% mud}. We can line up the outcomes here so they look like:

	10%	**40%**	**50%**
A	Mud	Puppies	Puppies
B	Mud	Mud	Puppies

The 10% and 50% sections are the same, so on INDEPENDENCE, this is really just a comparison between the outcomes at stake in the 40% section—and there, we've specified that you have a clear preference. So, you've got to choose A.

With this in mind, let's look at property (I): *for any lotteries A and B, $A \succ B$ iff $u(A) > u(B)$.*

Let's start with the proof from "$A \succ B$" to $u(A) > u(B)$. Suppose that $A \succ B$. Can we show that $u(A) > u(B)$? Well, $u(A)$ and $u(B)$ are just the probabilities of some lotteries OpW and OqW, such that $A \sim OpW$, $B \sim OqW$, and therefore, by our definition of utility, $u(A) = p$, and $u(B) = q$. But because $A \succ B$, it can't be that $p < q$—otherwise you'd be preferring a lower chance of the best outcome to a higher one. And it can't be that they're equal, either—otherwise you'd be preferring the same lottery over itself. So $p > q$, and hence $u(A) > u(B)$.

Now let's go the other direction: suppose that $u(A) > u(B)$. Can we show that $A \succ B$? Well, if $u(A) > u(B)$, then $p > q$, so OpW is a higher chance of the best world, compared with a lower chance in OqW. So $OpW \succ OqW$. So because $A \sim OpW$, and $B \sim OqW$, it must be that $A \succ B$.

Thus, property (I). Let's turn to property (II): namely, $u(ApB) = p \cdot u(A) + (1 - p) \cdot u(B)$. Here we basically just appeal to the following fact about compound lotteries. Suppose that $A \sim OqW$, and $B \sim OrW$. This means that you'll be indifferent between ApB and the compound lottery $(OqW)p(OrW)$. Overall, though, this compound lottery gives you a $pq + (1 - p)r$ chance of O, and W otherwise. So if we define s such that $pq + (1 - p)r = s$, $ApB \sim OsW$. So by our definition of utility, $u(ApB) = s$. And $u(A)$ was q, $u(B)$ was r, and s was just: $pq + (1 - p)r$. So $u(ApB) = p \cdot u(A) + (1 - p) \cdot u(B)$. Thus, property (II).

An example might help. Suppose you have the following lottery:

$$\{.3 \text{ chance of frogs}, .7 \text{ chance of flowers}\}.$$

What's the utility of this lottery? Well, we know from above that

$$\text{frogs} \sim \{.8 \text{ chance of puppies}, \ .2 \text{ chance of mud}\}$$

(call this lottery Y), such that $u(\text{frogs}) = .8$. And

$$\text{flowers} \sim \{.1 \text{ chance of puppies}, \ .9 \text{ chance of mud}\}$$

(call this lottery Z), such that $u(\text{flowers}) = .1$. So the compound lottery YpZ, where $p = .3$, is just a .3 chance of

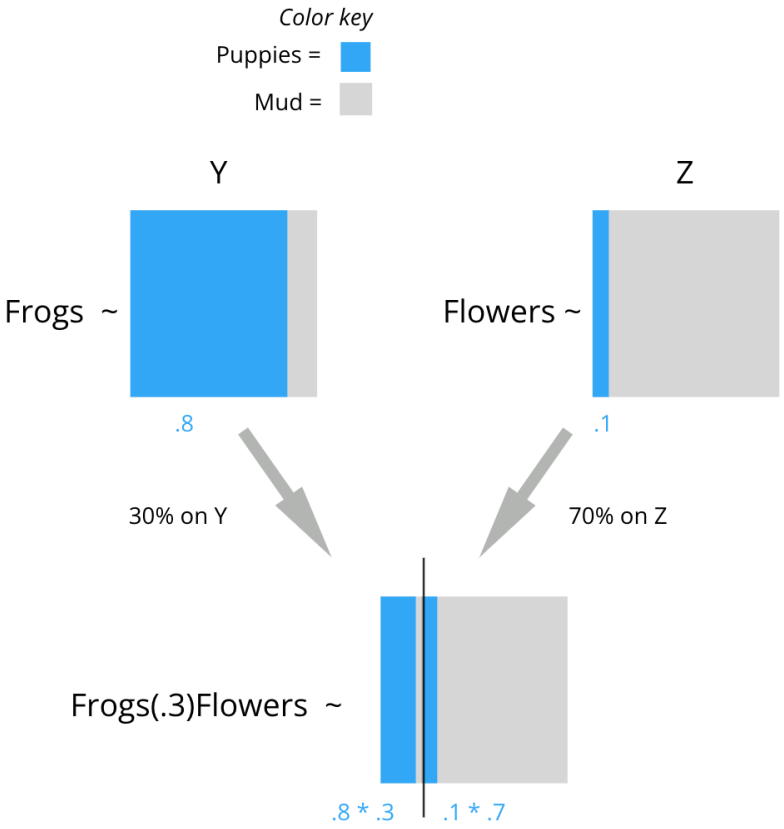
$$\{.8 \text{ chance of puppies}, \ .2 \text{ chance of mud}\},$$

and .7 chance of

$$\{.1 \text{ chance of puppies}, .9 \text{ chance of mud}\}.$$

But this just amounts to a $0.3 \cdot 0.8 + 0.7 \cdot 0.1$ chance of puppies, and mud otherwise. Thus, since frogs $\sim Y$ and flowers $\sim Z$,

$$u(\text{frogs}(.3)\text{flowers}) = .3 \cdot u(\text{frogs}) + (1 - .3) \cdot u(\text{flowers}).$$



This has been an informal and incomplete sketch, but hopefully it suffices to give some non-black-box flavor for how properties (I) and (II) fall out of the axioms above. I'm going to skip property (III), as well as the proof from properties (I)-(III) to the axioms—but see Peterson (2006) for more.

13 Separability implies additivity

Ok, so that's a bit about vNM. Let's turn to another, very general theorem, which focuses on the notion of "separability"—a notion that crops up a lot in ethics, and which, if accepted, often leads quickly to very EUM-ish and utilitarian-ish results.

Separability basically just says that your ranking of what happens at any combination of "locations" is independent of what's going on in the other locations—where "locations"

here can be worlds, times, places, people, person-moments, or whatever. If your preferences reflect this type of independence, and they are otherwise transitive, complete, and "reflexive" (e.g., for all A, $A \succeq A$), we can prove that they will also be *additive*: that is, your overall utility can be represented as the sum of the utilities at the individual locations.

To see why this happens, let's introduce a bit of terminology (here I'm following the presentation in Broome (1995, Chapter 4)—a book I recommend more generally). Let's say you got a set of alternatives A, and a (transitive, complete, and reflexive) preference ordering over these alternatives $\succeq$. Further, let's suppose that we've got an overall utility function U that assigns real numbers to the alternatives, such that $U(A) \geq U(B)$ iff $A \succeq B$. Finally, let's suppose that the alternatives, here, are vectors (i.e., lists) of "occurrences" at some set of "locations." Thus, the vector $(x_1, x_2, \ldots, x_n)$ has n locations, and the "occurrence" at location 1 is x_1. We'll focus on cases where all the alternatives have the same locations (a significant constraint).

Thus, for example, let's say you have three planets: 1, 2, and 3. Planet 1 can have mud, flowers, or puppies. Planet 2 can have rocks, corn, or cats. Planet 3 can have sand, trees, or ponies. So, ordering the locations by planet number, example alternatives would include: (mud, corn, ponies), (puppies, cats, sand), and so on.

Now suppose we fix on some subset of locations (a "subvector"), and we hold the occurrences at the other locations constant. We can then say that subvector X ranks higher than subvector Y, in a conditional ordering $\succeq'$, iff the alternative that X is extracted from ranks higher, in the original ordering $\succeq$, than the alternative Y is extracted from. That is, a conditional ordering is just the original ordering, applied to a subset of alternatives where we hold the occurrences at some set of locations fixed. And relative to a given conditional ordering, we can talk about how subvectors rank relative to each other.

