On being a random sample[1]

David Manley

It is well known that *de se* (or 'self-locating') propositions complicate the standard picture of how we should respond to evidence. This has given rise to a substantial literature centered around puzzles like Sleeping Beauty, Dr. Evil, and Doomsday[2]—and it has also sparked controversy over a style of argument that has recently been adopted by theoretical cosmologists.[3] These discussions often dwell on intuitions about a single kind of case, but it's worth seeking a rule that can unify our treatment of all evidence, whether *de dicto* or *de se*.

This paper is about three candidates for such a rule, presented as replacements for the standard updating rule. Each rule stems from the idea that we should *treat ourselves as a random sample*, a heuristic that underlies many of the intuitions that have been pumped in treatments of the standard puzzles. But each rule also yields some strange results when applied across the board. This leaves us with some difficult options. We can seek another way to refine the random-sample heuristic, e.g. by restricting one of our rules. We can try to live with the strange results, perhaps granting that useful principles can fail at the margins. Or we can reject the random-sample heuristic as fatally flawed—which means rethinking its influence in even the simplest cases.

1. Inward and outward

1.1. From de se to de dicto—and back

We often treat ourselves as though we are random samples drawn from a larger group of individuals. Sometimes we know a fact about the group and

---

[1] Precursors of this paper have been knocking around for more than ten years, and under a few different titles, including 'Self-location and the Existential Selection Effect'. Thanks are due to many people for discussion of the issues herein, or comments on one or another of its previous forms, especially Maria Aarnio, Frank Arntzenius, David Baker, Cian Dorr, Kenny Easwaran, James Joyce, Christopher Meacham, Sarah Moss, Eric Swanson, and Brian Weatherson. A special helping of gratitude goes to three people in particular whose input has been absolutely critical—John Hawthorne, Jacob Ross, and Charles Sebens.

[2] For more on these puzzles see, e.g., Bartha and Hitchcock 1999; Bostrom 2001, 2002a; Elga 2000, 2004; Leslie 1989, 1996.

[3] The key premise has been called 'typicality', 'mediocrity', 'the Copernican principle', and is sometimes even identified with the 'anthropic principle'. See, e.g., Linde 2007, Page 1996, Tegmark 2004, Vilenkin 2011, Guth 2000, Dyson, Kleban, and Susskind 2002; however, for a contrary view, see also Hartle and Srednicki 2007.

use it to draw conclusions about ourselves. And sometimes the reverse occurs—we start with a fact about ourselves and work our way out to conclusions about the group or even the world as a whole.

Suppose I have some medical symptoms, for example. I am worried about whether I have condition C, which I know is suffered by at least some people with my symptoms and medical history. But at first I have no idea whether C is common or rare among people in this reference class. Now consider two variations on the example:

> *Inward*: I learn that *most people* with my symptoms and history have C. I conclude that I probably have C.
>
> *Outward*: I learn that *I have C*. I treat this as (at least some) evidence that C is common among people with my symptoms and history.

In neither case do I treat the proposition that I have C as merely equivalent to the proposition that *someone* in my medical reference class has *C*—that is something I knew from the outset. In *Outward*, for example, learning that I have C is a much stronger piece of evidence than learning that *someone* in the reference class has C. It amounts to something more like *a randomly selected member* of the reference class has C.

At issue in these cases is the connection between what are called *de dicto* propositions (i.e. those about how the world is generally) and *de se* propositions (i.e. those about one's own place in the world). As I will be using the phrase, to *reason inward* is to draw *de se* conclusions from a piece of *de dicto* evidence. And to *reason outward* is to draw *de dicto* conclusions from a piece of *de se* evidence. In many cases, these two kinds of reasoning are entirely mundane and uncontroversial. But in other cases, they raise fundamental questions about the requirements of rationality and the nature of evidence.

1.2. Beyond entailment links

In easy cases, the relevant *de se* and *de dicto* propositions are linked by simple entailment. Suppose I learn that I am in building 2. Then I can conclude that *someone* is in building 2. Or, if I learn that *no one* is in building 2, I can conclude that I am not in building 2. In other cases, I can avail myself of background knowledge to make such inferences—for example, maybe I know that I am the only person in building 2. Then I will treat any *de se* hypothesis about myself as equivalent to a *de dicto* hypothesis about the unique occupant of building 2.

Other cases of inward and outward reasoning are trickier. For example, suppose we are wondering whether our galaxy has some unobservable feature F. Our best cosmological theory tells us that there are plenty of life forms in

the universe, only some of which will inhabit galaxies that are F. Now consider two variations on the example:

> *Inward**: Our best theory tells us that most life forms inhabit F-galaxies. We conclude that we probably live in an F-galaxy.
>
> *Outward**: We learn that our galaxy is F. We treat this as (at least some) evidence that most observers inhabit F-galaxies.

Again, we are tempted to treat ourselves as though we were chosen at random from among the various life forms in the universe. (It is this form of reasoning that crops up a lot in theoretical cosmology, as we will see.)

