World Domination in Decision Theory and Formal Epistemology

Stephen Yablo

July 17, 2021

Second of what is shaping to be a whole lot of drafts.
Advice and/or discouragement appreciated.

1 GAME

Let's play a game. I will describe it first at an abstract level, leaving until later (sections 2-6) the surprisingly vexed question of how to fill in details, given that our intuitions shift when the "same" game is expounded differently.

Your job is to choose between options A and B. A can done in one or more ways α, B in one or more ways β. The options overlap in that each way β of B-ing extends to two ways $\beta^\star$ and $\beta_\star$ of A-ing, and each way α of A-ing is a $\beta^\star$ or $\beta_\star$ for some way β of B-ing.[1]

Now the rewards. Each β wins you a penny, and likewise each $\beta_\star$. Both are as good as worthless, and *worthless* is the name we will know them by. The remaining αs — the $\beta^\star$s — win you a million dollars. The $\beta^\star$s are the *lucrative* αs.

This leaves a lot to the imagination, of course. But we can see already that A is a better option than B. I say this not because I have done expected utility calculations — those are not possible because we haven't been given probabilities. I say it because A in a good sense *dominates* B:

1. some ways α of A-ing — the $\beta^\star$s — win you more than any way β of B-ing does
2. no way β of B-ing wins you more than any way α of A-ing does.

I haven't said what kind of control you have over how A is realized. Again it doesn't much matter. Maybe an α is assigned to you randomly; maybe it's up to you which α occurs. A is a better choice than B either way, given dominance. I assume you would try to the extent possible to exercise option A in a lucrative way $\beta^\star$, rather than a worthless way $\beta_\star$. The argument for picking A over B does not depend on this.

2 SWITCHES

You are presented with two switches, one on the left and one on the right. Option B is flipping the left switch (Lefty). Option A is flipping either both switches (Lefty and Righty) or just Lefty (B in the obvious notation is L, A is $LR \vee L\bar{R}$.) Both switches must be flipped for the light to come on, which wins you \$1,000,000; one switch wins you a penny; flipping neither switch leaves you with nothing. Letting → be some suitable arrow,

[1] The overlap is for convenience. We can arrange for A and B to be incompatible, e.g. by letting A be done on Saturday and B on Sunday.

$$\begin{aligned} LR &\to \$1{,}000{,}000 \\ L\bar{R} &\to \$.01 \\ \bar{L}R &\to \$.01 \\ \bar{L}\bar{R} &\to \$.00 \end{aligned} \tag{1}$$

B's truthmakers β correspond to the ways λ of flipping Lefty; it will be simplest to pretend there is just one of these, and one way ρ of flipping Righty. A's truthmakers α are $\beta^\star = \lambda\rho$ and $\beta_\star = \lambda\bar{\rho}$. β will for convenience be identified with λ and the αs with $\lambda\rho$ and $\lambda\bar{\rho}$.[2] Pulling these threads together,

$A = LR \vee L\bar{R} =$ *You flip both switches or just Lefty*.
$B = L =$ *You flip Lefty* .
the βs are the ways λ of flipping Lefty.
the αs are the ways $\lambda\rho$ of flipping Lefty with Righty and $\lambda\bar{\rho}$ of flipping Lefty without Righty.
the $\beta^\star$s are the αs whereby you flip Righty. (2)
the $\beta_\star$s are the αs whereby you don't flip Righty.
each $\beta^\star$ gets (guarantees) you \$1,000,000.
no $\beta_\star$ gets you \$1,000,000; likewise no β gets you that much.

Does this way of filling in the details have the features we wanted, set out in section **1**?

- "each way β of B-ing extends to two ways α of A-ing"
    - yes, each λ extends both to $\lambda\rho$ and $\lambda\bar{\rho}$
    - these are both ways of A-ing, that is, of $(LR \vee L\bar{R})$-ing
- "each way α of A-ing is a $\beta^\star$ or $\beta_\star$, for some way β of B-ing"
    - yes, both $\lambda\rho$ and $\lambda\bar{\rho}$ (the αs) extend λ
    - α is a $\beta^\star$ ($\beta_\star$)-type extension if it contains ρ ($\bar{\rho}$)
- "no β wins you \$1,000,000, and likewise no $\beta_\star$"
    - it is not the case that the reward for flipping Lefty is \$1,000,000
    - nor is \$1,000,000 the reward for flipping only Lefty
- "the rest of the αs — the $\beta^\star$s — do win you a million dollars"
    - yes, \$1,000,000 is the reward for flipping Lefty and Righty

From this it seems that A is a better option than B. Done right, the $\lambda\rho$ way, it brings in \$1,000,000. No ways of B-ing are so lucrative. The reward for B-ing (done the only way it can be done, λ) is \$.01.

[2] Even bracketing this identification, flipping both cannot be considered a way of flipping Lefty; compare flipping Lefty within five miles of a burning barn. Flipping Lefty without Righty is not a way of flipping Lefty either. (The reason, looking ahead to section **13**, is that one doesn't flip Lefty *by* flipping it near a burning barn. Nor is it a truthmaker for *You flip Lefty* that you flip Lefty near a burning barn; the sentence isn't true by virtue of your doing that. For more on this and related topics (proportionality, relevance, exact truthmaking), see Yablo [2014], chapter 4; Fine [2017]; and Yablo [2022?].

3 APPLES

A more intriguing example builds on a case of Kit Fine's. *Alternative Eden* has infinitely many apples $\mathfrak{a}_i$. All but one are stamped with the Holy Seal of Approval, indicating it is safe to eat. The one remaining apple is $\mathfrak{badapple}$. Eating $\mathfrak{badapple}$ angers God to such an extent that he ruins your life and the lives of your children and theirs, never letting the matter drop as generation follows upon generation.

One gets into the neighborhood of the game sketched in section **1** by letting A be eating almost all of the apples (all but a finite number, "cofinitely many" in the jargon), and B be eating almost all apples distinct from $\mathfrak{badapple}$.[3] Eating almost all regular apples (apples distinct from $\mathfrak{badapple}$) gets you a penny, let's say. Eating almost all regular apples *and* $\mathfrak{badapple}$ *too* wins you knowledge of good and evil, valued conventionally at a million dollars. (The "$\mathfrak{bad}$" in our version means not that the apple is bad, but that it confers *knowledge* of badness — which is good.) The game may now be described as follows ($\mathfrak{b}$ is $\mathfrak{badapple}$):

A = *You eat almost all apples.*
B = *You eat almost all regular apples*—almost all apples $\mathfrak{b}$ aside.
each α is of the form: eating a certain cofinite array of apples, e.g. all but $\mathfrak{a}_5$.
each β is of the form: eating a cofinite array of regular apples, e.g. all but $\mathfrak{a}_5$.
each $\beta^\star$ is an α wherein you eat $\mathfrak{b}$.
each $\beta_\star$ is an α wherein you don't eat $\mathfrak{b}$.
each $\beta^\star$ gets you \$1,000,000.
no $\beta_\star$ gets you \$1,000,000; likewise no β gets you that much.
(3)

Does this way of filling in the details have the features we wanted?

- "each way β of B-ing extends to two ways α of A-ing"
  - yes, since eating a bunch of regular apples extends in two ways to eating a bunch of apples
  - one extension has you eating (chewing) $\mathfrak{b}$, the other has you eschewing $\mathfrak{b}$
- "each way α of A-ing is a $\beta^\star$ or $\beta_\star$, for some way β of B-ing"
  - yes, each way of eating almost all apples extends a way β of eating almost all $\neq \mathfrak{b}$
  - it's a $\beta^\star$ ($\beta_\star$)-style extension if α has you eating (not eating) $\mathfrak{b}$
- "no β wins you a penny, and likewise no $\beta_\star$s"
  - a penny is the reward for eating cofinitely many regular apples
  - and the reward too for eating cofinitely many apples none of which is $\mathfrak{b}$
- "all other αs — all $\beta^\star$s — win you a million dollars"
  - yes, \$1,000,000 is the reward for eating cofinitely many apples of which one is $\mathfrak{b}$

As before, A looks like a better option than B. Done right, it brings in \$1,000,000; there are no ways of B-ing that bring in \$1,000,000.

[3] If we want A and B to be incompatible, this can be arranged by putting A on one day and B on another. Or there could be two Alternative Edens. *Eden One* is where the act is done if A is selected over B; *Eden Two* is where it's done if B selected over A.

4 GAME

Consider now a game analogous to, but prime facie distinct from GAME. The specifics are laid out this time in terms of *worlds* w where a sentence is true, rather than ways for it to be true.

Your job is to choose between options A and B. A is true in all and only the as, B in all and only the bs. Given that S is true in w iff it has in w a truthmaker σ, w is an a iff some α obtains there, and a b iff some β obtains in w; but this is a side-remark, no part of GAME's definition.

Now the rewards. A-worlds a come in two types, those $\mathtt{a}^\star$ where you win $1,000,000, and those $\mathtt{a}_\star$ where you win $.01. B-worlds are typed correspondingly. They divide into those $\mathtt{b}^\star$ where you win $1,000,000, and those $\mathtt{b}_\star$ where you win $.01.

This leaves a lot to the imagination, of course. But we have seen no reason to think A preferable to B. I say this not because I have done expected utility calculations; that would not have been possible in the absence of probabilities . I say it because A does not *dominate* B in GAME as it did in GAME. The case for dominance in GAME was this, for randomly chosen αs and βs:

AB$_1$: α wins you at least as much, and sometimes more than, β does.

AB$_2$. $\neg$ (β wins you at least as much, and sometimes more than, α does).

The corresponding argument where GAME is concerned would proceed from analogous assumptions re randomly chosen as and bs that

AB$_1$. a wins you at least as much, and perhaps more than, b does

AB$_2$. $\neg$ (b wins you at least as much, and perhaps more than, a does).

But although AB$_2$ is correct, the "at least as much" claim in AB$_1$ is false — the $\mathtt{a}_\star$s win you *less* than the $\mathtt{b}^\star$s. For you by hypothesis win $.01 in each $\mathtt{a}_\star$, and $1,000,000 in each $\mathtt{b}^\star$.

Another way to see that AB$_1$ must fail is this. AB$_1$ assumes that there are "extra" A-worlds a, A-worlds where B is false.[4] How otherwise could there be A-worlds where you wind up richer than in any B-world? Yet as we're about to see, extra A-worlds simply do not exist in the worldly analogues of SWITCHES and APPLES. A world where you flip Lefty with or without Righty is ipso facto a world where you flip Lefty. (And vice versa.) That every A-world is a B-world — if the A-worlds just *are* the B-worlds — then clearly no A-worlds exist that make you richer than any B-world does.

5 SWITCHES

For a sense of how GAME might work in practice, we look for an analogue of SWITCHES centering on the *worlds* w where, say, *You flip Lefty and Righty, or Lefty without Righty* is true, rather than its ways of being true:

A = *You flip Lefty.*
B = *You flip Lefty and Righty, or else just Lefty*
each a is a w where you flip Lefty with Righty, or else without Righty
each b is a w where you flip Lefty
each $\mathtt{b}^\star$ is an a where you flip Righty in addition to Lefty
each $\mathtt{b}_\star$ is a a where you flip Lefty but not in addition Righty
in each $\mathtt{b}^\star$, you get $1,000,000.
in each $\mathtt{b}_\star$, you get at most a penny.

(4)

[4] As there were extra A-ways, αs not identical to any β.

This is similar to SWITCHES in that we are confronted in both with same options A, B; and the payoffs are the same as well, no matter how we conduct ourselves:

flip both switches	\$1,000,000
flip exactly one switch	\$.01
flip neither switch	\$0

Table 1: `SWITCHES` Payoffs

If the games are that similar, one might expect A to dominate B in `SWITCHES` as it did in SWITCHES. The argument in SWITCHES was, where λ (ρ) ranges over ways of flipping Lefty (Righty), α is either $\lambda\rho$ or $\lambda\overline{\rho}$, and $\beta = \lambda$,

AB_1 a randomly chosen way α of flipping Lefty with or without Righty wins you at least as much as, and perhaps more than, a randomly chosen way β of flipping Lefty

AB_2. $\neg$ (a randomly chosen way β of flipping Lefty wins you at least as much as, and perhaps more than, a randomly chosen way α of flipping Lefty with or without Righty)

The corresponding argument for `SWITCHES` would go like this

`AB`$_1$. a randomly chosen world `a` where you flip Lefty with or without Righty makes you at least as rich as, and perhaps richer than, a randomly chosen world `b` where you flip Lefty

`AB`$_2$. a randomly chosen world `b` where you flip Lefty makes you at least as rich as, and perhaps richer than, a randomly chosen world `a` where you flip Lefty with or without Righty.