Thus, for example, suppose we specify that planet 3 is going to have trees, and we focus on the conditional ordering over alternatives where this is true. Then we can say that the subvector (planet 1 with puppies, planet 2 with cats) ranks higher than (planet 1 with mud, planet 2 with rocks), relative to this conditional ordering, iff_{def} (puppies, cats, trees) $\succ$ (mud, rocks, trees).

We can then say that a subset of locations S is "separable," under an ordering $\succeq$, iff_{def} the ranking of subvectors at those locations is always the same relative to all conditional orderings that hold the occurrences at the other locations fixed. Thus, for example, suppose that some random stuff happens at planet 2 and planet 3—it doesn't matter what. If planet 1 is separable from these other planets, then if $(\text{puppies}, \text{whatever}_2, \text{doesn't-matter}_3) \succ (\text{mud}, \text{whatever}_2, \text{doesn't-matter}_3)$ for some values of "whatever" and "doesn't-matter," then that's also true for *all* values of "whatever" and "doesn't-matter." That is, if holding planet 2 and 3 fixed, switching from mud on planet 1 to puppies on planet 1 is ever an improvement to the overall situation, then it is *always* an improvement to the overall situation.

According to Broome, this means that a separable subset of locations can be assigned its own "sub-utility" function, which can be evaluated independent of what's going on at the other locations. (Broome seems to think that this is obvious, but I'm a little bit

hazy on it. I think the idea is that the ordinal ranking of occurrences at that subset of locations is always the same, so if the sub-utility function reflects this ranking, that's all the information you need about those locations. But the ordinal ranking can stay the same while the size of the "gaps" between outcomes changes, and I feel unclear about whether this can matter. I'll take Broome's word for it on this issue for now, though.) That is, if planet 1 is separable from the others, then $U(x_1, x_2, x_3)$ can be replaced by some other function $V(u(x_1), x_2, x_3)$, where $u(x_1)$ is the "subutility" function for what's going on at planet 1.

Now let's say that an ordering over alternatives is "strongly separable" iff_{def} every subset of locations is separable. And let's say that an ordering is "additively separable" iff_{def} it can represented as a utility function that is just the sum of the subutilities at each location: i.e., $U(x_1, x_2, \ldots, x_n) = u_1(x_1) + u_2(x_2) + \ldots + u_n(x_n)$. We can then prove:

An ordering is strongly separable iff it is additively separable.

(The proof here, according to Broome, is due to Gérard Debreu).

It's clear that additively separable implies strongly separable (for any subset of locations, consider the sum of their utilities, and call it x: for any fixed y, representing the sum of the utilities at the other locations, $x + y$ will yield the same ordering as you increase or decrease x), so let's focus on moving from "strongly separable" to "additively separable." And for simplicity, let's focus on the case where there are exactly three locations, and on integer utility values in particular, to illustrate how the procedure works.

Because the locations are strongly separable, we know that each location can be given its own sub-utility function, such that: $U(x_1, x_2, x_3) = V(u_1(x_1), u_2(x_2), u_3(x_3))$. So what we'll do is define subutility functions in each location, such that the overall utility function comes out additive. And we'll do this, basically, by treating an arbitrary subutility gap in the first location as "1," and then defining the subutilities at the other locations (and for everything else in location 1) by reference to what it takes to compensate for that gap in the overall ordering (Broome notes that we need an additional continuity condition to ensure that such a definition is possible, but following his lead, I'm going to skip over this). This gives us a constant unit across locations, which makes the whole thing additive.

In more detail: pick an arbitrary outcome at each location, and assign this outcome 0 utility, on the sub-utility function at that location. Now we've got $(0, 0, 0)$. And because the ordering is strongly separable, we can hold what's going on at location 3 fixed, and just make comparisons between sub-vectors at locations 1 and 2. So assign outcomes "1" in locations 1 and 2 such that: $(1, 0, 0) \sim (0, 1, 0)$. Assign "2" in the second location to the outcome such that $(1, 1, 0) \sim (0, 2, 0)$, "3" in the second location such that $(1, 2, 0) \sim (0, 3, 0)$, "-1" such that $(0, 0, 0) \sim (1, -1, 0)$, and so on. And note that because of separability, putting 0 in location 3 doesn't matter, here—we could've chosen anything. Thus, we've defined the subutilities at location 2 such that:

A1. $(1, b, c) \sim (0, b + 1, c)$ for all b and c.

Now do the same at the third location, using the gap between 0 and 1 at the first location as your measuring stick. I.e., define "1" in the third location such that $(1,0,0) \sim (0,0,1)$, "2" such that $(1,0,1) \sim (0,0,2)$, and so on. Again, because of strong separability, what's going on at location 2 doesn't actually matter, so:

A2: $(1,b,c) \sim (0,b,c+1)$ for all b and c.

And now, finally, do the same procedure at the first location, except using the gap between 0 and 1 in the second location as your basic unit. E.g., $(2,0,0) \sim (1,1,0)$, $(3,0,0) \sim (2,1,0)$, and so on. Thus:

A3: $(a+1,0,c) \sim (a,1,c)$ for all a and c.

But now we've defined a unit of utility that counts the same across locations (see footnote for more detailed reasoning).[1] And this means that the overall ordering has to be additive. This is intuitive, but one way of pointing at it is to note that we can now reduce comparisons between any two alternatives to equivalent comparisons between alternatives whose subutilities only differ at one location. Thus, for example: suppose you're comparing $(3,4,2)$ vs $(1,5,9)$, and you want to reduce it to a comparison between $(3,4,2)$ and $(x,4,2)$. Well, just shuffle the utilities in $(1,5,9)$ around until it looks like $(x,4,2)$. E.g., $(1,5,9) \sim (1,5-1,9+1) \sim (1,4,10)$. And $(1,4,10) \sim (1+8,4,10-8) \sim (9,4,2)$. So $(1,5,9) \sim (9,4,2)$. But comparing $(9,4,2)$ to $(3,4,2)$ is easy: after all, 9 is bigger than 3, and all the locations are separable. So because $(9,4,2) \sim (1,5,9)$, $(3,4,2) \prec (1,5,9)$. But this sort of procedure always makes the alternative with the bigger total the winner (Broome goes through a more abstract proof of this, which I'm going to skip).

How do we move from here to non-integer values? Well (again, assuming some sort of continuity condition), we can run this procedure for arbitrarily small initial value gaps—

[1] By linking A1 and A2 via their mutual reference to $(1,b,c)$, we know that:

A4: $(0,b+1,c) \sim (0,b,c+1)$, for all b and c.

But because locations 2 and 3 are separable, we can put in any other number at location 1 in A4, such that:

A5: $(a,b+1,c) \sim (a,b,c+1)$, for all a, b, and c.

But now set $b = 0$ in A5. This implies:

A6: $(a,1,c) \sim (a,0,c+1)$ for all a and c.

But now, by linking A6 with A3, via their mutual reference to $(a,1,c)$, we get:

A7: $(a+1,0,c) \sim (a,0,c+1)$, for all a and c.

And because locations 1 and 3 are separable from location 2, we substitute in "b" for "0", and get:

A8: $(a+1,b,c) \sim (a,b,c+1)$, for all a, b, and c

But now, equipped with A5 and A8, we're really cooking with additivity gas. Notably, we now know that:

A9: $(a+1,b,c) \sim (a,b+1,c) \sim (a,b,c+1)$, for all a, b, and c.

That is, we have a constant unit across locations.

e.g., .1, .01, .001, and so on. So successive approximations with finer and finer levels of precision will converge on the value of a given outcome.

Admittedly, the continuity condition here seems pretty strong, and it straightforwardly doesn't hold for finite sets of occurrences like the mud/flowers/puppies I've been using as examples. But I think the proof here illustrates an important basic dynamic regardless. And the overall result is importantly general across whatever type of "location" you like—generality that helps explain why addition shows up so frequently in different ethical contexts.