Can these transitions between *de dicto* and *de se* be justified by entailment links? It depends on the details. Suppose we can assume that while there may be plenty of life out there, there is only one species *exactly* like us—right down to every aspect of biology and culture. If we know that we have some unique feature U, we can treat the *de se* proposition *We inhabit an F-galaxy* as equivalent to the *de dicto* proposition *The U-people inhabit an F-galaxy*. With this entailment link in hand, no further inward or outward reasoning is needed. In place of any *de se* proposition, we can simply use a corresponding *de dicto* proposition about the U-people. (Suppose we treat every galaxy as having an equal chance of giving rise to the U-people, regardless of F-ness. If there are *more* F-galaxies, they are collectively more likely to have given rise to the U-people. And—in the other direction—if the U-people inhabit an F-galaxy, there are probably more F-galaxies.)

So far so good. But suppose instead that we know the universe is *really big*—it is a multiverse made up of many (but finitely many) universes. Moreover, scattered around this multiverse there are guaranteed to be many populations *exactly* like ours—right down to our biology, cultures, and experiences. Some inhabit F-galaxies and others inhabit non-F-galaxies. Given this, we can't use a unique feature to generate a proxy *de dicto* claim for every *de se* claim. Still, we don't lose our ability to reason inward when we learn statistical facts about populations like ours in cases like this:

> *Inward***: We learn that the vast majority of populations exactly like ours inhabit F-galaxies. We conclude that we probably inhabit an F-galaxy.

Likewise, it can be natural to reason *outward* even in the absence of entailment links:

> *Outward***: We learn that our own galaxy is F. We treat this as (some) evidence that most populations like ours inhabit F-galaxies.

Here the strongest *de dicto* hypothesis entailed by our *de se* evidence is something we already knew—viz. that a population just like ours inhabits an F-galaxy. And yet the inference seems perfectly natural. And note that this form of reasoning can be extended to provide purely *de se* evidence that bears on

cosmological theories. For example—if two cosmological theories differ on the expected proportions of observers like us in F-galaxies, the fact that our own galaxy is F is evidence for one theory over the other. Indeed, just this sort of outward reasoning has been by theoretical cosmologists in support of a range of hypotheses.

1.3. The plan

This paper is about hard cases of inward and outward reasoning—cases where entailment links and the standard updating rule cannot guide us. We will examine three rules that can take its place, and consider how well they reflect our judgments about both hard and easy cases.

Our rules can replace ordinary updating in the sense that they offer answers where simple *de dicto* conditionalization does not. But when considering each rule as a potential replacement, we will set aside the question whether it should carry the same normative force that is usually afforded conditionalization. One could hold that one's favorite rule—insofar as it goes beyond standard conditionalization—should be treated as something like a *reasonable strategy* rather than a *constraint on rationality*. It is worth seeking a coherent, systematic way to treat hard cases as well as easy ones, even if we doubt that our credences in hard cases are subject to norms of the same force.

Here is the plan for the rest of the paper. In the next we will consider how the random sample heuristic applies in cases of purely inward reasoning. But most of the hard cases in the literature—both in philosophy and in cosmology—involve what I have been calling 'outward' reasoning. And the application of the random sample heuristic is trickier in such cases. Accordingly, most of this paper will focus on strategies for outward reasoning—one that rejects outward reasoning beyond entailment links, and two that embrace it. We will conclude by discussing cases where even those principles break down, and asking whether there is reason, on balance, to prefer one of these approaches to the others.

2. Inward

2.1. Worldmate sampling

Here is a well-known case of inward reasoning beyond entailment links:

> *Dr. Evil.* The Philosophy Defense Force has a plan to defeat Evil in his impregnable battle-station—simply convince Dr. Evil that we have created a duplicate of him that is having exactly his experiences, and that this duplicate will be tortured unless the duplicate performs actions corresponding to surrender.

According to Elga, once Dr. Evil is convinced he has a duplicate, he should assign a .5 credence to the hypothesis that he is the original Dr. Evil. In other

words, he should treat himself as a random sample from the two individuals that have exactly his experiences (2004).

We need some jargon to state the principle Elga applies to this case. Take all my current evidence, including my apparent memories, and call that my 'complete evidential state' or 'CES'. (I will be assuming an internalist picture of evidence on which one's CES is given by one's current qualitative experiences and apparent memories, all internally individuated. One could also spell out all of the principles I will be discussing in more externalist terms—though this would yield different results in some cases, perhaps including this one.[4])

Next, a 'world' is a fully specific *de dicto* hypothesis, and a 'predicament' is a fully specific situation in which a subject has a CES in a world, at a time.[5] Now we can state:

> ELGA'S RULE. Assign equal credence to any two predicaments in the same world with one's CES.[6]

---

[4] For example, if we think of Dr. Evil's evidence (even after his duplication) as including the content of his memory that he built the battle-station, rather than the mere seeming-to-remember that he built the battle station, then a molecule-for-molecule duplicate embedded in a perfect replica of the battle-station will not have Dr. Evil's CES. Given this picture of evidence, Elga's rule would recommend continued certainty that he is the original.