`AB`$_2$ is true because some worlds where you flip Lefty with or without Righty make you a millionaire (on account of your flipping both); that's the best that occurs in any world, a fortiori as good as occurs in any world where you flip Lefty. But `AB`$_1$, if you think about it, is NOT true. Among worlds where you flip Lefty are some where you flip it with Righty; in these you come out a millionaire. No worlds whatever make you richer than that, so in particular none where you flip Lefty make you richer than that.

An easier way to see that `AB`$_1$ must fail is that it requires the existence "extra" A-worlds, A-worlds where B is false. But a world where you flip Lefty with or without Righty is ipso facto a world where you flip Lefty. And vice versa. That the A-worlds are just some (in fact, all) of the B-worlds entails that no `a` makes you richer than every `b`.

6 `APPLES`

Consider now a game like APPLES but prime facie distinct from APPLES. The rules are stated once again in terms of *worlds* where a sentence holds, rather than ways for it to hold,.

A = *You eat cofinitely many apples.*
B = *You eat cofinitely many regular apples* — apples distinct from $\mathfrak{badapple}$.
each $\mathtt{a}$ is a $\mathtt{w}$ where you eat cofinitely many apples, e.g., all but $\mathfrak{a}_5$
each b is a w where you eat cofinitely many regular apples, e.g., all but $\mathfrak{a}_5$
each $b^\star$ is an a where you eat $\mathfrak{badapple}$ (5)
each $b_\star$ is an a where you don't eat $\mathfrak{badapple}$
in each $b^\star$, you get \$1,000,000.
in each $b_\star$, you get at most a penny.

The new game is similar to APPLES in that we choose between the same two options (*A*, *B*) and the payoffs are unchanged as well:

eat cofinitely many apples including $\mathfrak{badapple}$	\$1,000,000
eat cofinitely many apples not including $\mathfrak{badapple}$	\$.01
eat fewer than cofinitely many apples	\$.00

Table 2: $\mathtt{APPLES}$ Payoffs

To run through a by now familiar dialectic, one might expect that *A* would be a better option than *B* in $\mathtt{APPLES}$ as it was in APPLES. Supporting *A* over *B* in APPLES, recall, was a dominance argument

AB$_1$. a randomly chosen way α of eating almost all apples wins you at least as much as, and possibly (if α involves $\mathfrak{b}$) more than, a randomly chosen way β of eating almost all apples $\neq \mathfrak{b}$

AB$_2$. $\neg$ (a randomly chosen way β of eating almost all apples $\neq \mathfrak{b}$ wins you at least as much as, and possibly more than, a randomly chosen way α of eating almost all apples)

The corresponding thought now would be

$\mathtt{AB}_1$. a randomly chosen $\mathtt{a}$ where you eat almost all apples makes you at least as rich, and possibly richer than, a randomly chosen way $\mathtt{b}$ where you eat almost all regular apples

$\mathtt{AB}_2$. $\neg$ (a randomly chosen $\mathtt{b}$ where you eat almost all regular apples makes you at least as rich, and possibly richer than, a randomly chosen $\mathtt{a}$ where you eat almost all apples)

$\mathtt{AB}_2$ is true because some worlds where you eat almost all apples make you a millionaire; that's the best that occurs in any world. But $\mathtt{AB}_1$ is NOT true. Among worlds where you eat almost all $\mathfrak{a}_i$s $\neq \mathfrak{b}$, there are some where you eat $\mathfrak{b}$ too; in these you're a millionaire. No worlds whatever make you richer than that. Hence in particular none where you eat almost all apples make you richer.

An easier way to see that $\mathtt{AB}_1$, must fail is this. $\mathtt{AB}_1$ requires the existence of "extra" *A*-worlds, *A*-worlds where *B* is false.[5] But any world where you eat almost all apples is ipso facto a world where you eat almost all of them $\mathfrak{badapple}$ aside. And vice versa. (A single apple can't make the difference between cofinitely many and not cofinitely many.) If the *A*-worlds just *are* the *B*-worlds, then clearly no *A*-world makes you richer than does any *B*-world.

[5] As there were extra *A*-ways, ways to eat almost all apples that were not ways of eating almost all apples $\neq \mathfrak{b}$.

7 Contradiction (?)

The story so far: Two game-types have been been sketched, one pitting A against B in a worldly setting ($\mathtt{G}_{AB}$), the other pitting A against B in a, so to speak, *wayward* setting ($\mathbf{G}_{AB}$). A dominance argument favors A over B in $\mathbf{G}_{AB}$, but no similar argument suggests itself for A over B in $\mathtt{G}_{AB}$. And there are reasons to doubt such an argument is possible: A and B are necessarily equivalent.

Why belabor all this? Players are confronted in each game with the same two options, A and B. And whichever switches they flip — apples they eat — they come away with the same reward. Decision problems that present you with the same options, and pay out the same regardless, would seem to be identical. (Where else would one look for a difference between them apart from options and payoffs?) If $\mathtt{G}_{AB}$ and $\mathbf{G}_{AB}$ present us with the same deliberative predicament, then there should presumably be a single truth as to what rational agents should do in that predicament. Is option A rationally preferable, or is it not?

There is, to say it again, a difference in how the options are *presented*. The choice between A and B was framed first in terms of *ways* of A-ing and B-ing, then in terms of worlds where A, B, obtain. The worlds are the same. Condition A (Lefty is switched with or without Righty; cofinitely many apples are eaten) is satisfied if and only condition B (Lefty is flipped; cofinitely many tree-of-life apples are eaten) is satisfied. That is why A looks no better or worse that B on the worldly framing.

But the *ways* in which these conditions are met are not the same. Eating all the apples including $\mathfrak{badapple}$ is a way of eating cofinitely many apples, but not a way of eating cofinitely many $\mathfrak{badapple}$ aside That is why A strikes us from a wayward perspective as a better option than B.

Can the duck/rabbit aspect here be written off as just a curiosity? A looks better than B seen through way-glasses, but not world-glasses. It would not be the first time beauty was chased backed to the eye of the beholder. But we are talking not about beauty here but decision making. And it is not clear that frame-relativity is a tolerable option in this case. Is there really no fact of the matter as to whether A is more advisable than B?

Whatever one thinks about relativism/non-factualism elsewhere, they are hard to swallow in a case like ours, where both perspectives are available and we have only the one chance to act. I for one would not want to get into the habit of defending my choices by saying that I was seeing the figure just then as a rabbit rather than a duck, or thinking of the glass for whatever reason as half-empty rather than half-full, or letting myself be guided by the operation's 80% success rate rather than the fact that one of five patients dies on the table. I would rather try to decide, when it comes anyway to worldly vs wayward framings, which of the two leads to the right results.

8 Intensionalism

To put the point less untendentiously: Suppose that eating cofinitely many apples is preferable to eating cofinitely many distinct from $\mathfrak{badapple}$, given that without $\mathfrak{badapple}$ you are out of luck reward-wise Then the standard worldly framing of A vs B is just wrong.

How would one attempt to establish this? A first pass argument, using APPLES rather than switches, might be constructed as follows:

[1] Eating almost all apples is a better option than eating almost all $\neq \mathfrak{b}$ ($A \succ B$).

[2] A and B are intensionally equivalent, so on the worldly framing identical.

[3] Intensionalism identifies options one of which is better than the other. [1][2]

[4] Intensionalism gets it wrong. [3]

The only really controversial premise here is [1]: $A \succ B$. Fortunately the logic of preference has a lot to tell us about statements of this type.[6] First a bit of ground clearing. Options X and Y are hyper-equivalent (short for "hyperintensionally" equivalent) if each way of exercising the one is a way of exercising the other, and vice versa.[7] It should be clear that

$$\begin{array}{l}\text{Eating almost all apples } (A) \text{ is hyper-equivalent to:}\\ \qquad\qquad\qquad \text{eating almost all} \neq \mathfrak{b}\ (B)\text{, or eating almost all and eating } \mathfrak{b}\end{array} \tag{6}$$

Writing $B^\star$ be the second disjunct on the last line — eating cofinitely many apples of which one is $\mathfrak{badapple}$ — (6) says that A is hyper-equivalent to $B \vee B^\star$. In symbols,

$$A \cong (B \vee B^\star). \tag{7}$$

This is interesting; for on the hand, hyper-equivalents would seem to be evaluatively indiscernible, while on the other, a disjunction's value would seem to flow in some natural, monotonic, way, from the values of its disjuncts. If A agrees in value with $B \vee B^\star$, then since $B^\star$ is clearly better than A (bringing the maximal reward every time), B will have to be worse than A, which is just what [1] says). Alternatively we could argue like this: B is worth *less* than $B^\star$, since it leaves you (unlike $B^\star$) sometimes with only a penny. But then B ought to be worth less too than $B \vee B^\star$. But this is the same (up to hyper-equivalence) as being worth less than A! Of course $A \succ B$ is the premise that needed bolstering.

Let's now try to develop the case for $A \succ B$ more carefully. The first assumption we'll need is that hyper-equivalents are equally valuable. Writing $X \approx Y$ for $X \succeq Y$ and $Y \succeq Y$,

$$\text{(HYC) if } X \cong Y \text{, then } X \approx Y. \tag{8}$$

(The label (HYC) is to remind us of what it says: hyperequivalence is a "congruence" with respect to preference- relations.) The second assumption goes back to Castañeda [1969]. von Wright had proposed in the early 1960s that $X \vee Y \succ Z$ only if $X \succ Z$ and $Y \succ Z$.[8] Castaneda finds this requirement too strict:

> [it] leads to the incomparability of $X \vee Y$ with Z, where, say, Value(X) = 1,000; Value(Y) = 20, and Value(Z) = 20. In a case like this it seems to be more satisfactory to say that the value of $X \vee Y$ is greater than the value of Z (Castañeda [1969], 258-9).

(Similarly it seems better to treat X as preferable to $Y \vee Z$ in cases where X and Y are tied and $X \succ Z$.) Applying this thought to the case where options agree in value because they're identical, we get *Castaneda's Principle*: If $X \succ Y$, then $X \succ \quad X \vee Y \quad \succ Y$.

Actually, we will need given (HYC) to state the principle a bit more cautiously. $X \vee Y$ is not going to be preferable to Y if the two are hyper-equivalent! And they will be hyper-equivalent, if X's ways of holding are a subset of Y's ways of holding. E.g., let $X = \$10$ and $Y = \$10 \vee \$1$. Then $X \succ Y$, it seems, but $X \vee Y \not\succ Y$, because Y and $X \vee Y$ are hyper-equivalent. The proper formulation of Castaneda's idea is

$$\text{(CAP) If } X \succ Y \text{, then } X \vee Y \succ Y \text{ (assuming } X \vee Y \ncong Y) \text{ and } X \succ X \vee Y \text{ (assuming } X \vee Y \ncong X) \tag{9}$$

And now we reasons as follows, starting from the fact that $B_\star$ (eating cofinitely many apples while NOT eating $\mathfrak{badapple}$) is quite definitely worse than $B^\star$ (eating cofinitely many apples while eating $\mathfrak{badapple}$):

[6] Notable contributions are Von Wright [1963], Rescher [1968], Chisholm and Sosa [1966], Castañeda [1969], Hansson [2001]

[7] Compare Fine's notion of *exact* equivalence in, for instance, Fine [2015].