14 From "separability implies additivity" to EUM

Does this result get us to expected utility maximization? It at least gets us close, if we're up for strong separability across probability space—a condition sometimes called the "Sure-Thing Principle." Here I'll gesture at the type of route I have in mind (see also Savage for a more in-depth representation theorem that relies on the sure-thing principle).

First, let's assume that we can transform any lotteries ApB and CqD (and indeed, any finite set of lotteries) into more fine-grained lotteries with the same number of equiprobable outcomes—all without changing how much you like them (strictly, we need something about "can be approximated in the limit" here, but I'll skip this). This allows us to put any finite set of lotteries into the "same number of locations" format we assumed for our "strong separability iff additive separability" proof above.

Thus, for example, suppose that either God will flip a coin for puppies vs. corn (lottery A), or he'll give you a 20% probability of flowers, vs. a 80% probability of frogs (lottery B). We can transform this choice into:

States	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
Probability	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Lottery A	Puppies	Puppies	Puppies	Puppies	Puppies	Corn	Corn	Corn	Corn	Corn
Lottery B	Flowers	Flowers	Frogs	Frogs	Frogs	Frogs	Frogs	Frogs	Frogs	Frogs

Now suppose your ranking over lotteries is representable via an overall utility function $U(l)$, and it's strongly separable across states. This means that we can give each state n a sub-utility function u_n. And because strong separability implies additive separability, we know that there is some $U(x_1, x_2, \ldots, x_n) = u_1(x_1) + u_2(x_2) + \cdots + u_n(x_n)$.

The next claim is just that, given that all the states here are equally probable, we should be able to permute the outcomes of a lottery across states, without affecting the overall utility of the lottery. The intuition here is something like: "regardless of how much you like A or B, if I offer you a coin flip for A vs. B, it shouldn't matter which one gets paired with heads, and which with tails" (this is the type of intuition I appealed to in the second argument in part II).

Now, this principle won't hold in many real-life choices. Suppose, for example, that you just ate a sandwich that you think has a 50% chance of containing deadly poison. Now Bob offers you two lotteries:

Lottery X: $10M conditional on poisoned sandwich, and nothing otherwise.

Lottery Y: nothing conditional on poisoned sandwich, and $10M otherwise.

Are you indifferent between X and Y? No. You definitely want lottery Y, here—even if you think the two possible states are equally likely—because the money is a lot more useful to you if you're still alive. That is, the "states" and the "prizes" here interact, such that you value prizes differently conditional on the different states.

This type of problem crops up a lot for certain arguments for EUM (for example, Savage's). But I've set up my "hanging out with God" scenario in order to avoid it. That is, in my scenario, the "prizes" are entire worlds, such that Lottery X vs. Y is actually switching from "a world where I have $10M but am dead" to "a world where I have $10M but am alive" (and also: "nothing and alive" vs. "nothing and dead"), which are *different prizes*. And the "states," with God, are just flips of God's coin, spins of God's wheel, etc—states that do not, intuitively, interact with the value of the "worlds" they are paired with.

To the extent you buy into this "hanging out with God" set-up, then, the argument can proceed: *in this type of set-up*, you shouldn't care which worlds go with "heads" vs. "tails," and similarly for other equiprobable states. Thus, we can permute the "worlds" paired with the different, equiprobable states in a given "lottery-that's-been-transformed-into-one-with-equiprobable-states," without changing its utility. And this implies that our subutility functions for the different equiprobable states have to all be the same (otherwise, e.g., if one subutility function gave a different value to "frogs" vs. "flowers" than another, then the overall total could differ when you swap frogs and flowers around).

So now we know that for every lottery L, there is a single sub-utility function u, and a "transformed" lottery L' consisting entirely of equiprobable, mutually exclusive states (s_1, s_2, etc), such that $U(L') = u(x_1) + u(x_2) \ldots u(x_n)$. But now it's easy to see that $U(L')$ is going to be equivalent to an expected utility, because the number of states that give rise to a given world will be proportionate to the overall probability of that world (though again, technically we need something about limiting approximations here). E.g., for Lottery B above, you've got: two states with flowers, and eight with frogs, so $U(\text{Lottery B}) = 2 \cdot u(\text{flowers}) + 8 \cdot u(\text{frogs})$. But dividing both sides by the total number of states, such that we get $U'(\text{Lottery B}) = .2 \cdot u(\text{flowers}) + .8 \cdot u(\text{frogs})$, will yield the same overall ordering. Thus, your preferences over lotteries are representable as maximizing expected utility.

Again: not an airtight formal proof. But I find the basic thought useful anyway. (I think it's possible that something like this is the basic dynamic underlying at least part of Savage's proof, once he's constructed your probability assignment, but I haven't checked).

15 Peterson's "direct argument"

Let's look at one further argument for EUM, from Peterson (2017), which requires taking for granted *both* a probability assignment *and* a utility function, and which tries to show that given this, you should maximize expected utility.

Peterson is motivated, here, by a dissatisfaction with arguments of the form: "If your choices about lotteries meet XYZ conditions, then you are *representable* as an EU-maximizer" (see also Easwaran (2014)). He wants an argument that reflects the way in which probabilities and utilities seem *prior* to choices between lotteries, and which can therefore *guide* such choices.

He appeals to four axioms:

> P1. *If all the outcomes of an act have utility u, then the utility of that act is u.*

That is, a certainty of getting utility u is worth u. Sounds fine.

> P2. *If Lottery A leads to better outcomes than Lottery B under all states, and the probability of a given state in A is always the same as the probability of that state in B, then A is better.*

Again, sounds good: it's basically just "if A is better no matter what, it's better." This leads to questions in Newcomb-like scenarios, but let's set that aside for now: the "hanging out with God" set up above isn't like that.

> P3. *Every decision problem can be transformed into one with equally probable states, in a way that preserves the utilities of the actions.*

This is the principle we appealed to last section (as before: strictly, you need something about limiting approximations, here, but whatever).

> P4. *For any tiny amount of utility t, there is some sufficiently large amount of utility b, such that if you make one of any two equiprobable outcomes worse by t, you can always compensate for this by making the other outcome better by b.*

For example, suppose that, in Lottery X, God will give you puppies if heads, and frogs if tails. But now he proposes a tweaked lottery Y, which will involve making one of the puppies in heads feel a moment of slightly-painful self-doubt. P4 says that there must be *some* way of improving the frogs situation, in tails (for example, by adding some amount of puppy happiness), such that you're indifferent between X and Y. And further, it says that for whatever utility amount the puppy self-doubt subtracted (t), and whatever utility the puppy happiness added (b), you're always indifferent between an original coin-flip lottery, and a modified one where the heads outcome is worse by t, and the tails outcome is better by b.

As I see it, the main problem for Peterson is that this axiom is incompatible with various utility functions. Suppose, for example, that my utility function caps out at 1. And suppose that I've got a lottery that gives me 1 if heads, and 1 if tails. Here, there's nothing I can do to tails that will compensate for some hit to my utility in heads; I've already max-ed tails out, so any loss on heads is a strict loss. And we'll get similar dynamics with bounded utility functions more generally. And even for unbounded utility functions, P4

isn't obvious: it implies that you're always happy to subtract t from some outcome, and to add b to another equiprobable outcome, no matter the original utilities at stake.

Suppose, though, that we grant P4 (and thereby restrict ourselves to unbounded utility functions—which, notably, cause their own serious problems). Then, we can show that t and b have to be *equal*—otherwise, we can make you indifferent between two lotteries where one yields better outcomes than the other no matter what. Thus, consider a case where $b > t$ (the same argument will work if $b < t$):

State:	**H**	**T**
Lottery X:	u(Puppies)	u(Frogs)

Now we transform X into the equally-valuable X′:

Lottery X′:	$u(\text{Puppies}) - t$	$u(\text{Frogs}) + b$

And now we transform X′ into the equal-valuable X″:

Lottery X″:	$u(\text{Puppies}) - t + b$	$u(\text{Frogs}) + b - t$

But now, because $b > t$, we've added a positive amount of utility to both outcomes. So by P2, we're required to prefer X″ over X. (If b was less than t, we'd have subtracted something from both outcomes, and you'd be required to prefer X over X″.)