[5] I adopt the term 'predicament' from Elga. Predicaments are *maximal situations* in which a subject can find herself, where this includes her world's being as it is. (As such, predicaments are 'world-bound'.)

[6] The original principle concerns the *indistinguishability* of predicaments, but gives rise to inconsistency if that relation is intransitive (Weatherson 2005, §4). (Think of a Sorites series of color experiences.) Putting things in terms of *sameness* of CES avoids this threat of inconsistency, but raises a related problem: arguably one is not always in a position to know when one has a given CES, and therefore whether one is following the rule. More generally, Williamson has cast doubt on the very possibility of stating doxastic rules whose conditions are *luminous*—viz. such that, whenever they obtain, one is in a position to know that they do (2000). Perhaps we can content ourselves with rules whose conditions are *lustrous*—such that, whenever they obtain, one is in a position to *justifiably believe* that they do (Berker 2008).

I will set aside this important debate in the text. My own view is that we should not be overly concerned with a subject's access to her status vis-à-vis doxastic norms: she can properly implement the relevant rules without being in a position to know that she has done so. But none of this undermines the usefulness of more-or-less 'subjective' norms. The trouble with highly 'objective' norms like 'believe only what you know' or 'believe the truth' is not that the involve non-luminous conditions, *per se*. It is that they fail provide the theorist with necessary or sufficient conditions for doing what one ought to do, doxastically. And they fail to offer much practical guidance for those who are trying to develop good doxastic habits.

So, if I learn that most predicaments with my CES have some feature, I should have a higher than even credence that I myself have that feature.

More formally, we can state this idea as a constraint applying to hypotheses (whether *de dicto* or *de se*) conditional on worlds.[7] Let 'E' specify my current CES, and 'p_E' be my posterior credence function upon having E. For any world W in which there is at least one predicament with my CES, let 'n_W(E)' be the number of predicaments that have E according to W. And let 'n_W(E&H)' be the number of predicaments-with-E that also exemplify H, according to W. (A predicament *exemplifies*—or is an *exemplar* of—H just in case H is true of the subject of that predicament, at the time of that predicament. A *de dicto* hypothesis that is true at a world is exemplified by every predicament in that world.) Here, then, is the rule:[8]

WORLDMATE SAMPLING (WS): $$p_E(\mathrm{H} \mid \mathrm{W}) = \frac{n_W(\mathrm{E} \,\&\, \mathrm{H})}{n_W(\mathrm{E})}$$

This rule supports our judgments in *Inward***: if we learn that the vast majority of populations exactly like ours inhabit F-galaxies, we should conclude that we probably inhabit an F-galaxy.[9]

2.2. WS and additivity

It is worth addressing the additivity problem faced by worldmate sampling. Suppose I assign some nonzero credence to worlds where there are infinitely many subjects with my CES, each tagged with a natural number. Then Elga's rule requires me to assign equal credences to countably many *de*

---

[7] The principles are not quite equivalent: one difference is that WS involves the assumption that $p_E(\mathrm{H} \mid \mathrm{W})$ is defined for any world containing a predicament with my CES, while Elga's rule is consistent with assigning a credence of 0 to every predicament in such a world and allowing $p_E(\mathrm{H} \mid \mathrm{W})$ to go undefined.

[8] This principle is far narrower than what is ordinarily called 'the principle of indifference'; in particular, it is not obviously subject to 'cube factory' worries of the sort raised by van Fraasen 1989 and discussed by White 2008 and Novack 2010.

[9] *Modulo* the following concern: if we consider infinitely many worlds containing predicaments with his CES, and assign the same infinitesimal (or zero) credence to each, we automatically satisfy WS whatever credence we give more general hypotheses like 'Our galaxy is F'. As Weatherson notes, we can fix this problem by making the principle govern multi-world hypotheses directly. For example:

> (WS*) For any hypotheses X and Y such that in every world where either is exemplified, X has exactly N times as many exemplars with one's CES as Y does: X deserves N times the credence of Y.

(This generalizes on Weatherson's suggestion, which compares only pairs of hypotheses with at most one exemplar per world.)

se hypotheses of the form 'Conditional on being in that world, I am tagged with *n*'. This conflicts with the principle of additivity. Meanwhile WS, as stated, simply yields undefined credences for every hypothesis conditional on such worlds.[10]

There are several ways for proponents of WS to respond. First, we could embrace the fact that WS insists on undefined credences in these cases—after all, what better credences are there? Unfortunately, the failure to define $p_E(\mathrm{H}|\mathrm{W})$ for any world will tend to infect $p_E(\mathrm{H})$ more generally, since the latter should equal the sum of every conditional probability of H given a world, weighted by the probability of the world. Thus if I assign some non-zero credence to a world with infinitely many subjects with my CES, some of whom are Italian, my credence in being Italian will go undefined. Better to sum only the *defined* conditional values, and let this constitute a lower bound for my credence in H—and likewise let the defined values for $p_E(\sim\mathrm{H}|\mathrm{W})$ constitute an upper bound.