[8] Von Wright [1963]

[1] $B^\star \succ B_\star$	$\beta^\star$ get you \$1,000,000; $\beta_\star$ gets you \$.01
[2] $B^\star \succ (B^\star \vee B_\star)$	[1], (CAP), $B^\star \ncong (B^\star \vee B_\star)$
[3] $B \cong (B^\star \vee B_\star)$	inspection
[4] $B^\star \succ B$	[2][3], (HYC)
[5] $(B^\star \vee B) \succ B$	[4], (CAP), $B \ncong (B^\star \vee B)$
[6] $A \cong (B^\star \vee B)$	inspection, (6)
[7] $A \succ B$	[3][4], (HYC)

The case this makes for $A \succ B$ — eating cofinitely many apples $\succ$ eating cofinitely many $\neq$ 𝖇 — seems pretty strong. Our anti-intentionalism is therefore on strong ground too; it rests mainly on $A \succ B$. The practice of running dominance arguments only in terms of worlds, and treating the failure of world-based dominance as somehow the end of the story, is beginning to look like a mistake.

9 Objections and Replies

Now we run through a series of objections, partly in hopes of countering them, partly to clarify what the difference is supposed to be between the orthodox (worldly) approach to decision-making and the approach in terms of ways.

9.1 The Same Worlds?

You keep on saying that the A-worlds are the B-worlds. Given that the αs are A's truthmakers, an A-world (B-world) is a w where you act α-ly for some α (β-ly for some β). Hence we have the same worlds twice only if

You act α-ish-ly in a world (for some α) iff you act β-ish-ly in that world (for some β). (10)

Is (10) plausible? Maybe not; for there are ways of A-ing whereby you eat more apples than contemplated in any way of B-ing. Let `w` for instance be a world where you eat *all* the apples. Eating all the apples is a way of eating cofinitely many apples, so `w` is a world where you act α-ly for some α. It does not look however like a world where you act β-ly for any β. Eating all the apples (𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊 included) is on your theory *not* a way of eating cofinitely many apples $\neq$ 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊.

Reply 1: This is a perceptive comment, but not so perceptive as to inspire doubts about the `a`s identity with the `b`s. It is true that `w` is not *by virtue of being a world where you eat all the apples* a β-world; for eating all of them is not a way of eating cofinitely many distinct from 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊. But `w` might qualify *another* way. `w` is also a world where you eat all the apples $\neq$ 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊, notwithstanding that that's not all you do there. Eating all the apples $\neq$ 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊 IS (unlike eating all the apples) a way of eating cofinitely many apples $\neq$ 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊. You do therefore enact a β in w. That β is eating all the apples $\neq$ 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊, not (what is admittedly *also* done) eating all the apples

But wait — *do* you perform that β? It may seem that to eat a bunch of apples $\neq$ 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊, you have to NOT eat 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊.

I reply that not eating 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊 is only *implicated* by "she ate all the apples $\neq$ 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊," not implied. *Eating all of the apples* $\neq$ 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊 may purport *conversationally* to be a complete description of your activities. But this has nothing to do with its literal content.[9] You may where literal content is concerned eat all the apples distinct from it, and then go ahead and eat 𝖇𝖆𝖉𝖆𝖕𝖕𝖑𝖊

[9] The debate in linguistics about covert exhaustivity operators could be relevant here; this is not the time.

too. Doing the both together is not a way of eating cofinitely many apples $\neq \mathfrak{badapple}$. But that does not stop it from taking place in *worlds* where you eat cofinitely many apples $\neq \mathfrak{badapple}$. A world where you β is one where you at least β, not (or not necessarily) a world where that is all you do.

9.2 Dominance Lost?

The claim was that A dominates B due to the existence of ways α of A-ing that earn you \$1,000,000, in the absence of ways β of B-ing that earn you that much. Isn't this called into question if there are ways of B-ing that *can* you earn you that much if done properly — if done, that is, along with $\mathfrak{badapple}$-eating? In fact *every* way of B-ing earns you that much if accompanied with $\mathfrak{badapple}$-eating. How does A still dominate B if each β leaves it still *open* to me to win \$1,000,000?

Reply 2: The worry is that A should not count as dominating B unless (#) there is a way α of A-ing such that α *guarantees* a better result than any β even *allows*. This condition is not met since eating all apples distinct from $\mathfrak{badapple}$ is a β that *allows* for a big payday (all βs do; they put only a lower bound on which apples are eaten). Hence eating all apples including $\mathfrak{badapple}$ does not (definitely) beat β. If (#) is an appropriate condition on dominance, then A does not dominate B.

Should we agree that A dominates B only if there is an α such that the best α-worlds are for all β better than the best β-worlds? I don't see why. α's superiority over the βs may lie not in its guaranteeing more than any β *allows*, but its guaranteeing more than any β *guarantees*. To see how (#) lets us down here, note that one could equally say of burping what has just been said in defense of B-ing. Burping too earns you \$1,000,000 when done right — "right" here meaning, after eating cofinitely many apples including $\mathfrak{badapple}$. I am not sure it should bother us that *eating cofinitely many apples of which one is* $\mathfrak{badapple}$ fails to (#)-dominate *eating cofinitely many* $\neq \mathfrak{badapple}$, when it does not (#)-dominate burping either, despite being clearly dominant in the sense that matters to rational decision-making.

Now, eating cofinitely many apples $\neq \mathfrak{badapple}$ is admittedly not as useless as burping. For although it's the αs, not the βs, that bring in the reward, each α *contains* a β. One can win the big bucks without burping, but not without B-ing.

This does admittedly give B-ing an advantage over burping. But the choice we face is between B-ing and A-ing. And when it comes to *this* choice, there is really no contest. That each β gets us *part* of the way to locking in \$1,000,000 can hardly prevent A from dominating B, if some αs get us *all* the way to locking in \$1,000,000 (and no β does).

It appears then that A does, in a good, deliberation-guiding, sense dominate B after all, never mind that for each way β of B-ing, some β-worlds are as good as they come. It's enough that choosing A potentially *guarantees* you a big-payday world; each $\alpha^\star$ suffices for one. Choosing B only keeps you in the running for a big-payday world. There are no βs whatever such that $\beta \mathrel{\Box\!\!\rightarrow}$ \$1,000,000. The best that can be said about any of them is that $\beta \mathrel{\Box\!\!\rightarrow}$ \$1,000,000 has non-zero probability. Depending on how closeness is judged, there's a *chance* you'd eat $\mathfrak{badapple}$ supererogatorily as it were, in addition to the cofinitely many regular apples you are called on to eat by β.

9.3 Just a Chance?

"There's a chance you'd eat $\mathfrak{badapple}$" — quite the understatement. It ignores how the B-chooser, Bina call her, would actually exercise her chosen option. Certainly there are no βs such that Bina *must* eat $\mathfrak{b}$ (and so win \$1,000,000) if she βs. But that is just to say just that β does not *strictly* imply winning \$1,000,000; $\beta \not\Rightarrow$ *\$1,000,000*. And the question of what Bina *would* do/win concerns not strict implication $\Rightarrow$ but counterfactual implication $\mathrel{\Box\!\!\rightarrow}$.

Knowing as she does that eating $\mathfrak{badapple}$ too will bring her a huge reward, it stands to reason that Bina would, if she β'd, supplement that so far non-renumerative activity with some $\mathfrak{b}$-eating. Which suggests that $\beta \mathrel{\Box\!\!\rightarrow}$ *\$1,000,000* has probability 1. Bina *could* if she played her cards wrong enact β without winning \$1,000,000. But she would play her cards right.

Why does this matter? Well, we assumed, when arguing that $A \succ B$, that *\$1,000,000* $\succ B$. We said in defense of this assumption that B doesn't *guarantee* Bina \$1,000,000, while \$1,000,000 obviously does guarantee her \$1,000,000. But, whether B guarantees her a big reward seems irrelevant to the question of choiceworthiness, if a big reward is what she'd get in fact. Pie-or-poison is every bit as good an option as pie, if the outcome either way is the same: you wind up eating pie. The causal decision theorist takes a similar line, when she defines an act X's expected utility as

$$\text{(CDT)} \quad \mathrm{U}(X) \;=\; \Sigma_O \quad \mathrm{pr}(X \mathrel{\Box\!\!\rightarrow} O) \times \mathrm{V}(O). \tag{11}$$

.

U($B^\star \vee$ B) and U(B) are going to agree, on this formula, if as we're contemplating pr($B \mathrel{\Box\!\!\rightarrow}$ \$1,000,000) = pr(($B^\star \vee B$) $\mathrel{\Box\!\!\rightarrow}$ \$1,000,000) = 1.

Reply 3: The objection assumes that it is up to the agent whether she eats only regular apples or also $\mathfrak{b}$. But the line between what's under Bina's control and what's left to chance can be drawn how we like. Imagine it goes like this: when Bina shows up with her basket, a chance mechanism half the time throws $\mathfrak{b}$ in, where it mixes undetectably with the other apples. What is in Bina's control is whether to β. But supposing she does, whether it is $\beta \wedge E\mathfrak{b}$ that obtains ($E\mathfrak{b}$ is eating $\mathfrak{badapple}$), or $\beta \wedge \neg E\mathfrak{b}$, is decided quite independently (pr($E\mathfrak{b}$) = pr($E\mathfrak{b} \mid \beta$) = 1/2). Castaneda draws attention to this kind of choice structure in the paper already mentioned. One

> can bring about ... $P \vee Q$ either by bringing about P directly, or by bringing about Q directly, or by [activating] a process that will end up either with the realization of P or with the realization of P. (Castañeda [1969], 263)

P in our case is $\beta \wedge E\mathfrak{b}$, and Q is $\beta \wedge \neg E\mathfrak{b}$. Bina can bring about their disjunction, but which disjunct obtains is decided by chance.

With the case filled out liked this, our earlier contention that $B^\star \succ B$ is clearly correct, for $B^\star$ always brings in \$1,000,000, while in B-worlds we get oftentimes much less, depending on the outcome of a chance process. Castaneda's Principle takes us from $B^\star \succ B$ to $B^\star \vee B \succ B$ as before — from which it is a short step by congruence to $A \succ B$.[10]

9.4 Control Issues

The chancy version of G_{AB} is underdescribed. Is it only if Bina chooses B that she lacks control over whether $\mathfrak{b}$ is eaten? Or does she lack it as well if she chooses A? Either way a problem arises.

Say it is only the B-chooser who loses control over $\mathfrak{b}$'s fate. Then it is not true after all that the same worlds are in play however Bina chooses. One choice (A) leaves her able to eat $\mathfrak{b}$, the other (B) deprives her of that ability. But the purported counterexample to intensionalism requires Bina to have the same possible futures before her, apple-selection-wise, whatever she decides.

If the chance mechanism takes control from the A-chooser too, then A's supposed advantage over B — done right, it wins you \$1,000,000 — evaporates. Whether A is done right is no more under the A-chooser's control than which kind of B-world she ends up in is under the B-chooser's control.

Now we see how the argument at the end of section **8** misfires. Castaneda's Principle needs to be understood a certain way. $X \succ Y$ does not ensure $X \vee Y \succ Y$ unless the agent is able to *take advantage* of the opportunity $X \vee Y$ presents — by steering events towards the ways of $X \vee Y$-ing that confer on X its advantage over Y. But if the agent is able to steer events in the proper direction, then Castaneda's Principle fails in another way. It no longer follows from $X \succ Y$ that $X \succ X \vee Y$ since the agent will exercise the $X \vee Y$ option by X-ing. This destroys the argument since it starts with precisely

[10] For the first step we need that $B^\star \vee B \ncong B$. How is this established? Take any β you like, $\beta \wedge E\mathfrak{b}$ is a way for $B^\star$ to hold, and hence a way for $B^\star \vee B$ to hold. But although $\beta \wedge E\mathfrak{b}$ is *true* in some nearby B-worlds, it is not a *way* for B to hold, given the irrelevance to B of whether $\mathfrak{badapple}$ is eaten.

this inference: $B^{\star} \succ B_{\star}$ (line [1]) is taken to imply $B^{\star} \succ (B^{\star} \vee B_{\star})$ (line [2]). Assuming complete control, the $B_{\star}$ sub-option, which brings in only a penny, will never be exercised. If $B_{\star}$ is never exercised, it would be better surely to say that $B^{\star} \approx (B^{\star} \vee B_{\star})$.