(We can also think of this as an additional argument for intuitions along the lines of "you should be indifferent about saving -1 life on heads, and +1 on tails." I.e., we can ask whether there's at least *some* number n of lives we could save on tails, such that you always become indifferent to saving -1 on heads and $+n$ on tails. But then, if n isn't equal to 1, and you apply the same reasoning to -1 on tails, $+n$ on heads, then we can make you indifferent to strict improvements/losses across both heads and tails.)

Equipped with $b = t$, together with P3, we can transform any lottery into a lottery with equiprobable states, where all of them yield the same utility. Thus, suppose $u(\textsf{Puppies}) = 9$, and $u(\textsf{Frogs}) = 4$, and we're trying to assess the utility of 20% you get puppies, 80% you get frogs (call this lottery R). Well, splitting it into equiprobable cases, and then successively adding/subtracting some constant unit of utility (e.g., our b and t—let's use 1) across states, we get:

State	**S1**	**S2**	**S3**	**S4**	**S5**
Probability	20%	20%	20%	20%	20%
Original R	9	4	4	4	4
R′	8	5	4	4	4
R″	7	5	5	4	4
R″	6	5	5	5	4
R‴	5	5	5	5	5

By P1, $u(R''') = 5$. So because $R''' \sim R$, $u(R) = 5$, too. And this will always be the average value across states—i.e., the sum of utilities, which is held constant by the shuffling, divided by the number of states (which, for a lottery with equally probable states, is just the expected utility—for n equally probable states, the probability of a state is just $1/n$, so

$U(R) = u(\text{S1})/n + u(\text{S2})/n + \dots)$.

Peterson's argument is similar in various ways to the "separability" argument I gave in the previous section. To me, though, it feels weaker. In particular, it requires a very strong assumption about your utility function. Still, to people attracted to positing, up front, some "constant units" that they always value the same (for example, if they think that adding a sufficiently sad puppy always makes an outcome worse by the same amount, and adding a sufficiently happy puppy always makes an outcome better by that same amount), Peterson's argument helps elucidate why expected utility maximization follows naturally.

OK: that's a look at three different theorems/arguments that try to move from "XYZ constraints" to "EUM" (or, "representable as EUM"). Presumably I lost a ton of readers along the way—but hopefully there are a few out there who wanted, and were up for, this level of depth.

In each of these cases, though, we took the probability part of EUM for granted. In the next (and final) essay in this series, I'll look at ways of justifying this part, too.

Part IV

Dutch books, Cox, and Complete Class

Previously in sequence: Skyscrapers and madmen; Why it can be OK to predictably lose; VNM, separability, and more

This is the final essay in a four-part series on expected utility maximization (EUM). This part focus on theorems that aim to justify the subjective probability aspect of EUM, namely: Dutch Book theorems; Cox's Theorem (this one is still a bit of a black box to me); and the Complete Class Theorem (this one also supports EUM more broadly). I also briefly discuss Savage, Jeffrey-Bolker, and a certain very general argument for making consistent trade-offs on the margin—both across goods, and across worlds.

16 Comparing with the urns

So we've seen three ways of arguing for EUM—an argument from the vNM axioms, an argument from the general connection between separability and additivity, and Peterson's "direct argument." In all of these cases, though, we had to assume some probability assignment. Let's look at that assumption more directly.

The "hanging out with a coin-flipping, urn-pulling God" set-up made the assumption of a probability assignment relatively innocuous, in virtue of the fact that basically everyone wants to be a standard probabilist about things like coins, urns, and spinning wheels. For other types of propositions, though (e.g., "what's the chance that some human walks on mars before 2100?"), some people, and some theories of probability (see here), start saying: "no, you can't put probabilities on things like that."

Still, fans of EUM often do. Indeed, they start putting probabilities on basically any kind of proposition you want—probabilities often understood to express some subjective level of confidence, and hence called "subjective probabilities." This section briefly describe a more way of thinking about this practice I often use in my own life (I also gestured at this in part II). Then I turn to some prominent theorems that fans of subjectivity probability often look to for support.

Suppose that God has already created a world. He's told you some stuff about it, but you don't know everything. Indeed, maybe God lets you descend into the world, live a particular life for a while, and then return to heaven, equipped with whatever knowledge of the world that life gave you.

And suppose that once you're back in heaven, no matter what world God has created, you always love heavenly ice-cream with the same level of passion, and strongly prefer it over nothing (feel free to pick a more meaningful prize if you'd like). Now God starts asking you questions of the following format: "would you rather I give you a heavenly ice-cream cone if (a) X is true of the world I created, or if (b) I pull a red ball out of an urn,

where the fraction of red balls is p?" (see Critch (2016) for more on this sort of technique).

Here, if we assume that you always want the ice-cream cone in the scenario you find more likely, we can then say that if you choose (a), then you think X more than p likely; if you choose (b), less; and if you're indifferent, then p is your probability on X.

What's more, if you're a standard probabilist about urns, I'm optimistic about making sure that you're a standard probabilist about "humans on mars"-ish things as well, on pain of various inconsistencies, guaranteed losses of probability on stuff you like, and other forms of silliness. I haven't worked through all the details here, but see footnote for examples.[2]

Now, this whole procedure can feel like a backwards way of assigning probabilities. Don't I need to know my probability on X, in order to choose between (a) and (b)? And fair enough: really, faced with (a) and (b), you still need to do some kind of brute "which is more likely" mental maneuver in relationship to X and p—a maneuver that we haven't really elucidated.

Still, though, even if this isn't a good *definition* of probability, I find thinking about subjective probabilities in terms of (a) vs. (b)-like choices a helpful tool—and one that makes "but you can't assign a probability to X" seem, to me, quite an unattractive form of protest. If you're going to make sensible (a)-vs.-(b)-like choices, then you effectively *have* to assign such probabilities—or at least, to act like you have.

And I prefer this way of thinking about subjective probability to salient alternatives like "fair betting odds" (see next section). In particular, thinking in terms of betting odds often

[2]Suppose that you try to give two inconsistent probabilities to a single proposition. For example, you say that you're indifferent between <ice cream if humans on mars by 2100> and <ice cream if red ball, where 10% of balls are red">; but you're *also* indifferent between <ice cream if humans on mars by 2100> and <ice cream if red ball, where 90% of balls are red>. Now, by transitivity, you're indifferent between an uncontroversially 90% chance of ice-cream, and an uncontroversially 10% chance—which is a pure loss of ice-cream probability.

Or suppose you try to violate the following probability axiom: $p(A \text{ or } B) = p(A) + p(B)$, where A and B are mutually exclusive. For example, consider an urn with 30% red balls, 30% blue balls, and 10% green balls. Suppose you say <ice cream if Trump is next US president> $\sim$ <ice cream if red ball>, <ice cream if Biden is next US president> $\sim$ <ice cream if blue ball>, but <ice cream if Trump or Biden> $\sim$ <ice cream if red, blue, or green ball>. Now we can start you off with <ice cream if red, blue, or green ball>, swap for <ice cream if Trump or Biden>, swap that for the two-ticket combo of <ice-cream if Trump> and <ice-cream if Biden>, then swap each of those for red ball and blue ball tickets, respectively. Thus, in combination, you've given up your green ball probability of ice cream, for nothing.

Or suppose you refuse to make these choices. "No, this is too subjective, I demand some kind of frequency, or propensity, or something. Without that, (a) and (b) are just incomparable." But now we get the same incompleteness issues I discussed in part 3. Notably, for example, if incomparability makes you OK with trading, then you can end up trading an uncontroversially 90% chance of ice-cream, for an "ice cream if humans on mars" ticket, for an uncontroversially 10% chance of ice-cream—thus, again, giving up ice-cream probability for free.