However, maybe it isn't right to impose undefined credences in the relevant infinitary worlds. A second option is to add 'where defined' to the principle, and go permissivist about the rest. Compare the idea that it's perfectly fine—perhaps obligatory—to treat one's lottery ticket as equally likely to win as any other ticket, as long as the lottery is finite. In a countably infinite case, however, one additive distribution is as good as any other. Relatedly, we could adopt WS in its original form but hold that it has the force of a *ceteris paribus* rule that must be followed unless trumped by a stronger rule like additivity. (Or vice versa, depending on one's feelings about countable additivity. One might wonder: what's so important about adding up that particular cardinality of infinity?)

Another option is to be permissivist only about *some* infinitary cases. For example, in the absence of numerical proportions of $n_W(\mathrm{E\&H})$ to $n_W(\mathrm{E})$ there may sometimes be a natural substitute. For example, consider a world in which, among the countably subjects that have my CES, countably many exemplify H and countably many don't. Now, all these subjects are embedded in an ordinary spacetime, and if one takes any point in that spacetime and

[10] Weatherson also shows how WS gives rise to a version of the 'shooting room' paradox. (See 2005, §8.) It's not clear to me why the kind of case he considers is any more trouble for the proponent of WS than the possibility of a shooting-room setup is for *anyone*. Indeed, the intuition that such a setup is possible has been used to motivate giving up countable additivity. Perhaps the idea is that the standard shooting-room setup must avail itself of the possibility of a 'random sample' from a countable infinity, whereas Weatherson's version aimed at WS merely requires a deity with the capacity to create infinitely many beings at speed.

considers spheres around that point of increasing size, one finds the ratio of H-subjects to non-H subjects always approaches 2/3. In that case, it seems plausible to assign a credence of 2/3 to being H, conditional on that world.[11]

Alternatively, we could allow credences to go *imprecise* in the problematic cases. Views differ about how to model a proper belief state in a case where having a specific credence assignment is arguably inappropriate: we could, for example, treat it as a *set* of functions rather than as a single credence function. Brian Weatherson has argued that WS wrongly treats *uncertain* questions as *risky*, but he suggests a related rule: any two predicaments with my CES should be assigned the same *set* of credences by these functions (2005; §6). Since this rule is consistent with each function satisfying additivity in the problem cases,[12] proponents of WS could insist on full-strength WS for finite cases and co-opt this idea only for infinite cases. This would lead to a very limited amount of imprecision in the ordinary case, such as my credence in being Italian.

Finally, Jacob Ross has suggested that conflicts between principles like WS and countable additivity might be best thought of as generating rational dilemmas: situations where two genuine constraints on rationality cannot both be satisfied (2010 §5). For those who have independent reasons to recognize the existence of conflicting rational requirements, this may not be such a bitter pill to swallow.[13]

2.3. WS and belief dynamics

It is well known that the dynamics of *de se* beliefs cannot properly be modeled by a flat-footed application of standard conditionalization.[14] Consider the following example from Frank Arntzenius: Jane is omniscient about the *de dicto* facts, watching a clock that she's certain is accurate (2003: 367). At first

---

[11] Thanks again to Cian Dorr for this point, and for discussion of the issues in this section.

[12] Think of each credence function as a kind of 'committee member' in one's internal 'committee of uncertainty'. This principle requires that committee members' assignments collectively balance out. If one committee member assigns credence *n* to my being the subject tagged with the number 1, another committee member must assign that credence to my being the subject tagged with the number 2.

[13] For discussion see Priest 2002 and Christensen 2007.

[14] I am assuming a treatment of self-locating belief along the lines of Lewis 1979. For some recent discussion of the general problem, along with proposals at varying levels of generality—some of which entail WS—see Halpern 2006, Titelbaum 2008, Meacham 2008, Meacham 2010, Schwarz, 2012, Moss 2012.

she is certain that it is 6am. Her credence that it is 7am is therefore zero, so later when the clock reads '7am' and she becomes certain that it is 7am, Jane does not reach that belief by the standard updating rule.[15] After all, diachronic conditionalizing involves zooming in on the portion of her previous epistemic space that is consistent with her new evidence; it has no mechanism allowing her to *gain* credence in hypotheses she has already ruled out.[16]

Now suppose that, rather than trying to apply conditionalization to *de se* beliefs, Jane were simply to apply WS. She would then conform her credence in *it is 7am* conditional on W to the expected fraction of predicaments with her CES in W who in fact exemplify *it is 7am*. She knows that W holds and also that there is exactly one predicament in W with her CES. So $n_W(\mathrm{E}) = 1$. And that predicament also exemplifies the hypothesis that it's 7am—so $n_W(\mathrm{E\&H}) = 1$. Jane ends up certain that it is 7am.