Reply 4 The $X \succ Y \Rightarrow X{\vee}Y \succ Y$ side of Castaneda's Principle is acceptable, you're saying, only if it is up to the agent how the relevant options are exercised, that is, which of the possible ways of X-ing, Y-ing, etc are enacted. If it is up to the agent, though, then the other side of Castaneda's Principle — $X \succ Y \Rightarrow X \succ X{\vee}Y$ fails, in a way that directly undercuts the section **8** argument.

But, the $X \succ Y \Rightarrow X{\vee}Y \succ Y$ sub-principle assumes at most that the agent has *some* control (it's enough, really, that X-ing could occur, through the agent's efforts or not). While the objection to $X \succ Y \Rightarrow X \succ X{\vee}Y$ assumes that the agent has *complete* control and will never Y given that X is better. This is hardly a vice like grip we're caught in. It is not all that embarrassing if CAP applies, and the argument for $A \succ B$ goes through, only when X is neither precluded nor guaranteed.

The issue this objection really raises is how to square our hyperintensionality-favoring contentions about qualitative preferability with the practice of quantitative decision theory, taken almost universally to be an *intensional* theory that wants no truck with fine-grained distinctions. How can A can be preferable to B, when — given that the eat-almost-all-apples-distinct from $\mathfrak{b}$-chooser would eat $\mathfrak{b}$ too, "on the side" as it were — the expected utilities agree? An excellent question which we begin to grapple with (grappling is all we'll manage unfortunately) now.

10 Expected Utility

Our principal claim: A is a better option than B, notwithstanding the fact that A and B are intensionally equivalent. In support of this we offered a kind of ways-based dominance argument (section **8**). That argument may seem far from decisive, relying as it does on assumptions like *Hyperequivalent Congruence*

(HYC) if $X \cong Y$, then $X \approx Y$.

and *Castaneda's Principle*

(CAP) If $X \succ Y$, then $X{\vee}Y \succ Y$[11] and $X \succ X{\vee}Y$[12]

These are *comparative* principles. And as we know, qualitative and comparative principles have come in for much less discussion than the quantitative principles studied in causal decision theory, e.g.,

$$\begin{array}{l} \text{(UCP) } X \succ Y \text{ iff} \\ \qquad \mathrm{U}(X)\ (= \Sigma_O\ \mathrm{pr}(X \mathrel{\Box\!\!\rightarrow} O) \times \mathrm{V}(O)) \quad > \quad \mathrm{U}(Y)\ (= \Sigma_O\ \mathrm{pr}(Y \mathrel{\Box\!\!\rightarrow} O) \times \mathrm{V}(O)). \end{array} \tag{12}$$

If $A \succ B$ as claimed, then U(A) ought other things equal to be greater than U(B). And now we run into trouble. A can come out with a higher expected utility than B only if $A \mathrel{\Box\!\!\rightarrow} O$ differs in probability from $B \mathrel{\Box\!\!\rightarrow} O$ for some outcomes O. This will be difficult to arrange on standard, Lewis-Stalnaker type, theories of counterfactuals.

The reason is that X's contribution on standard theories to the proposition expressed by $X \mathrel{\Box\!\!\rightarrow} Y$ in a world is the set $f(w, X)$ of near-enough worlds where X is true. This set does not change if we substitute an X' intensionally equivalent to X. But then $A \mathrel{\Box\!\!\rightarrow} O$ comes out expressing the same set-of-worlds proposition as $B \mathrel{\Box\!\!\rightarrow} O$. Which makes it hard to see how the two counterfactuals can differ in probability or hence how U(A) can exceed U(B). A's presumed superiority over B comes into conflict with the fact, if it is one, that U(A) $\not>$ U(B).

[11] Assuming $X{\vee}Y \ncong Y$.

[12] Assuming $X{\vee}Y \ncong X$.

I say "if it is one" because the case just made for U(A) $\not>$ U(B) is not air-tight, for reasons we now begin looking into. A and B obtain, recall, in highly specific ways, indicated above with αs and βs. Suppose that the apples are $\mathfrak{a}_1$, $\mathfrak{a}_2$, etc., with $\mathfrak{badapple}$ coming first ($\mathfrak{b} = \mathfrak{a}_1$). Then A's ways of holding (the $\beta_\star$s and $\beta^\star$s) are representable as assignments to each natural number n of ✓ (if $\mathfrak{a}_n$ is to be eaten) or ✗ (if $\mathfrak{a}_n$ is not to be eaten). The αs come out looking like this:

	$\mathfrak{a}_1$	$\mathfrak{a}_2$	$\mathfrak{a}_3$	$\mathfrak{a}_4$	$\mathfrak{a}_5$	....
$\beta^1_\star$	✗	✓	✓	✓	....	
$\beta^\star_1$	✓	✓	✓	✓	....	
$\beta^2_\star$	✗	✗	✓	✓	....	
$\beta^\star_2$	✓	✗	✓	✓	....	
$\beta^3_\star$	✗	✓	✗	✓	....	
$\beta^\star_3$	✓	✓	✗	✓	....	
$\beta^4_\star$	✗	✗	✗	✓	....	
$\beta^\star_4$	✓	✗	✗	✓	....	
....	...	...	...	...	....	
$\beta^n_\star$	✗	...	...	...	....	
$\beta^\star_n$	✓	...	...	...	....	
....	...	...	...	...	....	

Table 3: Ways for A to hold

The βs resemble the $\beta_\star$s, except that $\mathfrak{b}$ ($= \mathfrak{a}_1$) is out of the picture; it is not a fruit that even *can* figure in ways of eating almost all apples $\neq \mathfrak{b}$:

	$\mathfrak{a}_1$	$\mathfrak{a}_2$	$\mathfrak{a}_3$	$\mathfrak{a}_4$	$\mathfrak{a}_5$	....
β(*1*)	—	✓	✓	✓	✓	...
β(*2*)	—	✗	✓	✓	✓	....
β(*3*)	—	✓	✗	✓	✓	...
β(*4*)	—	✗	✗	✓	✓	...
β(*5*)	—	✓	✓	✗	✓	...
β(*6*)	—	✗	✓	✗	✓	...
β(*7*)	—	✓	✗	✗	✓	...
β(*8*)	—	✗	✗	✗	✓	...
....	—	...	...	...	....	....
$\beta(n)$	—	...	...	...	...	....
....	—	...	...	...	...	....

Table 4: Ways for B to hold

That A and B have such varied possible implementations suggests that they are "underlyingly" disjunctive, even if not disjunctive on the surface. This is important since disjunctiveness turns out to complicate matters when it comes to decision-theoretic calculations. What is pr(A $\square\!\!\rightarrow$ O) supposed even to *mean*, when different αs have different probabilities of producing a given outcome?[13]

[13] The title of the next section is a play on Kratzer's "How Specific is a Fact?" which strikes similar themes (Kratzer [1990]). See also Kratzer [1989] and Kratzer [2002].

11 How Specific is an Act?

The probability of $\alpha \mathrel{\Box\!\!\rightarrow} O$ is likely to differ from that of $\alpha' \mathrel{\Box\!\!\rightarrow} O$. This is one of the reasons that Lewis elects, in his own version of causal decision theory, to make acts/options A as specific as possible — subject to a certain condition, that which option is realized is under the agent's control.

> Suppose we have a partition of propositions that distinguish worlds where the agent acts differently .... Further, he can act at will so as to make any one of these propositions hold; but he cannot act at will so as to make any proposition hold that implies but is not implied by (is properly included in) a proposition in the partition. *The partition gives the most detailed specifications of his present action over which he has control.* Then this is the partition of the agents' alternative options. Say that the agent realises an option iff he acts in such a way as to make it hold. Then the business of decision theory is to say which of the agent's alternative options it would be rational for him to realise. (Lewis [1981], 7)

He adds in a footnote that "They are his narrowest options. Any proposition implied by one of them might be called an option for him in a broader sense, since he could act at will so as to make it hold. But when I speak of options, I shall always mean the narrowest options." Reserving v for narrowest — henceforth simply "narrow" — options, X is an option in Lewis's broad sense (it's a broad$_L$ option) iff it is implied by some v.

Broad options will have to be accommodated somehow. I am not normally deciding between options so specific as to exhaust what I have control over. But Lewis's sense is too broad to be useful. An agent can act so as to make a broad$_L$ option X hold, but not normally so as to make X fail. $X =$ *Someone blinks* is a broad$_L$ option, since it's implied by *I blink*, and hence by any narrow option v that has me blinking. But decision theory should not be offering me advice about whether to see to it that someone blinks as opposed to no one blinking Which of these obtains is not up to me; for I cannot bring it about that no one blinks.

Is there no way to restrict the term "broad option" to Xs such that it is under my control whether or not X obtains? There is. Call X broad if it is a *disjunction* of narrow options v, e.g., X might be $v_1 \vee v_2$. Whether $v_1 \vee v_2$ holds is under my control for I can bring it about that it does hold, and that I can bring it about that it doesn't. I do the first by v_1-ing, or else by v_2-ing. I do the second by declining both to v_1 and to v_2.

This does not get us all the way out of the woods, however. Yes, broad options construed as disjunctions of vs are within my power to enact or not, as I choose. But how is that choice supposed to be made? You might think it should be on the basis of the options' expected utilities. But to calculate the expected utility of a disjunction $v_1 \vee v_2$, we will need to know for each outcome O the probability of $((v_1 \vee v_2) \mathrel{\Box\!\!\rightarrow} O$. And there is not much agreement these days about how to evaluate, much less calculate the probabilities of, counterfactuals with disjunctive antecedents. Fans of *Simplification*

$$\text{(SDA)} \quad (P \vee Q) \mathrel{\Box\!\!\rightarrow} R \vDash (P \mathrel{\Box\!\!\rightarrow} R) \wedge (Q \mathrel{\Box\!\!\rightarrow} R) \tag{13}$$

would seem committed to

$$\text{(PDA)} \quad \mathrm{pr}((v_1 \vee v_2) \mathrel{\Box\!\!\rightarrow} O) \leq \min(\mathrm{pr}(v_1 \mathrel{\Box\!\!\rightarrow} O), \mathrm{pr}(v_2 \mathrel{\Box\!\!\rightarrow} O)) \tag{14}$$

To go by (PDA), $\mathrm{pr}((v_1 \vee v_2) \mathrel{\Box\!\!\rightarrow} O)$ ought to be zero when v_1's potential outcomes are different enough from v_2's. Suppose for instance that $\mathrm{pr}(v_1 \mathrel{\Box\!\!\rightarrow} O_1) = 1$ and $\mathrm{pr}(v_2 \mathrel{\Box\!\!\rightarrow} O_2) = 1$, where O_1 and O_2 are incompatible. Then $\mathrm{pr}((v_1 \vee v_2) \mathrel{\Box\!\!\rightarrow} O_1)$ is zero (since it is bounded above by $\mathrm{pr}(v_2 \mathrel{\Box\!\!\rightarrow} O_1) = 0$) and $\mathrm{pr}((v_1 \vee v_2) \mathrel{\Box\!\!\rightarrow} O_2)$ is 0 (since it is bounded above by $\mathrm{pr}(v_1 \mathrel{\Box\!\!\rightarrow} O_2) = 0$).

Perhaps then we should consider v_1 and v_2's possible outcomes separately, factoring in only later the probability of $(v_1 \vee v_2)$-ing the one way as opposed to the other. The utility of $v_1 \vee v_2$ on this view is the average value of v_1's possible outcomes and v_2's, weighted by (i) the probability the agent will $v_1 \vee v_2$ by v_i-ing, and (ii) the probability that v_i-ing (v_2-ing) will result in O:

$$\mathrm{U}(v_1 \vee v_2) \;=\; \Sigma_{i=1,2} \quad \mathrm{pr}(v_i \mid v_1 \vee v_2) \quad \times \quad \Sigma_O\, (\mathrm{pr}(v_i \,\Box\!\!\rightarrow O) \times \mathrm{V}(O)). \tag{15}$$

Assuming as before that Σ_O $\mathrm{pr}(v_i \,\Box\!\!\rightarrow O) \times \mathrm{V}(O)$ is $\mathrm{U}(v_i)$, this comes close to a formula proposed long ago by Jeffrey:[14]

$$\mathrm{U}(v_1 \vee v_2) = \quad \Sigma_{i=1,2} \quad \mathrm{pr}(v_i \mid v_1 \vee v_2) \times \mathrm{U}(v_i). \tag{16}$$

.