Or suppose you say: "I hate these fuzzy subjective things, I always choose the urn option." Then, not only will you'll end up choosing a one-in-a-zillion chance of ice-cream (God has access to *very* large urns) over <ice-cream if X-fairly-plausible-thing-that-you've-decided-can't-be-given-a-probability>, but you'll also choose a one-in-a-zillion chance of ice-cream over <ice-cream if *not* X-fairly-plausible-thing>. Thus, in conjunction, if we imagine a zillion ball urn with one red ball and one blue ball, and we pair "X" with red ball and "not X" with blue ball, you'll have preferred a "two in a zillion chance of ice cream" over a certainty of ice cream.

Or suppose you try to just "prefer the urn option other things equal" (see "ambiguity aversion"). Again you get negation problems. I.e., either your urn-derived probabilities on "X or not X" don't add up to 100% (such that you end up preferring a less-than-certain urn-based ice-cream to a certain non-urn one), or you can't, actually, consistently favor the urn option: any preference you give to the urn option, relative to X, comes at the cost of your preference for the urn option, relative to not-X.

requires thinking about *paying* as well as receiving money, and/or about *different* amounts of money, in different circumstances. I find this more cognitively burdensome than just asking questions like: "would I rather win {blah thing I definitely want} if X, or if a coin flip came up heads?" And such questions skip over complications to do with risk aversion and the diminishing utility of money: there's no downside to "betting," here, and the only thing that matters is that you want {blah thing} at all (and the same amount in both cases).

That said, note that this sort of set-up depends importantly on your *always* loving {blah thing} with the same level of passion, such that the only variable we're changing is the likelihood of getting it. As ever with these hanging-out-with-God scenarios, this isn't always realistic: in the real world, you might well prefer <ice-cream if the sandwich I just ate was normal> over <ice-cream if the sandwich was poisoned>, even if you think that sandwich very likely to have been poisoned, because you like ice-cream more if you're alive. In practice, though, even if you're a "enough with these unrealistic thought experiments" type, you can generally just *check* whether your preference for a given prize (e.g., $10K) actually varies in the relevant way. Often it won't, and the method will work fine even outside of thought-experimental conditions.

At least in the form I've presented it, though, this method isn't a formal argument or theorem. I'll turn to some of those now.

17 Dutch books

Let's start with what is maybe the most influential argument for probabilism: namely, the so-called "Dutch book theorem." This isn't my favorite argument, but it's sufficiently common that it's worth mentioning and understanding.

Consider a proposition X, and a ticket that says: "I'll pay you $1 if X is true." The basic set-up of the Dutch book argument is to require you to specify a "fair price" for any ticket like this—i.e., a price where you're equally happy to *buy* a ticket, at that price (and hence get some chance of $1), or to *sell* one (and hence incur some chance of owing $1).

Perhaps you're thinking: "wait, specifying a fair price on everything sounds like a pretty EUM-ish thing to do. Are you already assuming that I'm some sort of expected utility maximizer? And in particular, are you assuming that my utility in money is linear?"

That is, indeed, often the backdrop picture (we make the price of the ticket small, so as to make it plausible that your u($) is linear, on the margin, for amounts that small)—and it's one reason I like this approach less than e.g. the one in the previous section. But strictly, we don't need a backdrop like this. Rather, we can just prove stuff about fair prices like these, and leave their connection with "probability" and "utility" open.

In particular: we can prove that if your ticket prices fail to satisfy the probability axioms, you are vulnerable to accepting a series of trades that leave you with a guaranteed loss (we also prove the converse: if you satisfy the probability axioms, you're immune to such losses—but I won't focus on that here, and a glance at the SEP suggests that some further conditions are required). We imagine these trades being made with a "Dutch Bookie,"

who doesn't know anything more than you about X, but who takes advantage of the inconsistencies in your fair prices.

I'm not going to go through the full proof (see here), but as an example: let's show that if you violate the third probability axiom (e.g., $p(A \text{ or } B) = p(A) + p(B)$, if A and B are mutually exclusive), you can get Dutch booked. And let's think of your prices in terms of the fraction of $1 they represent (e.g. 30 cents = .3).

Suppose that A is "Trump is the next US President," and B is "Biden is next US president." And suppose your fair price for A is .3, and your fair price for B is .3 (see current Predictit odds here), but your fair price for "Trump or Biden is the next US president" is .7. Thus, the bookie gives you .3 for a Trump ticket, .3 for a Biden ticket, and he sells you a "Trump or Biden" ticket for .7. So you're at -.1 prior to the election; you'll get a dollar if Trump wins, a dollar if Biden wins, and you'll pay a dollar if Trump or Biden wins. Oops, though: now, no matter who wins, you'll be at -.1 at the end of the night, too (if it's Trump or Biden, you and bookie will swap a dollar for a dollar, leaving you where you started; and if neither wins, no money will change hands). A guaranteed loss.

Similarly, if you had .3 on Trump, .3 on Biden, and .5 on Trump-*or*-Biden, then the bookie will sell you a Trump ticket and a Biden ticket, and then buy a Trump-or-Biden ticket, such that, again, you'll be down -.1, with $1 coming if Trump-or-Biden, but paying out $1 if Trump, and $1 if Biden. Bad news.

There's a large literature on this sort of argument. On one interpretation, its force is pragmatic: obey the probability axioms (at least with your fair prices), or you'll lose money. This interpretation has to deal with various objections to the effect of: "what if I just refuse to post fair prices?" Or: "what if there are no Dutch bookies around?" Or "what if I can foresee what I'm in for, and refuse the sequences of trades?" (Or maybe even: "what if there are 'Czech bookies' around, who are going around buying and selling in a way that would guarantee me *profit*, if my prices violate the probability axioms?"). The vibes here are similar to some that come up in the context of money-pumps, but I haven't tried to dive in.

On another interpretation, the argument "dramatizes an inconsistency in your attitudes." For example: even granted that "Trump" and "Biden" are mutually exclusive, you'll pay a different price if I offer you "Trump" and separately "Biden," instead of "Trump or Biden," and this seems pretty silly. If that's the argument, though, did we need all the faff about bookies? (And as Hajek discusses, we'll have trouble proving the reinterpreted "converse" theorem—i.e., maybe it's true that if your fair prices violate the probability axioms, you necessarily have inconsistent attitudes, but is true that if you have inconsistent attitudes, you necessarily violate the probability axioms?)

My main worry about Dutch book arguments, though, is that I feel like they leap into a very EUM-ish mind-set, without explaining how we got there. In particular, they operate by making "what fraction of blah utility would you pay, to get blah utility if X?" a proxy for your probability on X, by treating money as a proxy for utility. But if I'm just starting out on my road to EUM, I haven't necessarily constructed some utility function that would allow me to answer this sort of question in a way that I understand (and saying: "well,

your utility per dollar is linear on the margin for small amounts of money, right?" won't disperse my confusion). Indeed, some of the most common mechanisms for constructing my utility function (e.g., vNM) *require* some antecedent probabilism to get going.

18 Cox's theorem

Let's turn, then, to a different argument for probabilism, which doesn't assume such EUM-ish vibes: namely, Cox's theorem (original paper here).

Cox's theorem assumes that you want to have some notion of a "degree of belief" or "plausibility," which can be represented by a real number assigned to a proposition (this has a little bit of a "hey wait, why do that?" vibe, but I'll set that aside for now). It then shows that if you want your "plausibility" assignment—call this $\mathsf{plaus}(x)$—to obey basic logic, along with a few other plausible conditions (see here for a summary), then your plausibilities have to be "isomorphic" to standard probabilities: i.e., there has to be some reversible mapping from your plausibilities to standard probabilities, such that we can construct a Bayesian version of you, whose probabilities track your plausibilities perfectly.

(Or at least: maybe we can? There are apparently problems with various statements of Cox's theorem—including the original, and the statement in Jaynes (1979)—but maybe they can be fixed, but maybe doing so requires an obvious additional principle? I definitely haven't dug into the weeds, here—but see e.g. Horn (2003) on "R4".)

What are these extra "plausible conditions"? There are three:

1. Given some background information A, the plausibility of B and C is some function of the plausibility of B given A, and the plausibility of C given (A and B). That is:

$$\mathsf{plaus}(B \wedge C|A) = F(\mathsf{plaus}(B|A),\ \mathsf{plaus}(C|A \wedge B)),$$

   for some function F.