Similarly, suppose Dr. Evil knows that he will be duplicated in a short time. At first he should be certain that he is the original Dr. Evil, but it seems to many of us that he should lose his certainty when he arrives at the time for duplication (say, t_2). Conditionalizing on his *de se* evidence that it is now t_2 won't work, because his prior credence in its being t_2 is zero. But using WS instead, his new credence that he is the original (conditional on any world consistent with his knowledge) will equal the fraction of predicaments with his CES in that world that are the original—namely, ½.[17]

In effect, WS allows one to distribute *de se* beliefs within worlds using only *de dicto* priors. This means, in effect, that one can dispense with the need for conditionalizing on de se priors—at least when it comes to purely *de se* shifts in

---

[15] In our notation, the update rule is $p_E(H) = p_{OLD}(\mathrm{H} \mid \mathrm{E})$. Assuming that the latter value is defined as $p_{OLD}(\mathrm{E\&H}) / p_{OLD}(\mathrm{E})$ rather than treated as primitive, the result is undefined because $p_{OLD}(\mathrm{E}) = 0$. Those who prefer primitive conditional credences for *de se* hypotheses would still face the question of where these values should come from: see fn 30. The principles that follow could be treated as proposals for constraining primitive conditional credences, rather than updating rules.

[16] This example involves a loss of *certainty*. If Jane can't be certain that the clock reads 6am, we could replace this with the belief that the clock *seems* to read 6am'. But maybe Jane should not be *certain* even about how things seem to her; perhaps she should leave open the possibility that she's having a cognitive hallucination, and in fact the clock seems to read 7am. But in that case, updating on 'the clock seems to read 7am' should be good evidence that she's having a cognitive hallucination. And that's the wrong result too. See Schwarz, 2012.

[17] In a context where Dr. Evil is distributing credences over infinitely many worlds, yielding an unconditional credence of ½ for being the original will require an assumption about how the space of worlds is partitioned, or else the use of WS* in place of WS. (See footnote 9 below.)

belief. These results are suggestive, at least for those who agree with the intuitions that WS was formulated to capture.[18] However, this is not a complete solution to the problem of *de se* updating, because WS only yields credences in *de se* hypotheses conditional on worlds. It does not tell us how to arrive at credences in those worlds, or any other *de dicto* hypotheses—and thus does not by itself allow us to arrive at unconditional *de se* credences.

For example, suppose Dr. Evil knows that the Philosophy Defense Force tossed a coin and will create a duplicate of him at t_2 only if the coin came up heads. In that case, how should he integrate his *de dicto* priors (constrained by the chances) with the *de se* evidence that it's t_2? "That's easy," one might think, "He should just use the chances to distribute his *de dicto* credences among the worlds, and then use WS to distribute his *de se* credences within worlds. He should end up with a credence of ¾ that he is the original." That does seem to be the most natural answer—but §3 presents some reasons to think it's wrong.

3. Outward: INVARIANCE

The natural idea we just encountered is that there should be no outward reasoning except what is forced by entailment links: one should arrive at *de dicto* credences using only *de dicto* priors and *de dicto* evidence. Accordingly, for purposes of comparing *de dicto* hypotheses, one's *de se* evidence will be equivalent to its strongest *de dicto* entailment: roughly, the fact *that this CES obtains*. (This idea been defended by Halpern 2006; Meacham 2008.)

More formally, let E′ be the strongest *de dicto* fact entailed by E. Then the rule for arriving at *de dicto* credences is simply to conditionalize on E′:

INVARIANCE: $p_E(\mathrm{H}) = p(\mathrm{H} \mid \mathrm{E}')$ for any *de dicto* hypothesis H.

This guarantees that outward reasoning can only occur using entailment links. And it is consistent with the use of worldmate sampling for *de se* credences. One would simply use *de dicto* conditionalization to evaluate world-hypothesis, and then distribute each world's credence value among the predicaments that exemplify E in that world.[19] This combination can be stated as a general rule for any H, whether *de se* or *de dicto*. Worldmate sampling

[18] There are other solutions to the problem of *de se* updating that self-consciously avoid the results yielded by WS, such as the 'shifted conditionalizing' recommended by Schwarz (2012). I won't try to evaluate these alternatives here: this paper concerns the options available to those who are moved by the intuitions behind WS.

[19] This is, at least, how a proponent of WS would implement what Meacham calls 'compartmentalized conditionalizing'. See also Halpern 2006 for a very similar approach that does not apply to possible cases of simultaneous duplication.

gives us $p_E(\mathrm{H}\,|\,\mathrm{W})$ for each world. Then this value gets weighted by the posterior probability of that world as supplied by INVARIANCE; for the unconditional credence in H, we sum the results.[20]

This is a very natural way to handle inward and outward reasoning. Unfortunately, it yields some highly counterintuitive results. To begin with, recall *Outward***:

> Our two remaining cosmological theories T and T* agree that there are many populations exactly like ours. But T predicts that most populations like ours will inhabit F galaxies, while T* predicts the opposite. At some point we discover that our own galaxy is F.