But now, how will the agent divide her credence (conditional on their disjunction) between v_1 and v_2? Suppose to keep things simple that she expects herself to go for *whichever of the two is rationally preferable*, that is, whichever has the higher expected utility. This is fair enough but it leaves broad options with no real work to do. Even when officially the choice is between broad options, say, $v_1 \vee v_2$ and $v_3 \vee v_4$, the decision process bottoms out at narrow options v_i, Why not follow Lewis in that case and run our calculations on v_1, v_2, v_3, and v_4 right from the start?

This does not do away with disjunctive options, exactly, but it does leave them very much on the sidelines; they come out exactly as valuable as their best disjunct. Deciding between $v_1 \vee v_2$ and $v_3 \vee v_4$ turns out to be "really" a matter of deciding among the four v_is, then checking which of $v_1 \vee v_2$, $v_3 \vee v_4$ has the winning v_i as a disjunct. In effect the decision-maker is portrayed as adopting a maximax rule. The question from her perspective is, what's the best that could happen. Eating cake is no better an option than eating cake or poison, given that the agent exercises the latter option by eating cake. That is not how we ordinarily think of it and not how the decision theorist *should* think of it. A theory that does not recommend cake over cake-or-poison has no business advertising itself as a theory of rational choice.

12 Interventionism

Another, more recent, approach to the problem of disjunctive options gets its probabilities from a Pearl-style system of structural equations. One starts with a "causal model" $\mathscr{M}$ built on variables X, Y, Z, etc. laid out in a directed graph. Z's value is fixed by its parents' values according to the model's governing equations, e.g., $Z := X \vee Y$. (The values of X and Y derive analogously from their own parents' values, until we reach "exogenous" variables whose values are given outright.) To evaluate, say, $(\neg X \wedge \neg Y) \,\Box\!\!\rightarrow \neg Z$, we plug 0 in for X and for Y (thus verifying the antecedent) and check whether the the resulting value for $\neg Z$ is 1. Indeed it is $(Z = (0 \vee 0) = 0)$. So the counterfactual is reckoned true.

This works well for counterfactuals — like $(\neg X \wedge \neg Y) \,\Box\!\!\rightarrow \neg Z$ — with *conjunctive* antecedents. But to calculate expected utilities for options like *eating cofinitely many apples*, we will need to assign probabilities to counterfactuals with *disjunctive* antecedents. Ray Briggs in an important paper observes that the interventionist has no obvious way even of evaluating such counterfactuals:

> The most striking difference between the similarity and causal modeling accounts is that the similarity account does, and existing causal modeling accounts do not, allow us to assign truth values to *arbitrary* counterfactuals. Galles and Pearl consider a language in which the only expressible counterfactuals have the form
>
> $P_1 \wedge ... \wedge P_n \,\Box\!\!\rightarrow Q_1 \wedge .... \wedge Q_m$,
>
> where P_1,...., P_n and Q_1,...., Q_m are atomic (Briggs [2012], 147).[15]

[14] Albeit Jeffrey was talking, as an evidential decision theorist, about expected *value* rather than expected utility.

[15] Think of the P_is as assigning numerical values to n distinct variables, while the Q_js assign values to m variables causally downstream from these.

How are we to extend the theory to disjunctive counterfactuals like $(\neg X \vee \neg Y) \mathrel{\Box\!\!\rightarrow} \neg Z$ —- equivalently to $(P_1 \vee P_2) \mathrel{\Box\!\!\rightarrow} Q_3$, where the P_is assign 0 to X and Y, and Q_3 assigns 0 to Z? (E.g., we might in the two-button case want to evaluate *If the left-hand button hadn't been pressed, or the right-hand button hadn't been pressed, the light would have stayed off.*)

Suppose following Briggs that we let the world role by played by saturated causal models — made up of graphs, equations, and assignments of numbers to variables. If both buttons were pressed in fact (L=R=1), then the closest world where Lefty isn't pressed is the tweaked model where L is set to 0, R is left at 1, and other values are assigned according to the equations; so $\neg L \mathrel{\Box\!\!\rightarrow} \$0$.[16] What then about $(\neg L \vee \neg R) \mathrel{\Box\!\!\rightarrow} \$0$?

The suggestion is to think about this, like Stalnaker, in selection-function terms: $f(w, \varphi)$ contains the φ-worlds that have got to be ψ, for $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ to hold in w. The question is, what should $f(w, \varphi_1 \vee \varphi_2)$ be? If we want $(\neg L \vee \neg R) \mathrel{\Box\!\!\rightarrow} \neg\$1{,}000{,}000$ to mean, as it should, that we get \$1,000,000 only if both buttons are pressed — much less if any buttons are left unpressed — then $f(w, \varphi_1 \vee \varphi_2)$ should be $f(w, \varphi_1) \cup f(w, \varphi_2)$.

Briggs' union rule has the result that the worlds that have to be $\$0$ for $(\neg L \vee \neg R) \mathrel{\Box\!\!\rightarrow} \$0$ to hold are the ones that have to be $\$0$ for $\neg R \mathrel{\Box\!\!\rightarrow} \$0$ to hold, *and* the ones that have to be worthless for $\neg R \mathrel{\Box\!\!\rightarrow} \neg\$1{,}000{,}000$ to hold. $(\neg L \vee \neg R) \mathrel{\Box\!\!\rightarrow} \$.01$ winds up entailing $(\neg L \wedge R) \mathrel{\Box\!\!\rightarrow} \$.01$, $(L \wedge \neg R) \mathrel{\Box\!\!\rightarrow} \$.01$, and $(\neg L \wedge \neg R) \mathrel{\Box\!\!\rightarrow} \$.01$, since $f(w, \neg L \vee \neg R) = f(w, \neg L) \cup f(w, \neg R)$, subsumes each of $f(w, \neg L)$, $f(w, \neg R)$, and $f(w, \neg L \wedge \neg R)$ $(= f(w, \neg L) \cap f(w, \neg R))$.[17]

Suppose quite generally that $f(w, X)$ is obtained by unioning $f(w, \chi)$ as χ ranges over X's possible truthmakers, what we have called its ways of being true. Since logical equivalents can differ in their truthmakers (they need not be hyper-equivalent), they will not always be substitutable in the antecedents of counterfactuals. Briggs puts it like this:

> The concept of truthmaking generates a sense of equivalence which is narrower than the ordinary classical one, and which I will call exact equivalence, following Fine. Two sentences are exact equivalents [hyper-equivalents, in our terminology] iff they have the same truthmakers. Not every pair of classically equivalent sentences is exactly equivalent; for instance, $(\neg X \wedge Y) \vee (\neg X \wedge \neg Y))$ is classically, but not exactly, equivalent to X. Although exact equivalents can be freely substituted in the antecedents of counterfactuals, classical equivalents cannot. (155)

Take again the two switch case, with the twist that we win \$1,000,000, if *either* button is pushed. If the first hadn't been pushed ($\neg L$), we would still have got rich. But it's false that we would still have got rich if: either the first hadn't been pushed or neither had been pushed $\neg L \vee (\neg L \wedge \neg R))$, since the light would have been off it neither had been pushed. And yet $\neg L \vee (\neg L \wedge \neg R))$ holds in the same worlds as $\neg L$.

The upshot is that it is not true, on intervenist theories of counterfactuals, that "X's contribution to the proposition expressed by $X \mathrel{\Box\!\!\rightarrow} Y$ in a world is just the set of nearby worlds where X is true." Of course that set of worlds *is* "the same for any X' intensionally equivalent to X." But antecedents are often thought nowadays to contribute something more fine-grained than a single set of worlds—a set of truthmakers, perhaps, or a set of alternatives, or a set of sets of worlds.[18]

For a couple of reasons, then, the argument in section **10** from the "intensionality of counterfactuals" to the intensionality of utility does not succeed in showing that $A \not\succ B$. The first is that counterfactuals are often thought nowadays to be hyperintensional. And even if they are intensional, expected utility may not be. The utility of a disjunction of acts derives, one may think, from the

[16] Think of $\$0$ here as covering the case where you get a penny and the case where you get nothing.

[17] This may or may not be the result we ultimately want. For further subtleties see Ciardelli et al. [2018], Lassiter [2017], and Lassiter [2018].

[18] There is a large literature on these matters and I'm no expert. See among others Fine [2012], Santorio [2018], Khoo [2018], and Santorio [2021]. Briggs' theory is akin in some ways to what Fine proposes. For Pearl's own take on disjunctive actions, see Pearl [2017].

utilities of the acts taken one by one. And intensional equivalents are liable to differ in their disjuncts (e.g., $XY \vee X\bar{Y}$ unlike X has XY as a disjunct, and unlike X can be done by XY-ing).

13 By-Ways

How do the worldly and wayward approaches to decision-making relate, and how do we decide between them? A good place to start is with Jeffrey, because although a world-theorist he falls at times into the language of ways.

In section **8** we attributed to Jeffrey the idea that an unspecific option R's value goes with the average value of its realizations R_i, weighted by their probabilities conditional on R. The *Stanford Encyclopedia of Philosophy* explains it like this:

> the desirability of a proposition, including one representing acts, depends both on the desirabilities of the different ways in which the proposition can be true, and the relative probability that it is true in these respective ways... Let $\{R_1, R_2, ..., R_n\}$ be one amongst many finite partitions of the proposition R; that is, sets of mutually incompatible but jointly exhaustive ways in which the proposition R can be realised... The desirability of R according to Jeffrey, denoted *Des(R)*, is given by:
>
> **Jeffreys' Equation** $$Des(R) = \Sigma_i\ Des(R_i) \times \mathrm{pr}(R_i/R) \tag{17}$$

A question may have occurred to you. Doesn't it make Jeffreys' Equation circular that to determine P's desirability, one must already know the desirability of its realizations P_i? Of course one can try to apply it again to the P_i realizations P_{i_j}, but this gets us no further ahead since the problem arises again for them.

How much this should bother us depends on whether the problem keeps on arising forever. Jeffrey would say that it doesn't. The process bottoms out at P's *maximally specific* realizations: the worlds W that verify P. Worlds are not multiply realizable, making them the perfect place for the P, P_i, P_{i_j}, $P_{i_{j_k}}$,...sequence to come to a halt.

> It is only the complete consistent novels that can be said to have nonprobabilistic values: the desirability of a proposition will be a probability-weighted average of the values of the possible worlds in which it would be true (Jeffrey [1990], 209).

Assuming the number of worlds is finite (otherwise we have to integrate), Jeffreys' equation boils down to something totally well-defined:

$$Des(R) = \Sigma_W\ Des(W) \times \mathrm{pr}(W/R) \tag{18}$$

But, if worlds for Jeffrey are ways (the *ultimate* ways, no less), then his relation to (16) — $\mathrm{U}(\mu_1 \vee \mu_2) = \Sigma_{i=1,2}\quad \mathrm{pr}(\mu_i \mid \mu_1 \vee \mu_2) \times \mathrm{U}(\mu_i)$ — needs to be reconsidered. The idea in (10) is not that R's utility can be obtained by subdividing the R-worlds however you like.[19] It's that we can obtain the utility of a a disjunction of narrow options by subdividing its worlds into the ones verifying the first disjunct, the ones verifying the second, and so on.

This is all somewhat obscured by the fact that Jeffrey speaks as we do of "ways for R to be true." What he has in mind by ways for R to hold is importantly different from what we've meant by the phrase in this paper. R's ways of being true for us are R's potential *truthmakers* — the facts *by virtue of* which R is, or could be, true. These facts correspond in the case of option-statements like *Bina*

[19] This way lies "partition invariance."

eats an apple to the more specific acts *by* doing which Bina could exercise the option, e.g., eating *this* apple in particular. The FIRST LAW OF WAYS, as we're conceiving them here, is that

(FLW) ρ counts as a way for R to hold only if one R's by ρ-ing —
or, in the case of non-action sentences, the *world* R's (verifies R) by ρ-ing. (19)

Let's call this, Ii you can forgive the cuteness, the BY-LAW. Ways of the type it governs will be BY-WAYS, though often we'll just say WAYS.