Jaynes mentions an "argument from elimination" for this, but I'm pretty happy with it right off the bat (though I'm admittedly influenced by some kind of background probabilism). That is, it does just seem that given our current background evidence, the plausibility of e.g. *Trump wins the next election and Melania is vice president* should be some function of (I) the plausibility of *Trump wins the next election*, and (II) the plausibility of *Melania is vice president* | *Trump winning the next election*.

2. The plausibility of not-$B|A$ is some function of the plausibility of $B|A$. That is:

$$\mathsf{plaus}(\text{not-}B|A) = G(\mathsf{plaus}(B|A)).$$

Again, sounds great. The plausibility of *Trump doesn't win the next election* (conditional on our current background evidence) seems very much a function of the plausibility of *Trump wins the next election* (conditional on that evidence).

3. F and G are both "monotonic"—that is, they are either always non-decreasing, or non-increasing, as their inputs grow.

Also looks good. In particular: intuitively, the plausibility of B *and* C should go up as the plausibility of B goes up, and as the plausibility of C|B goes up. And the plausibility of not-B should go *down*, as the plausibility of B goes up.

From this, though (modulo the complications/problems mentioned above), we can get to the "isomorphic to Bayesianism" thing. I'm not going to try to go through the proof (which I haven't really understood), but I'll gesture at one part it: namely, the derivation of the "product rule" $p(B \wedge C|A) = p(B|A) \cdot p(C|A \wedge B)$.

As I understand it, the basic driver here is that because $\wedge$ is associative (that is, $\mathsf{plaus}(A \wedge (B \wedge C)) = \mathsf{plaus}((A \wedge B) \wedge C)$), we can show that F needs to be such that $F(x, F(y,z)) = F(F(x,y),z)$ (see here). And then we can prove, using a lemma from Aczél (see p. 16 here) that this means there is some (reversible?) function W such that $W(f(x,y)) = W(x) \cdot W(y)$, such that if we treat plug in $x = B|A$ and $y = C|A \wedge B$, we get the product rule we wanted.

I wish I understood the "lemma" part of this better. Talking about it with someone, I was able to get at least some flavor of what (I'm told) is the underlying dynamic, but I'm sufficiently hazy about it that I'm going to relegate it to a footnote.[3] Overall: I'm putting this one down as "still a pretty black box," and I'd love it if someone were to write a nice, intuitive explanation of what makes Cox's theorem work.

Assuming it does work, though, I think it's a substantive point in probabilism's favor. In particular, it's getting to probabilism from assumptions about the dynamics of the "plausibility" that seem to me weaker than—or at least, *different from*—the probability axioms. Thus, (1) and (2) above are centrally about what the plausibility of different propositions *depends on* (rather than how to calculate it), and (3) is about a very high-level and common-sense qualitative dynamic that plausibility calculations need to satisfy (e.g., things like "as a proposition gets more plausible, its negation gets less plausible"). And because of its more purely formal (as opposed to pragmatic) flavor, Cox's theorem suffers less from objections of the form "what if I just don't do deals with the Dutch bookie?" and "what if there aren't any Dutch bookies around?" (though this also renders it more

[3]Suppose we have some function $F(x_1, x_2, \ldots, x_i)$. And suppose that this function is associative, in the sense that the order of the inputs doesn't matter, and that the inputs can only take on a finite number of values. (Maybe it works in the infinite case too? Not sure.) Further, suppose we number the possible input values 1 through n. We can then put each x_i that gets fed into F through some further function, H, which spits out a vector with n slots, all of which are zero except for the slot corresponding to the value of x_i, which gets a 1. Thus, if $x_i = 52$, and 52 got given the number "3" in the process of numbering possible inputs to F, the third slot in the vector created by $H(x_i)$ gets a 1 (and the rest get zeros). But now, if you put all the inputs to F through H, and then add up the resulting vectors, you get an overall vector that effectively "counts" how many inputs to F took on the different possible values. And since the order doesn't matter, this overall vector preserves the information needed to get to the original output of F. That is, there is some function Z such that $Z(H(x_1) + H(x_2), \ldots, H(x_i)) = F(x_1, x_2, \ldots, x_i)$, and if Z is invertible (do we know this?) there's some $Z'(F(x_1, x_2, \ldots, x_i)) = H(x_1) + H(x_2), \ldots, H(x_i)$. So (modulo questions about invertibility?) this establishes some general connection be associativity and being some function away from an addition-flavored thing. And then if we exponentiate, we get to a multiplication flavored thing. And I'm told that Cox on the product rule is kind of a more specific version of this. Or something. I'm definitely waving my hands, here, and unsurprised if I'm making mistakes.

vulnerable to: "who cares about these constraints?").

19 The Complete Class Theorem

Let's look at one final way of trying to justify probabilism (as well as something like EUM): namely, the Complete Class Theorem (here I'm drawing heavily on a simplified version of Abram Demski's explanation here; see also these notes, and the original paper from Wald).

The set up here is something like the following. God presents you with a set of worlds, W, and tells you that you're going to be born into one of them $(W_1, W_2,$ etc$)$. Further, each of these worlds comes with a set of observations $(O_1, O_2, \dots)$ that you'd make, if you were in that world. However, multiple worlds are compatible with the same set of observations. Your job is to choose a policy that outputs actions in response to observations (you're also allowed to choose "mixed policies" that output actions with some probability). And when you take an action in a world, this yields a given amount of utility (the Complete Class Theorem starts with a utility function already available). And for simplicity, let's assume that the set of worlds, observations, and actions are all finite.

Thus, for example, suppose that you get one util per apple, and carrots are nothing to you. And your possible worlds are "normal better farm" world, where planting apples yields ten apples, and planting carrots yields ten carrots; and a "weird worse farm" world, where planting apples yields five *carrots*, and planting carrots yields five *apples*. In both worlds, your observations will be: an empty field, ready for planting. And your possible actions are: <plant apples, plant carrots, do nothing>.

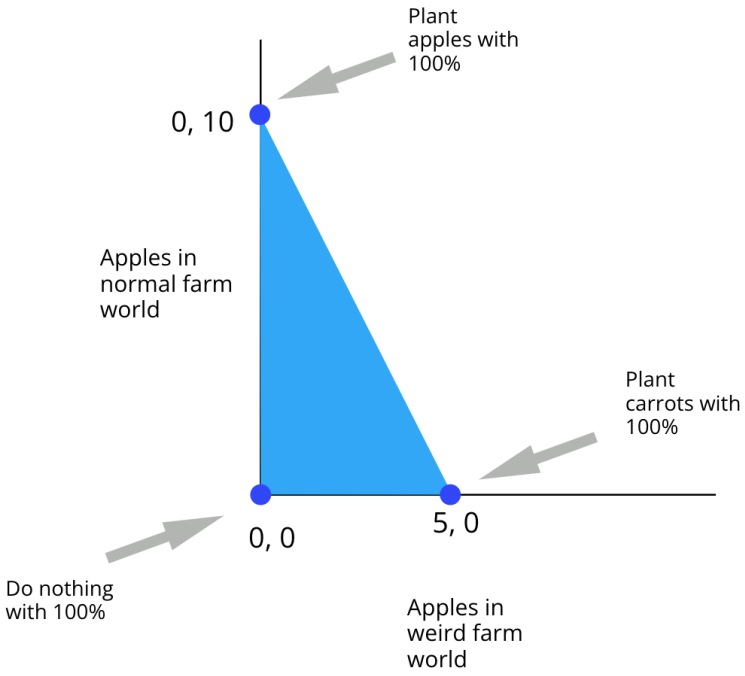
The complete class theorem says that your policy is pareto optimal (that is, there's no policy that does better in at least one world, and at least as good in all the others) *iff* there is *some* non-dogmatic probability distribution over worlds (i.e., a distribution that gives non-zero probability to all worlds) such that you're maximizing expected utility, relative to that distribution. That is, not throwing away utility for free, in some worlds, is equivalent to being representable as doing (non-dogmatic) EUM.