Here the strongest *de dicto* evidence entailed by our discovery is something we already knew—namely that at least one galaxy is F. So, according to invariance, the *de se* evidence that our own galaxy is F should not alter our credences in T and T*. But that seems wrong: we are inclined to treat it as evidence that theory T is correct, as though we had been selected at random from all the populations like ours. And indeed, this kind of outward reasoning beyond entailment links has become common in theoretical cosmology.

But things get even worse for INVARIANCE, as I will illustrate some simple cases where one's priors are settled by the objective chances. In the first type of case, INVARIANCE tells me I do not get any relevant evidence, but intuitively I do. And in the second, INVARIANCE tells me I *do* get evidence, but intuitively I do not.

3.1. Too little evidence

Consider a variant on LIGHTS that involves fewer subjects but proceeds in two-steps:[21]

> (TWO-STEP) The gods toss a fair coin. If *heads*, one subject is created in a well-lit room; *tails*, two subjects are created in separate rooms—one well-lit and the other dark. (Their experiences are no more fine-

---

[20] Putting this together gives us:

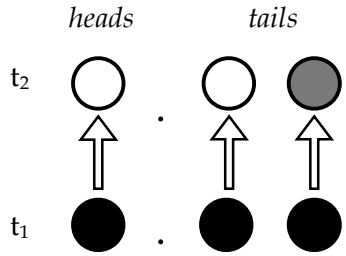
$$\text{WS-INVARIANCE: } p_E(\mathrm{H}) = \sum_W \left(p(\mathrm{W} \mid \mathrm{E}') \frac{n_W(\mathrm{E}\,\&\,\mathrm{H})}{n_W(\mathrm{E})} \right)$$

As with worldmate sampling, this way of putting things assumes additivity. Things go undefined if I assign a zero or infinitesimal credence to any hypotheses that contains a predicament with my CES, and that therefore I have not ruled out. We could be more flexible by conjoining invariance with (WS*) from fn. 9.

[21] Bostrom presents a number of similar (and ingenious) cases in his 2001, 2002a, and 2002b.

grained than this.) Every subject wakes with eyes shut at t_1. I wake up and a moment later open my eyes. The lights are on.

TWO-STEP



Here, circles of the same shade represent predicaments with the same CES. At t_1, all I have to go on is the Principal Principle, so I assign equal credence to *heads* and *tails*. When I see that the lights are on, I learn that I am not the person in the dark room. But this is merely *de se* evidence—on either outcome, someone will see lights. So INVARIANCE blocks any credence from seeping across the dotted line. (This is why Meacham calls this approach 'compartmentalized conditionalization'.)

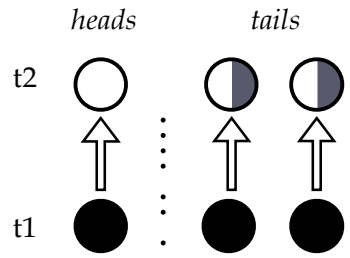
So I retain equal credences in *heads* and *tails*. But it seems quite clear that finding the lights on is evidence for *heads*—after all, given *tails* I might have found them off! Indeed, imagining myself in this situation, I'm not sure I would be able to restrain myself from reasoning this way. Whatever my credence at t_1—we are imagining it should be 1/2, though one of the other rules we consider will challenge this—it is hard to deny that I get evidence for *heads* at t_2.

Even worse, the invariantist has to posit a counterintuitive asymmetry between TWO-STEP and:

> (CHANCES) As in TWO-STEP, except that if *tails* there are two subjects. For each of them, there is an independent 50% objective chance of seeing lights.[22]

[22] This case is structurally similar to my 'black and white room' example, discussed as a potential problem for Meacham's view in [work that reveals the author's name omitted].

CHANCES



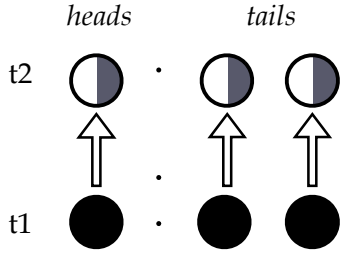
This time the *tails* outcomes split into four worlds, so seeing the lights on actually *does* rule out a *de dicto* hypothesis—viz. there are two subjects who both see darkness. So this time conditionalizing on my strongest *de dicto* evidence gives me evidence for *heads*, which is the right result. [23] But it is bizarre for a rule to require that I shift my credence towards *heads* in CHANCES but not in TWO-STEP.