How do WAYS related to the ways (henceforth `ways`) that Jeffrey is interested in? R's `ways` of being true are the members R_i of any old partition of the set of R-worlds — any old set of propositions that are mutually incompatible and jointly exhaustive of those worlds. It is clear from the *Stanford Encyclopedia* piece quoted above that R_i need have nothing to do with how or why R obtains. The passage continues

> if P is the proposition that it is raining, then we could partition this proposition ... according to whether we go to the beach or not.

As we ordinarily think of it, *It's raining* is not true for different reasons, or in virtue of different facts, according to whether we go (or not) to the beach. Raining-and-beach-going is a `way` P_i for it to rain, relative to the right partition, but it is not a way π_i for it to rain. Going in the other direction, it makes no difference to why *We went to the beach* is true that that it rains (or doesn't) when we get there. Beach-going-followed-by-rain is a `way` of going to the beach, but it is not a way of going, any more than being rained on while dancing is a way of dancing.

Where does this leave us? The crucial point about `ways` is that they lump an act's *concomitants* (including, notably, its aftermath) in with how the act gets performed in the first place. Jeffrey is slightly confusing on this, since a lot of his examples involve WAYS:

> The agent is trying to comfort a lady whose cat has been killed. This may consist in any of a variety of acts, such as giving her another cat, holding her hand, or saying "He was getting old and stiff, anyway." And there are many ways of performing the last-mentioned act, of saying the words, some of which would be more likely to produce comfort than others: variations in volume of voice, proximity of speaker to hearer, and facial expression might all be important. (Jeffrey [1990], 178)

But although WAYS do motivational work for Jeffrey, they have no special role to play in the theory. R_i still counts as a way for R to hold, even when it is obtained from R by tacking on as a conjunct a randomly chosen other proposition S. "When a proposition can come true in more than one way [`way`!], it is, in effect, a gamble..." (87). A gamble on what? One is gambling, in picking R over $\neg R$, that cases of R—propositions partitioning the R-worlds — are on (weighted) average better than cases of $\neg R$—propositions partitioning the $\neg R$-worlds.

> any analysis into mutually exclusive, collectively exhaustive cases can be refined by taking a further proposition into account. Relative to a single proposition R, the cases are R, $\bar{R}$. But nothing prevents us from making use of the fact that each of these cases can in turn be split into two: R becomes two cases,
>
> RS, $R\bar{S}$
>
> and $\bar{R}$ becomes two cases,
>
> $\bar{R}S$, $\bar{R}\bar{S}$.
>
> And each of these four cases can in turn be split into two by considering a third proposition, T, etc. Then the distinction between propositions that can be true in only one way and [the rest] is relative to the number of cases that we see fit to consider. (Jeffrey [1990], 87)

An agent does not bring it about that R by bringing it about that RS, or bringing it about that $R\bar{S}$. So Jeffrey can only be talking here about `ways`.

14 Finesse

Not every distinction worth remarking on is worth jumping up and down about. Some distinctions, real though they may be, are better finessed. Wouldn't it be simpler, other things equal, to run all `ways` together, as Jeffrey does, lumping WAYS of R-ing in with `ways`, things like R-ing within five miles of a burning barn or R-ing followed by applause?

Lewis evidently thought that other things weren't equal. He wanted to limit ways of doing a thing to the finest-grained options μ such that it's up to the agent which is enacted. He did not disagree with Jeffrey about μ being itself a gamble — better or worse according to how likely it is to lead, in the circumstances, to good results. The question is what is allowed to go into the possible "circumstances" over which one distributes credences en route to deciding how to act. Lewis wants to limit them to facts outside the agent's control. Jeffrey models decision-makers as having views also facts within their control, e.g., what they are about to do.

Some object to this on the basis that overconfidence about what I am going to do can undermine the calculations that recommend it as the best thing to do.[20] Some like Levi maintain that "deliberation crowds out prediction."[21] Some worry that deliberation gets itself tangled up in feedback loops if the choice between μ_i and μ_j depends on a prior estimate of the likelihood of choosing μ_i (μ_j).[22]

All of these are, or could be, reasons for setting WAYS — how one accomplishes a thing — apart from `ways` — states of the world might wherein the thing is done. Feedback loops in particular come in for a lot of discussion in recent years, by people who (unlike me) have ideas about them.[23] I want consider a different reason, closer to our present concerns.

Among the `ways` of B-ing — of eating all but finitely many apples $\neq \mathfrak{b}$ — are some, like $\alpha_\forall =$ eating all the apples, where I eat $\mathfrak{b}$ as well. ω is NOT, by contrast, a WAY of B-ing. This again is because eating $\mathfrak{b}$ plays no part in its coming about that you B, that is, eat almost all apples distinct from $\mathfrak{b}$. Of course eating $\mathfrak{b}$ is *compatible* with B-ing; but that is a far cry from being done potentially in the *act* of B-ing.

What was said here of $\alpha_\forall$ applies as well to all the αs that get you rich. They are compatible with B-ing, but they are not the sort of thing that B-ers do as such. They are not done in the act of B-ing, but only that of A-ing. Now, which of these sounds like a stronger point in favor of X-ing — that it doesn't *rule out* getting rich, or that you can get rich precisely by X-ing, i.e., by doing what you do to X? The second, surely. That X doesn't preclude good outcomes hardly seems like a recommendation at all, unless the competition does preclude them. I suppose some kind of homage is paid to blinking, when you remark that it's over so quickly that there is still time to pursue one's dreams. But it's the damning with faint praise kind. There is nothing faint about the praise it pays X-ing to say that it can be done, *at one's discretion*, in a dream-fulfilling way.

Finessing the `way`/WAY distinction costs us something option-assessment-wise. It would be one thing if good `ways` of doing a thing reflected as well on it as good *ways* of doing it (good ξs by which one can X). But this does not seem to be the case. Certainly it is not the case where other normative notions are involved. Lying is not made more permissible by the fact that there are worlds where you lie harmlessly, say, because no one hears you. It does not say much for the rationality of believing the first thing anyone tells you that there are good `ways` of doing it, e.g., the `way` it's done in worlds where the first thing you're told is deeply, profoundly true.

But let's return to a case from the last paragraph. Blinking is not made much advisable, we said, by the fact you can eat all the apples (earning thereby \$1,000,000) after you're done. Why should it

[20] See Jeffrey [1990], ch11, and Jeffrey [1977].

[21] Joyce [2002]

[22] Skyrms [1990].

[23] Egan [2007], Arntzenius [2008], Joyce [2012], Spencer [2021].

make B-ing — eating almost all apples $\neq \mathfrak{b}$ — advisable that you can eat $\mathfrak{b}$ when you're done (earning thereby the same amount)?

Here is why, you might say: to win $1,000,000, you must *at least* B. B-ing is *necessary*, albeit not sufficient, for winning the million. And indeed every lucrative way of A-ing has a way of B-ing as a part. But again, the choice we face is not between B-ing and blinking, it's between B-ing and A-ing. That each β gets us *part* of the way to the jackpot does favor B over burping. But does it boost B's standing vis-a-vis A-ing? That it gets us part of the way to $1,000,000, even an essential part of the way, is a plus for B. But it can hardly bring B up to A's level; for A-ing gets us (can get us) *all* the way to $1,000,000. Alternatively we can put it in dominance terms Every other way of A-ing makes you rich. NO ways of B-ing make you rich. This seems a strong point in favor of A-ing.

Decision theory ranks options on the score of rational preferability, and advises us to act in accord with the ones that are highest-ranked. What is it to act "in accord with" X? Orthodox decision theory, when it recommends X, is steering us towards acts *consistent* with X. X is preferable to Y if acts consistent with X lead on balance to better results than acts consistent with Y. The approach explored here ranks options based on how they are liable be *exercised*. X is preferable to Y if to X leads on balance, averaging over the ways ξ of X-ing, to better results than Y-ing does. The point of recommending X over Y is to steer us, not towards acts consistent with X, but towards ways of X-ing — towards the acts by performing which one Xs. I take it that the second approach is more in line with ordinary notions of advice and advisability.

To see why, consider again the two button example, the version where both buttons must be pressed to turn the light on and win the big reward. Neither button is down to begin with. Among our options are (i) pressing the left button (L), and (ii) either pressing both buttons, or pressing just the one on the left ($LR \vee L\bar{R}$). (ii) may strike us as the better option, since to LR (the jackpot-winner) is a way of ($LR \vee L\bar{R}$)-ing, but not of L-ing. LR is not precluded by L; it is even partly accomplished by L-ing (albeit no more than $L\bar{R}$-ing is partly accomplished by L-ing). (ii) is better than (i), on the hyperintensional approach, for just this reason.

How do the options compare from an intensionalist perspective? If not precluding renumerative behavior is the best an option can hope for, then fact is that L achieves this just as well as $LR \vee L\bar{R}$. $L \vee \bar{L}$ and $L \vee \bar{R}$ do not for that matter preclude getting rich either. They are just as good, then, from a Jeffreyian perspective as LR, assuming as seems reasonable that the agent's `way` of $L \vee \bar{R}$-ing is to LR.

"Choose this, as it will (done right) win you $1,000,000" is by ordinary standards better advice than "Choose that, as it doesn't stop you from winning $1,000,000." Hyperintensionalism gives the first, this-done-right-will-make-you-rich sort of advice. Intensionalism's advice takes the second form: this does not definitely make you poor. Insofar as the first sort of advice is more useful, hyperintensionalism does a better job of pinning down what is (was) to be done than its traditional rival intensionalism.

15 Intensional about Evidence

All kinds of propositions can stand in preference relations and be assessed for desirability. The ones that function in Jeffrey as options — those to the effect that an agent does, or tries to do, thus and such — are special only from the perspective of the agent, when she is deciding whether to do thus and such, and/or the rest of us when we are evaluating her decision. Propositions play a doxastic role too, obviously, as objects of belief and credence. And they function epistemically as pieces of evidence, and hypotheses potentially confirmed by that evidence. If intensional equivalents can differ as *options*, then it stands to reason that they might differ too as pieces of *evidence*, or as *conjectures* that a given body of evidence may (or may not) support.

An infinitary example will be given in a minute. But let's first consider something simpler. Suppose that P and Q are evidentially quite independent; neither supports the other. To learn that P — the penny came up heads – - would not change your views about Q — the quarter came up heads —

in the slightest. But, what if you learned instead that $P \vee Q$ (one coin or the other came up heads), or, to make the alternative evidence intensionally equivalent to P, that either the penny came up heads or it and t he quarter both came up heads?

This might well change your attitude about the quarter. There is a decent chance after all that $P \vee PQ$ holds thanks to the fact that PQ, hence thanks in part to the fact of the quarter coming up heads. Granted, the news that $P \vee PQ$ may be uninformative about the quarter. Imagine our informant Smyth has evidence only for P; the second disjunct (PQ) is tacked on for no good reason later. Smyth would resemble in that case Gettier's Smith, who starting out with the justified false believe that *Jones owns a Ford*, arrives at *Jones owns a Ford or Brown is in Barcelona* by "select[ing] three place-names quite at random,"

> construct[ing] the following three propositions:
>
> - Either Jones owns a Ford, or Brown is in Boston;
> - Either Jones owns a Ford, or Brown is in Barcelona;
> - Either Jones owns a Ford, or Brown is in Brest-Litovsk.

and adopting all three as further beliefs. Smyth's testimony carries no information about the quarter if she observed only the penny and arrived at *The penny came up heads and the quarter did* by selecting three coins quite at random and adopting as further beliefs

- Either the penny came up heads or the penny and the nickel did;
- Either the penny came up heads or the penny and the quarter did;
- Either the penny came up heads or the penny and the dime did

But the bare possibility of $P \vee PQ$ carrying no information about Q doesn't count for much. To ignore the second disjunct, one would have to be certain of it. Perhaps Smyth's own informant testified either that the penny came up heads, or that it and the quarter both did (the bar was noisy and Smyth couldn't be sure). Smyth *might* have heard that the penny and quarter both came up heads and that seems like enough for evidential relevance vis a vis the quarter.[24]

16 Consistency and Confirmation

You may object that E's evidential value depends only on what it's consistent with, e.g., Bayesians think it's a function of which worlds E is consistent with. This is the same obviously for P and $P \vee PQ$. People do often *say*, "it's consistent with our evidence that the dog was hungry," as if this meant our evidence *supported* the hunger hypothesis.