The proof from "representable as doing EUM" to "pareto optimal" is easy. Consider the probability distribution relative to which your policy is doing EUM. If your policy isn't pareto-optimal, there's some alternative policy that does better in some world, and worse nowhere. And because your probability distribution is non-dogmatic, it's got some probability on the "does better" world. Thus: there's some alternative policy that would get more expected utility, relative to your probability distribution. But we said that your probability distribution was maximizing expected utility—so, contradiction.

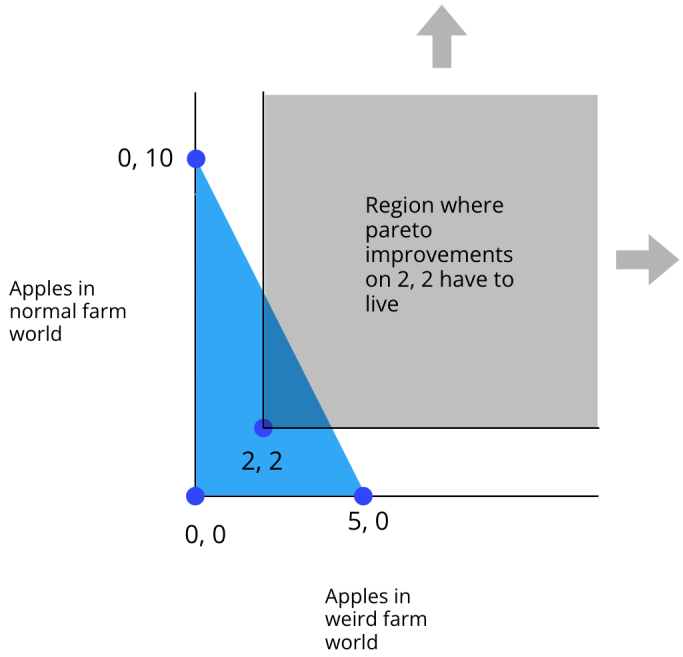
The proof from "pareto optimal" to "representable as doing EUM" is a bit more complicated, but it comes with a nice geometric visualization. Basically, each policy is going to be equivalent to a vector of real-numbers, one for each world, representing that policy's utility (or, for mixed policies, the expected utility) in that world. If there are n worlds, we can represent the set of available policies as points in an n-dimensional space. And

because you can pursue mixed policies, this set will be convex—e.g., for any two points in the set, the set also contains all the points on the line between them (the line drawn by mixing between those policies with p probability).

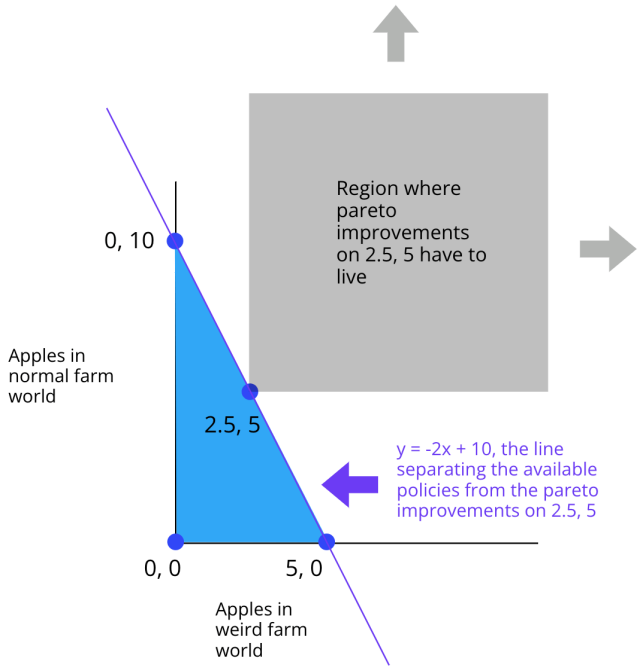
To illustrate, here is the space of available policies for the farming example above:



What's more, for any given point/policy in this space, we can define a region where a pareto-improvement on that policy would have to live—a region that would generalize the notion of "up and/or to the right," in two dimensions, to n dimensions. Thus, for a policy that yields two apples, in expectation, in each world, this region would be the grey box below:



Thus, my policy being pareto optimal is equivalent to the "pareto improvements on my policy" space being disjoint from the space of available policies. And this means that there will be some hyperplane that separates the two (this follows from the "hyperplane separation theorem," but it's also quite intuitive geometrically). For example, if my policy is $(5, 2.5)$—i.e., 50% on planting apples, 50% on planting carrots—then the hyper-plane separating the available policies from the pareto improvements on my policy is the line $y = -2x + 10$:



But we can use the vector normal to this hyperplane to construct a probability distribution, relative to which the policy in question is maximizing expected utility. In particular (here my understanding gets a little hazier): hyperplanes are defined as a set of points $(x_1, x_2, \ldots, x_n)$ that satisfy some equation of the form $a_1x_1 + a_2x_2 + \cdots + a_nx_n = c$, where the vector $(a_1, a_2, \ldots, a_n)$ is normal to the hyperplane. And the hyperplane separation theorem says that there will be some hyperplane, defined by vector v and some constant c, such that for any point x in the available policy space, and any point y in the pareto-improvement space, $v \cdot x \leq c$, and $v \cdot y \geq c$. What's more, because we chose the hyperplane to intersect with my policy, my policy is an x such that $v \cdot x$ actually just *equals* c. Thus, no other point in the available policy space yields a higher constant, when dot-product-ed with v.

But we can "shrink" or "stretch" a normal vector defining a hyperplane, without changing the hyperplane in question, provided that we adjust the relevant constant c as well. So if we scale v and c such that all the elements of v are between 0 and 1, and add up to 1 (I think we can ensure that they're non-negative, due to the 'up and to the right'-ness of the pareto-improvements region?), then we can treat the re-scaled vector as a probability distribution (here I think we need to assume that all the worlds are mutually exclusive). And thus, since the points x in policy space represented utilities in each world, the dot product $v \cdot x$ will represent the expected utility of that policy overall, and the constant c (the expected utility of my policy) will be highest expected utility that any available policy can achieve, relative to that probability distribution.

For example: in our weird farm example, the relevant hyper-plane for $(2.5, 5)$—and indeed, the relevant hyperplane for all pareto-optimal points—is defined by the line $y = -2x + 10$, which we can rewrite as $2x + y = 10$: i.e., the set of points w such that $(2, 1) \cdot w = 10$, where $(2, 1)$ is the vector normal to the line in question. Dividing both sides by 3, then, we get $(2/3, 1/3) \cdot w = 10/3$. Thus, the relevant probability distribution here is $2/3$rds on a "weird farm" world, and $1/3$rd on a "normal farm" world, and the expected utility of a policy that yields 2.5 given weird, and 5 given normal, is, indeed, $10/3(2/3 \cdot 2.5 + 1/3 \cdot 5)$.[4]

As ever in this series, this isn't a rigorous statement (indeed, jessicata brings up a possible counterexample here that I haven't tried to dig into, related to policies that are pareto-optimal, but only EU-maximizing relative a *dogmatic* prior). Assuming that the basic gist isn't wildly off, though, to me it gives some non-black-box flavor for how we might get from "pareto optimal" to "representable as doing EUM," and thus to "pareto optimal" iff "representable as doing EUM."

How cool is this "iff", if we can get it? Philosophically, I'm not sure (though geometrically, I find the theorem kind of satisfying). Here are a few weaknesses salient to me.