3.2. Too much 'evidence'

Consider:

(CHANCES-2) As in CHANCES except that, however the coin lands, for every subject there is a 50% objective chance of seeing lights.[24]

CHANCES-2



Again, I begin with equal credences for the coin toss. Now suppose I see lights at t_2. This intuitively furnishes me with no evidence at all regarding the coin toss. But INVARIANCE rejects this intuition. After all, I do get *de dicto* evidence when I see lights—namely, *that someone sees lights*, an outcome that was more likely given *tails*. In particular, there are two distinct possible *heads* worlds and four distinct *tails* worlds. Each of the *heads* worlds initially has a credence of

[23] I am not here endorsing the particular credence assignment recommended by INVARIANCE—namely 4/7 in *heads*.

[24] This is a variant of an example used by Cian Dorr in connection with Sleeping Beauty, in Dorr (2002).

1/4 and each of the *tails* worlds has a credence of 1/8.[25] When I see that the lights are on, I rule out one *heads* world and one *tails* world—the ones where both subjects fail to see lights.[26] And renormalizing gives me a credence of 3/5 in *tails*.

With larger numbers, this result is more dramatic. Suppose the coin toss settles whether 1 or a million subjects will be produced, and every subject is randomly assigned a number between 1 and 100 (with replacement, of course). In that case, I will have equal credences in the outcomes of the coin toss until I see my number, at which point I will be nearly certain that *tails* was tossed. After all, the fact that *someone* sees the number 42 rules out a very high proportion of *heads* worlds, and a very low proportion of *tails* worlds.

Another disconcerting feature of these results is the fact that I will end up preferring tails *no matter what I see* when I open my eyes. And this is something I could predict with my eyes closed, though at that point INVARIANCE restrains me from adopting what I know will be my future credence. This requires a particularly egregious violation of the Reflection Principle, which requires us to conform to our future expected credences. Admittedly, there are cases where one simply cannot avoid violations of Reflection—but these involve the possibility of memory loss, losing track of time, having one's brain tampered with, and so on. [27] The invariantist has no such excuse.

Given these problems, it is worth seeing what happens if we allow outward reasoning beyond entailment links.

---

[25] I am assuming something like WS to arrive at these credences in the *tails* worlds: but this is only for convenience. The asymmetry only requires that I not assign 1/4 to the tails world with two dark rooms. And why in the world would I do that?

[26] For anti-haecceitist reasons, the compartmentalizer may deny that there are two distinct worlds where one observer sees lights and the other does not. In that case there are 3 possible *tails* worlds—one with an initial credence of 1/4 and two with an initial credence of 1/8. I rule out one of the latter, so the result is the same. (Alternatively, one could tell a story in which the incubator inconspicuously marks the subjects 'A' and 'B'.)

[27] See, e.g., Arntzenius 2004. Reflection requires that "the agent's present subjective probability for proposition A, on the supposition that his subjective probability for this proposition will equal *r* at some later time, must equal this same number *r*" (van Fraassen 1984, p 16). This is more plausible if one adds a condition like 'If an agent is certain that she will not lose her memory, come to doubt the veracity of her memories, or become cognitively impaired or brainwashed...' But even the weakened principle is violated by INVARIANCE.

4. Interlude: hypothetical priors

Before presenting the two strategies for outward reasoning, it is worth addressing some issues about belief dynamics that will influence how we state the rules. In particular, we will be examining cases where the subjects are required to access conditional 'priors' despite not having any literally temporal priors. In other words, they face a version of the problem of old evidence.

Here is a case that nicely illustrates the problem:

> (VASECTOMY) Prior to meeting my mother, my father flipped a coin. Iff *heads* was tossed, he would undergo an irreversible vasectomy. This would make the chances of ever conceiving a child very slim.

Suppose I know the setup but have no other relevant evidence. At the time of the toss, the objective chance of *tails* was 1/2. But that should no longer be my credence in *tails*. Presumably the Principal Principle (Lewis 1980) does not apply because I have inadmissible evidence—namely that I exist. To arrive at a credence for tails, I would like to integrate the background chances with the probability that *I exist* conditional on each coin toss.[28] But I can't update on something I've always known.[29]

In what follows I will assume that in cases where subjects have no priors, they can still use 'hypothetical' or 'hypothetical ' priors—roughly, a credence function that encodes one's epistemic norms as applied to propositions in the absence of any evidence at all, including the relevant 'old evidence'.[30] And, importantly, VASECTOMY illustrates that one's *very existence* can count as evidence that needs to be bracketed.

One might think that use of hypothetical priors need only stand in for temporal priors in cases where the latter are unavailable for use, as when

---

[28] It might help to allay some concerns if we make it a feature of the example that I came into existence knowing about the coin toss set-up. (Someone might suggest that, if I learn about the setup late in life, then as an ideally rational agent I would have always had a conditional credence in *I am told that my father flipped a coin, etc*, given that *my father flipped a coin, etc. and the coin came up heads*. But it is hard to see anyway how I would assign the intuitively correct credences here, with my existence as background knowledge. I would have to mimic the kinds of hypothetical priors we are about to discuss.)