But "it's consistent with our evidence that the dog was hungry" is at the same time a *strange* thing to say. E does not make much of a case for H, if the best one can claim for it is consistency with E. Imagine our evidence consists of facts about the Battle of Hastings, e.g., that it occurred in 1066. This is consistent with the hungry-dog hypothesis, but it hardly supports that hypothesis.

Perhaps what people mean when they say "it's consistent with our evidence that H" is that E is *more* consistent with H than with $\neg H$ ($\mathrm{pr}(H|E) > \mathrm{pr}(\neg H|E)$. This does not offer H much support either though, unless a serious and continuing effort is being made to gather H-relevant evidence. (As often it has, when people say this kind of thing.)

The claim might be that H *remains* more consistent with our evidence than $\neg H$ — or becomes *increasingly* consistent with E relative to $\neg H$ — the more H-relevant evidence we gather. The stronger E gets, the more the EH-worlds outnumber/outweigh $E\bar{H}$-worlds.

[24] There are issues here about conditioning on *all* one learns, including facts about the causal provenance of one's evidence conceived more narrowly. More on this later hopefully.

This is not going to move us all by itself, though, for a body of evidence can be expanded along prejudicial lines. Who is to say that the speaker has not adopted a policy of beating about in H-friendly bushes, to the exclusion neighboring bushes where H-disfavoring evidence might lurk?[25] To address this one needs presumably to consider counterfactuals of the form

- If H, evidence for it would be found in these bushes
- If $\neg H$, evidence for it would be found in those bushes
- If E, we would encounter the fact that E in H-favoring bushes.
- If E obtained, we would encounter the fact that E in H-disfavoring bushes.

But these counterfactuals are arguably themselves hyperintensional, for reasons of a sort already seen. *If Hamlet's author was a literary genius, it would say so in a book arguing that the same man wrote Hamlet and Novum Organum* seems not terribly probable — much less probable than *If Hamlet's author was a literary genius, e.g., the literary genius Francis Bacon, it would say so in a book arguing that the same man wrote Hamlet and Novum Organum.*

A similar dynamic is at work with the two hypotheses in APPLES: $A =$ *Eve has eaten almost all of the apples*, $B =$ *She has eaten almost all of the apples distinct from* $\mathfrak{b}$. Imagine that someone walks in with the news E that Eve has eaten $\mathfrak{badapple}$. We know that Eve decides, for each apple $\mathfrak{a}_i$, whether to eat it by flipping a fair coin — a different coin for each apple. This being so it tells us nothing about which apples she's eaten apart from $\mathfrak{b}$ that Eve has eaten $\mathfrak{b}$. B is not confirmed in the slightest by evidence E.

This is just what we'd expect of course. What does seem a bit surprising is a result that follows from E not confirming B, if confirmation is intensional on the right hand side. A and B are cointensional; so, if confirmation is intensional on the right hand side, E does not count in favor of A either. Eve's eating $\mathfrak{b}$ smiles no more on her eating a lot of apples *of which* $\mathfrak{b}$ *is one*, than on her eating a lot of $\mathfrak{b}$-distinct apples.

But now, how could it not be good news for the hypothesis that most φs are ψs that *this* φ is ψ? Would it not have been worse news for the hypothesis, had it turned out not to be ψ? When E is better news for a hypothesis than $\neg E$, we usually do not hesitate to call it *good* news for that hypothesis, granted that the goodness may be slight or unquantifiable.

"E is better news for A than $\neg E$ would be" is meant to express a *comparative* judgment, like the judgment that $R \succ S$. I am not suggesting that pr$(A|E)$ >pr$(B|\neg E)$; the conditional probabilities may be equal, or ill-defined. The suggestion is just that (in a hopefully obvious notation from qualitative/comparative probability theory), $A|E \succ A|\top \succ A|\neg E$ — notwithstanding the existence of an intensionally equivalent B such that $B|E \not\succ B|\top \not\succ B|\neg E$.[26]

17 Hyperintensionalism about Evidence

This is only intuition-pumping, you may say. What is the *argument* supposed to be? I am not sure I have one. But here are some relevant-seeming *considerations*.

Homogeneity Supplementing *Eve eats* $\mathfrak{b}$ with sufficiently many other facts of the same form (*Eve eats* $\mathfrak{a}_i$) does give us evidence for *Eve eats almost all the apples* (A). Indeed enough such facts entail *Eve ate almost all the apples*, given that the apples are the $\mathfrak{a}_i$s. But one fact of the form *Eve eats* $\mathfrak{a}_i$

[25] Salow [2018]

[26] On comparative probabilities, see Walley and Fine [1979], Fishburn [1986], Hawthorne et al. [2016], Ding et al. [2021]. (None of this literature embraces hyperintensionality to my knowledge.)

form carries no more weight than another! If *Eve eats* $\mathfrak{b}$ were not relevant, it is hard to see how it plus a bunch of no-more-relevant facts could provide decisive support for A.[27]

Analogy That Eve eats $\mathfrak{b}$ does count in favor of Eve's eating *all* the apples, if only in the dodged-a-bullet sense; the hypothesis is refuted if she *doesn't* eat $\mathfrak{b}$. It counts too obviously in favor of her eating some of the apples, or at least n of the apples, or within n of all the apples. How can E support these hypotheses but not A, though all are to the effect that there at least so many, or thus and such a proportion of, φs are ψs?

Polarity Logicians have the notion of an atom p occurring *positively* in complex formula φ. p occurs positively in $p \wedge q$, $p \vee q$, and $q \supset p$; negatively in $\neg p$ and $p \supset q$; and neither positively nor negatively in $p \equiv q$. The notion is defined syntactically, but it has in finitary languages a clearly semantic upshot: toggling p to true can make φ true but never false. This result does *not* for perhaps guessable reasons extend to infinitary languages.[28] But it still holds "in spirit"; the truth oof positively occurring subformulas of φ bears in a clear intuitive sense favorably on φ.[29] Now, *Eve eats* $\mathfrak{b}$ occurs positively in the obvious propositional rendering of *Eve eats almost all the apples*. E to that extent bears favorably on A.[30]

Overdetermination One way to establish *Eve eats* $\mathfrak{b}$'s relevance would be to show that it figured in a minimal truthmaker — a fact making A such that no weaker fact makes A. This will not be possible since A has no minimal truthmakers; one can always subtract, for each apple $\mathfrak{a}$, the sub-fact that $\mathfrak{a}$ in particular was eaten. How is relevance established when a sentence's truth is overdetermined via a chain of progressively weaker facts? The answer is that A's truth would not have been *as* overdetermined had Eve specifically not eaten $\mathfrak{b}$. The set of A's obtaining truthmakers, had Eve eschewed $\mathfrak{b}$ but otherwise eaten just what she did eat, is a proper subset of the actual set.[31]

Grounding The *proportion* of apples eaten does not grow when we move $\mathfrak{b}$ from the uneaten column to the eaten column. But the *set* of apples eaten does grow, and its relative complement shrinks. Proportions depend on relative cardinality. Cardinality depends in turn on set-membership; a set's size in the how-many sense (size$_{\#}$) is monotonically grounded in size in the membership sense (size$_{\epsilon}$). *Eve eats* $\mathfrak{b}$ (E) bears positively on *Eve eats almost all the apples* (A) because

> *Eve eats almost all the apples* says that a set exceeds its complement in size$_{\#}$.
> Size$_{\#}$ is monotonically grounded in size$_{\epsilon}$.
> *Eve eats* $\mathfrak{b}$ bears positively on the set's size$_{\epsilon}$, and negatively on its complement's size$_{\epsilon}$.

Surprise It comes as quite a surprise, when almost all φs be ψ, that a randomly chosen φ turns out not to be ψ. That the φ in question turned out to be ψ would not be surprising at all. How could it not be good news for a hypothesis H, that things turn out a way that is (on H) just what you would expect, rather than a way that would be (on H) surprising?

Explanation Confirmation has often been linked to explanation.[32] (Crucially the notion of explanation at work here is not — anyway not assumed to be — reducible to probability-raising. Correct

[27] How do the points composing a sphere contribute to its volume, when each point is of measure zero? Well, the points do the job collectively, and one is exactly as helpful as another. (This relates to the issue of what principled difference there could be between countable and uncountable additivity that would justify accepting one while rejecting the other) (Easwaran [2013]).

[28] p occurs positively in the obvious propositional rendering φ of *There are infinitely many true atoms*, but verifying p alone can never flip φ from false to true.

[29] In the case of φ ($\approx$ *There are infinitely many true atoms*) we might try to explain this as follows. φ is redistribution-equivalent to the conjunction over all n of *There are at least n atoms*. And p for each n occurs positively in *There are at least n atoms*.

[30] *Eve eats* $\mathfrak{b}$ does not occur at all in the obvious propositional rendering of B.

[31] Yablo [2022?]

[32] White [2005]

predictions are not per se explanatory, even if they do reduce surprise in some crude probabilistic sense.)

1. Insofar as *H* would help to explain *E*, *E* (if true) is evidence for *H*.
2. For almost all the apples to be eaten potentially helps to explain why 𝔟 was eaten.
3. For almost all the 𝔟-distinct apples to be eaten does nothing to explain why 𝔟 was eaten.
4. So *E* is evidence for *A* but not (on these grounds, anyway) for *B*.

Such an argument doesn't get off the ground unless explanation is hyperintensional; but the case for this has been made many times.[33] It is suggested as well by the hyperintensionality of counterfactuals on the not implausible hypothesis that explanation is counterfactual-involving.[34]

18 Total Evidence, Perspicuous Evidence

That completes our short list of "considerations" favoring, or seeming to favor, *E supports A* over *E supports B*. One might be curious too about the support if any that *A* offers to *E* — that *Eve eats almost all the apples* offers to *Eve eats* 𝔟. If evidential support is understood (as by Bayesians) in terms of conditional probability, then *A* will have to support *E* too — $\text{pr}(X|Y) > X$ iff $\text{pr}(Y|X) > Y$.

This is an intriguing possibility but also, for reasons already hinted at, a troubling one.

Suppose we are considering whether or not to accept *Eve eats* 𝔟 (*E*𝔟), when a trusted informant says: *Almost all apples are eaten* (*A*). *A* counts in favor of *E*𝔟, we're imagining, so learning it should in some sense dispose us towards *E*𝔟 — more anyway than learning *B* (𝔅𝔞𝔡𝔞𝔭𝔭𝔩𝔢 *aside, almost all apples are eaten*) would.

The problem is that *A* on anyone's account is a priori deducible from *B*! Who is to say that our informant wasn't told *B* by *her* informant, inferring *A* from it before passing the news on to us? It makes no sense in that case to raise our confidence in *E*𝔟, since *B* does not support *E*𝔟 in the slightest. Another way to think of it is that the *A*-testimony I receive can support *E*𝔟 only it carries information about 𝔟; and this is unlikely if my informant obtained *A* from *B*. Importantly the topic-changer could be myself. There is nothing to stop me from inferring *A* from *B* before drawing conclusions about 𝔟𝔞𝔡𝔞𝔭𝔭𝔩𝔢.[35]

The issue here is slightly reminiscent of one discussed by golden-age confirmation theorists like Carnap. Imagine I am looking for evidence supporting some some hypothesis *H*. Unfortunately all I have in my evidence book is a fact *E* that is thoroughly independent of *H*. But wait... *E* implies $E \vee H$, which does support *H*![36] I therefore decide to tell myself (truthfully) that $E \vee H$. With no additional investigation, I have gratified my desire for evidence in favor of *H*.

Where have I gone wrong? Carnap appeals in cases like this to a methodological "principle which seems generally recognized, although not always obeyed." He calls it the *Principle of Total Evidence*:

> if we wish to apply ... the theory of probability to a given knowledge situation, then we have to take as evidence *E* the total evidence available to the person in question at the time in question, that is to say, his total knowledge of the [situation] (Carnap [1947], 138)

[33] Sober [1982], Siebel [2011]

[34] How to reconcile IBE reasoning with the supposed primacy of Bayesian update is of course a huge question (Roche and Sober [2013], McCain and Poston [2017]).