First, the theorem (at least as I've presented it) assumes various EUM-ish vibes up front. Centrally, it assumes a real-valued utility function. But also, because we're doing mixed

[4] And we can see why this is the probability distribution rationalizes any point on the pareto frontier: if you've got $2/3$rds on weird farm, and $1/3$rd on normal farm, then the expected number of apples you get from planting apples vs. carrots is the same: $1/3$rd $\cdot$ 10 for planting apples, and $2/3$rd $\cdot$ 5 for planting carrots. So you're happy with a policy that involves any p probability of planting apples, and $1 - p$ of planting carrots.

strategies (which themselves involve a bit of probability already), the utility of a policy in a given world is actually its expected utility right off the bat, such that the theorem is really about not throwing away *expected utility* for free, rather than utility proper. There may be ways of weakening these assumptions (though to me, getting around the latter one seems hard, given the centrality of convexity to the proof), but absent weakening, we're not exactly starting from scratch, EUM-wise.

Second, the theorem *doesn't* say a thing that you might casually (but quite wrongly) come away from it thinking: namely, that *given* a probability distribution and a utility function, if you don't maximize expected utility, you're doing something pareto-dominated. That's false. Suppose, in the example above, you're 99% that you're in a normal farm world, and 1% that you're in a weird farm world. The EV-maximizing thing to do, here, is to plant apples with 100% (EV: 9.9 utils). But planting carrots with 100% (EV: .05 utils) is pareto-optimal: nothing else does as well in the weird farm world. Put another way (see jessicata's comment here), CCT says that you can *rationalize* pareto-optimal policies as EU-maximizing, by pretending that you had a certain (non-dogmatic) probability over worlds (and that if you *can't* do this, then you're throwing away value for free). But once you actually have your utility function and probabilities over worlds, considerations about pareto-optimality don't tell you to do EUM with them. (And come to think of it: once we're pretending to have probability distributions, why not pretend that our probability distributions are dogmatic?)

Finally: how hard is it be pareto optimal/representable as doing EUM, anyway? Consider any set of observations, O, like the observation of sitting at your desk at home. And consider a random inadvisable action A, like stabbing the nearest pencil into your eye. And now consider a world W where you see your desk, and then, if you stab your pencil into your eye, the onlooking Gods will reward you maximally, such that any policy that puts less than 100% on eye-stabbing in response to desk-seeing does worse, in W, than 100% on eye-stabbing (thanks to Katja Grace for suggesting this sort of example). Boom: stabbing your pencil into your eye, next time you sit down at your desk, is pareto optimal, and representable as doing EUM. But it looks pretty dumb. And what's more, this seems like the type of counterexample we could generate for *any* action (i.e., X particular pattern of random twitching) in response to *any* observations, if the space of possible worlds that contain those observations is sufficiently rich (i.e., rich enough to contain worlds where the Gods reward that particular action maximally). That is, if you've got sufficiently many, sufficiently funky worlds, Pareto-optimality might come, basically (or entirely), for free.

Even granted these worries, though, I still count the complete class theorem as another point in favor of EUM. In particular, in toy examples at least, pareto-optimality does in fact rule out significant portions of policy space, and it's at least somewhat interesting that all and only the pareto-optimal points can be interpreted as maximizing (non-dogmatic) EU. Pareto-optimality, after all, is an unusually clear yes, rationality-wise. So even a loose sort of "pareto-optimal iff EUM" feels like it gives EUM some additional glow.

20 Other theorems, arguments, and questions

OK, these lasts two essays have been a long list of various EUM-ish, theorem-ish results. There are lots more where that came from, offering pros and cons distinct from those I've discussed here. For example:

- Savage seems to me a particularly powerful representation theorem, as it does not require assuming *probabilities* or *utilities* up front. I haven't been able to find a nice short explanation of Savage's proof, but my hazy and maybe-wrong understanding is that you do something like: construct the probability assignment via some procedure of the form "if you prefer to have the good prize given X than the good prize given Y, then you think X more likely than Y"; and then you get to EU-maximizing out of something like separability. The big disadvantage to Savage is that his set-up requires that any state of the world can be paired with any prize, which is especially awkward/impossible when "worlds" are the prizes. Thus, for example, for Savage there will be some action such that, if it's raining all day everywhere, you have a nice sunny day at the park; if the universe is only big enough to fit one person, you create a zillion people; and so on. (In a "hanging out with God" ontology, this would correspond to a scenario where God creates one world on the left, but doesn't tell you what it is, and then he offers you choices where he creates an entirely distinct world on the right, depending on which world he created on the left—where the assumption is that you don't care about the world on the left). This gets kind of awkward.
- My understanding is that Jeffrey avoids this problem by defining overall expected utilities for "propositions," and treating an action as a proposition (e.g., "I do X"). One problem with Jeffrey, though, is that he relies on a not-especially-intuitive "impartiality" axiom (Jeffrey remarks: "The axiom is there because we need it, and it is justified by our antecedent belief in the plausibility of the result we mean to deduce from it")—but maybe this is OK (I can kind of see the appeal of the axiom if I squint). Another issue that his framework fails to assign you a unique probability distribution—though some people (at least according to the SEP) seem to think that it can be made to do this, and others (see here and here) might think that failing to determine a unique probability distribution isn't problem. Overall, I'm intrigued by Jeffrey (especially because he avoids Savage's "sunny day given rain" problems) and would like to understand the proof better, but I haven't yet been able to easily find a nice short explanation, and this series is long enough.
- Yudkowsky, here, gestures at a very general argument for making consistent trade-offs on the margin, on pain of pareto-domination (thanks to John Wentworth for discussion). E.g., if you're trading apples and oranges at a ratio of 1:2 at booth A (i.e., you'll trade away one apple for more than two oranges, or an orange more than half an apple), but a ratio of 1:5 at booth B, then you're at risk of trading away an apple for three oranges at booth A, then four oranges for an apple at booth B, and thus giving away an orange for free. It's a further step from "your ratios on the margin need to be consistent" to "you have to have a well-behaved real-valued utility function overall" (and note that the relevant ratios can change depending on how many apples vs. oranges you already have, without risk of Pareto-domination—a fact that I think Yudkowsky's

presentation could be clearer about). But overall, we're getting well into the utility-function ballpark with this argument alone. And what's more, you can make similar arguments with respect to trade-offs across worlds. E.g., if you're trading tickets for "apples given rain," "oranges given rain," "apples given sun," and "oranges given sun," the same sort of "your ratios need to be consistent" argument will apply, such that in some sense you'll need to be giving consistent marginal "weights" to all four tickets—weights that can be treated as representing the EU gradient for an additional piece of fruit in a given world (though which won't, of themselves, determine a unique probability-utility decomposition).

- Other EUM-relevant arguments and theorems I'm not discussing include: Easwaren's (2014) attempt to derive something EUM-ish without representation theorems (though looks like his approach yields an incomplete ordering, and I tend to like completeness); the "accuracy" arguments for Bayesianism (see e.g. Joyce (1998) and Greaves and Wallace (2006)); and much else.

There are also lots of other questions we can raise about the overall force of these arguments, and about EUM more broadly. For example:

- *What about infinities, fanaticism, and other gnarly problem cases?* (My current answer: yeah, tough stuff, definitely haven't covered that here.)
- Are there alternatives to EUM that are comparably attractive? (My current answer: not that I'm aware of, though I haven't looked closely at the possibilities on offer (though I did at some point spend some time with Buchak (2014)). We know, though, that any alternative will have to violate various of the axioms I've discussed.)
- How useful is EUM in the real world? Obviously, we can't actually go around computing explicit probabilities and utilities all the time. So even if EUM has nice properties in theory, is there any evidence that it's an actively helpful mental technology in practice? (My current answer: it's an open empirical question how much thinking in explicitly EUM-ish ways actually helps you in the real world, but I think it's pretty clearly useful in at least some contexts—e.g., risk assessment—and worth experimenting with quite broadly; see Karnofsky (2021) for more.)

Clearly, there's a lot more to say. My main hope, in this series, has been to give some non-black-box flavor for the types of dynamics that cause EUM to show up at an ideal of decision-making from a variety of different angles—a flavor of the type my younger self was looking for, and which I hope can support emotional and philosophical clarity when making EUM-ish decisions. There is, in my opinion, something quite elegant here—and in a sense, quite deep. Few abstractions are so structurally relevant to our thought and action. So even if you continue to hold EUM as a distance, I think it's worth understanding.