[35] Or I could engage in a kind of aboutness-laundering, telling you that *B* in hopes that "the information" will make its way back to me as *A*.

[36] Bayesians will say this is because $\text{pr}(H|E \vee H) > \text{pr}(H)$. A hypothesis is always probabilified by its logical consequences, extremal values aside.

Not only does $E\vee H$ fall short of my full evidence, that evidence (E) screens $E\vee H$ off from H. $\text{pr}(H|E\vee H)$ exceeds $\text{pr}(H)$, but $\text{pr}(H|E\&(E\vee H)) = \text{pr}(H|E) = \text{pr}(H)$.

Totality is a matter of logical strength. My total evidence is all that I know, or all I have just learned, not something logically weaker than that. But there is in our $E\vee H$ example a subject matter dimension as well. E in itself is off-topic (or so we may suppose); it has topicality thrust upon it by the device of adding H as a disjunct. To bring out the contrast more clearly, let $E\&H$ be the new disjunct rather than H. We then get a statement $E\vee EH$ that is logically equivalent to E but has a more encompassing subject matter.

Does $E\vee EH$ differ evidentially from E? One certainly hopes not, if $E\vee EH$ was inferred from E. But if we learn directly that $E\vee EH$, matters are perhaps not so clear. Imagine E says that your coin came up heads, H that mine did. If I am told that either your coin came up heads, or both of our coins did, this *may* seem like good news for my coin coming up heads. One way for $E\vee EH$ to hold is via its second disjunct; and its second disjunct says in part that my coin came up heads.

Return now to B ($\mathfrak{b}$ *aside, almost all apples are eaten*). B considered in itself is off-topic when it comes to $E\mathfrak{b}$. But we can "make" it topical by adding a disjunct that is less indifferent to $\mathfrak{b}$. The new disjunct could be $E\mathfrak{b}$ itself; in that case the resulting disjunction is intensionally weaker than (properly implied by) B. But it could also be A (Almost all the apples are eaten), resulting in a disjunction $B\vee A$ that is intensionally equivalent to B.

Either approach yields a sentence with additional ways of being true, beyond B's ways, whereby $\mathfrak{b}$ was eaten. The suggestion is that this is *itself* a kind of weakening. Enlarging the field of events capable of witnessing a sentence's truth makes that truth easier to arrange. Whether the new verifiers bring in new worlds is a separate question — this occurs with $E\mathfrak{b}\vee B$, but not $A\vee B$, and not for that matter with A itself.[37] A spurious evidential link with $E\mathfrak{b}$ can be manufactured not only by *intensionally* weakening B (to $E\mathfrak{b}\vee B$), but also by *merely-topically* weakening it (to $A\vee B$, or to A itself (the two are hyper-equivalent).[38]

From this it seems we need a second principle, to the effect that one should not update on *topical* weakenings E^- of one's evidence E either—not even if E^- is intensionally just as strong as E. Another way to put it is that the evidence we update on should not have gratuitous, "extra" ways ω of being true that were thrown in for no good reason. A story will be needed though, about what a no-good reason looks like for giving ω a spot among our evidence-statement's truthmakers.

Here is tentative first thought about this. Let T be our total evidence rightly so-called, the evidence we ought to be updating on. ω has no right to be included among T's truthmakers if the process by which T reaches us can be seen in advance not to involve the fact (if it is one) that ω.

Can E^- be T by this standard? No, for ω was introduced in effect by fiat when we inferred E^- from E. If ω was not involved in the process by which E was learned, then the same holds for E^-, since inference from E does not give ω a way of getting in on the causal action. Our new principle—call it the *Principle of Perspicuity*, though it might be seen as just a ways-adjusted form of *Total Evidence*—should therefore be this:

> take as your evidence an E with no ways of being true beyond those potentially implicated in the process by which E is learned.

This principle helps to explain why A strikes us as supporting $E\mathfrak{b}$ more than B does. Assuming with *Perspicuity* that an evidence-statement's ways of being true are limited to states of the world potentially implicated in our acquiring that evidence, the fact of $\mathfrak{b}$ being eaten

(a) figures potentially in the process by which A is learned, but
(b) does not figure potentially in the process by which B is learned.

[37] A has the same truthmakers as $A\vee B$: the set of all ω-sequences of ✓'s and —'s containing finitely many —'s.

[38] A is in another sense *stronger* than B; it "says more" (Holliday's phrase) by having a larger subject matter. I speak of weakening, because the larger subject matter comes in the form of additional ways of being true—additional "disjuncts"—and adding disjuncts is (like removing conjuncts) a paradigmatic form of weakening,

Small wonder then that A strikes us as supporting $E\mathfrak{b}$, while B doesn't. To gain confidence in $E\mathfrak{b}$ based on evidence potentially sent your way by the fact that $E\mathfrak{b}$ seems only reasonable. To gain confidence in $E\mathfrak{b}$ on the basis of evidence that would have come to you by the same route whether H held or not — not so much.

19 Last Thoughts

I said that evidential support were understood (as by Bayesians) in terms of conditional probability, then A has to support E if E support A — $\text{pr}(X|Y) > X$ iff $\text{pr}(Y|X) > Y$.

But the Bayesian explication is out of the question here. For one thing, conditional probability is ordinarily seen as intensional and A and B are intensionally equivalent. A deeper reason is brought out by Builes' Paradox and the critical reaction to it.[39] Builes and Dorr et al point out that $\text{pr}(E|A)$ cannot differ from $\text{pr}(E|B)$, if (as seems inevitable in this case) $\text{pr}(A|B) = \text{pr}(B|A) = 1$.[40]

Those who persist in believing that *Almost all apples are eaten* ought in some sense to boost our confidence in $\mathfrak{b}$ *is eaten* will have to pursue this idea using some sort of non-Bayesian apparatus still undevised. [41] Resisting world domination (putting ways where worlds were) is necessary for this purpose, or so we have argued. But it is nowhere near sufficient. I don't know if probability theory can be run on states of the world, or more or less specific ways for the world to be. If so that could be just the ticket.

References

Frank Arntzenius. No regrets, or: Edith piaf revamps decision theory. *Erkenntnis*, 68(2):277–297, 2008.

Rachael Briggs. Interventionist counterfactuals. *Philosophical studies*, 160(1):139–166, 2012.

David Builes. A paradox of evidential equivalence. *Mind*, 129(513):113–127, 2020.

Rudolf Carnap. On the application of inductive logic. *Philosophy and phenomenological research*, 8(1):133–148, 1947.

Hector-Neri Castañeda. Ought, value, and utilitarianism. *American Philosophical Quarterly*, 6(4): 257–275, 1969.

Roderick M Chisholm and Ernest Sosa. On the logic of" intrinsically better". *American Philosophical Quarterly*, 3(3):244–249, 1966.

Ivano Ciardelli, Linmin Zhang, and Lucas Champollion. Two switches in the theory of counterfactuals. *Linguistics and Philosophy*, 41(6):577–621, 2018.

Yifeng Ding, Wesley H Holliday, and Thomas F Icard III. Logics of imprecise comparative probability. *International Journal of Approximate Reasoning*, 132:154–180, 2021.

Cian Dorr, John Hawthorne, and Yoaav Isaacs. Solving a paradox of evidential equivalence. *Mind*, page fzaa022, forthcoming. doi: 10.1093/mind/fzaa022.

Kenny Easwaran. Why countable additivity? *Thought: A Journal of Philosophy*, 2(1):53–61, 2013.

[39] Builes [2020], Dorr et al. [forthcoming]

[40] They argue from Popper's Multiplicative Axiom.

[41] Comparative probability may be helpful here (Walley and Fine [1979], Fishburn [1986], Hawthorne et al. [2016], Ding et al. [2021]). But "confidence-boosting" needs it would seem a metric aspect to be useful.

Andy Egan. Some Counterexamples to Causal Decision Theory. *Philosophical Review*, 116(1): 93–114, 2007.

Kit Fine. Counterfactuals without possible worlds. *The Journal of Philosophy*, 109(3):221–246, 2012.

Kit Fine. Angellic content. *Journal of Philosophical Logic*, pages 1–28, 2015.

Kit Fine. Truthmaker semantics. *A companion to the philosophy of language*, pages 556–577, 2017.

Peter C Fishburn. The axioms of subjective probability. *Statistical Science*, pages 335–345, 1986.

Sven Ove Hansson. Preference logic. In *Handbook of philosophical logic*, pages 319–393. Springer, 2001.

James Hawthorne et al. A logic of comparative support: Qualitative conditional probability relations representable by popper functions. In *The Oxford handbook of probability and philosophy*. Citeseer, 2016.

Richard C Jeffrey. A note on the kinematics of preference. *Erkenntnis*, 11(1):135–141, 1977.

Richard C Jeffrey. *The logic of decision*. University of Chicago press, 1990.

James M Joyce. Levi on causal decision theory and the possibility of predicting one's own actions. *Philosophical Studies*, 110(1):69–102, 2002.

James M Joyce. Regret and instability in causal decision theory. *Synthese*, 187(1):123–145, 2012.

Justin Khoo. Disjunctive antecedent conditionals. *Synthese*, pages 1–30, 2018.

Angelika Kratzer. An investigation of the lumps of thought. *Linguistics and Philosophy*, 12:607–653, October 1989.

Angelika Kratzer. How specific is a fact. *University of Massachusetts*, 1990.

Angelika Kratzer. Facts: Particulars or information units. *Linguistics and Philosophy*, 25(5-6): 655–670, December 2002.

Daniel Lassiter. Complex antecedents and probabilities in causal counterfactuals. In *21st amsterdam colloquium*, pages 45–54, 2017.

Daniel Lassiter. Complex sentential operators refute unrestricted simplification of disjunctive antecedents. *Semantics and Pragmatics*, 11:9, 2018.

David Lewis. Causal decision theory. *Australasian journal of philosophy*, 59(1):5–30, 1981.

Kevin McCain and Ted Poston. Best explanations. *Best explanations: New essays on inference to the best explanation*, page 1, 2017.

Judea Pearl. Physical and metaphysical counterfactuals: Evaluating disjunctive actions. *Journal of Causal Inference*, 5(2), 2017.

Nicholas Rescher. The logic of preference. In *Topics in Philosophical Logic*, pages 287–320. Springer, 1968.

William Roche and Elliott Sober. Explanatoriness is evidentially irrelevant, or inference to the best explanation meets bayesian confirmation theory. *Analysis*, 73(4):659–668, 2013.

Bernhard Salow. The externalist's guide to fishing for compliments. *Mind*, 127(507):691–728, 2018.

Paolo Santorio. Alternatives and truthmakers in conditional semantics. *The Journal of Philosophy*, 115(10):513–549, 2018.

Paolo Santorio. Simplification is not scalar strengthening. In *Semantics and Linguistic Theory*, volume 30, pages 624–644, 2021.

Mark Siebel. Why explanation and thus coherence cannot be reduced to probability. *Analysis*, 71 (2):264–266, 2011.

Brian Skyrms. Ratifiability and the logic of decision. *Midwest studies in philosophy*, 15:44–56, 1990.

Elliot Sober. Why logically equivalent predicates may pick out different properties. *American Philosophical Quarterly*, 19(2):183–189, 1982.

Jack Spencer. An argument against causal decision theory. *Analysis*, 81(1):52–61, 2021.

Georg Henrik Von Wright. *Norm and action*. Routledge and Kegan Paul, London, 1963.

Peter Walley and Terrence L Fine. Varieties of modal (classificatory) and comparative probability. *Synthese*, pages 321–374, 1979.

Roger White. Explanation as a guide to induction. *Philosophers' Imprint*, 5(2):1–29, April 2005.

Stephen Yablo. *Aboutness*. Princeton University Press, 2014.

Stephen Yablo. Relevance without minimality. In Dirk Kindermann, Peter van Elswyk, and Andy Egan, editors, *Unstructured Content*. OUP, 2022?