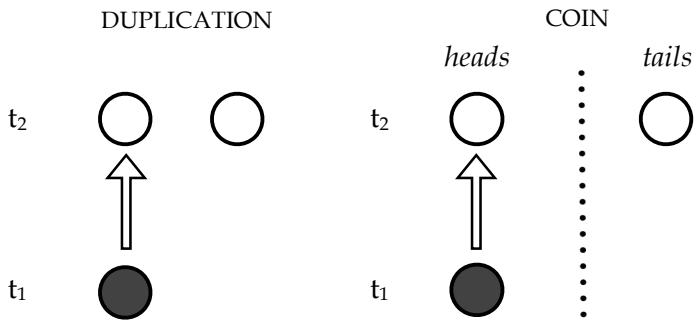
[29] Some, e.g. Pust 2007 have argued that 'Cartesian' knowledge—knowledge of a sort that cannot be doubted—can never be treated as evidence. I take our intuitions in cases like this to be sufficient reason to reject this outright prohibition, and to require that a solution to the problem of old evidence can accommodate even Cartesian evidence.

[30] For some discussions of the problem, see Earman 1992, ch. 5; Glymour 1980, ch. 3; Howson and Urbach 1989, 272-75; Joyce, ch. 6.

someone has just come into existence. But this brings us to our second point about priors. Consider the following case, modeled on the 'Shangri La' case in Arntzenius (2003):

> (COIN) The gods tossed a coin. If *heads*, I get created at t_1 and live happily ever after. If *tails*, I get created at t_2 with false memories of being at t_1 and then destroyed. My CES at t_2 would be the same either way.

The structural resemblance with a Dr. Evil-style case of certain duplication can be illustrated as follows:



Assume *heads* comes up in COIN. In that case, I start off certain that I am not going to die, but it seems that I should start to worry that I am going to die when I reach t_2—just as in the case of Dr. Evil. After all, as Frank Arntzenius puts it, "you know that you would have had the memories that you have either way and hence you know that the only relevant information that you have is that the coin was fair" (356).

However, to make good on this intuition, we must give up on the rule that my *de dicto* credences should only change as a result of conditionalization. After all, I start off certain in the *de dicto* hypothesis that *heads* came up but then start to doubt this when I reach t_2: conditionalization will not allow this. Moreover, my failure to adhere to conditionalization is not explained by any *involuntary* loss of information—as when one gets hit over the head and develops amnesia. By hypothesis, I proceed normally through time with no adverse cognitive events. I do become *worried* that my memories are not veridical; but that only happens *because* I fail to update by conditionalization. The epistemic possibility of false memories is a *symptom* of my violation of the rule, not its *source*.[31] In short, conditionalization not only fails to model the dynamics of my *de se* beliefs; it also fails to model the dynamics of my *de dicto* beliefs.

[31] Neither is this like a case where a new *de dicto* degree of belief impinges on one's cognition with the force of evidence, as in Jeffrey conditionalization—one *arrives* at

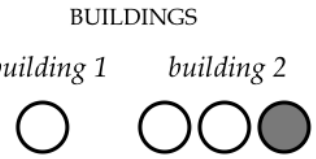
To handle this case, we can use an updating rule that *never* actually appeals to one's temporally prior conditional credences, but involves a complete update on hypothetical priors at every point instead.[32] On this approach, one's current credences are generated from one's complete current evidence along with one's current hypothetical priors. Note that the idea is not to *revert* to the use of hypothetical priors only when one is uncertain as to whether you have real past priors. The heart of the difficulty with COIN is that one only *becomes* uncertain as to whether one has any past priors *as a result* of abandoning them!

In my statements of the two strategies for outward reasoning, I will take the 'hypothetical prior approach' for granted. But for those who prefer to treat updating as genuinely diachronic, there are ways of amending both principles to use real priors—if they are available.[33]

5. Outward: transworld sampling

Consider the following self-location problem:

> (BUILDINGS) In building 1, the gods create one subject that sees lights. In building 2, the gods create three subjects, but only two see lights. (Their experiences are no more fine-grained than this.) Knowing the protocol, I wake and see lights.



Worldmate sampling tells me to treat each of the lights-seeing predicaments as equally likely in every world consistent with me evidence—so I should be twice as confident that I am in building 2 as that I am in building 1.

But why, exactly? Here are two potential answers:

---

one's new credences by reasoning. (If the coin is weighted, arriving at the right credence that one is the duplicate will involve some calculation!)

[32] Meacham calls this 'hp-conditionalizing' (2008:248).

[33] See fn. 44. However, those motivated by the sense that information loss of this kind is irrational will likely have a similar reaction to DUPLICATION, and so reject WS. In that case they will be interested in the discussion that follows only insofar as it contains arguments *against* certain ways of generalizing on WS!

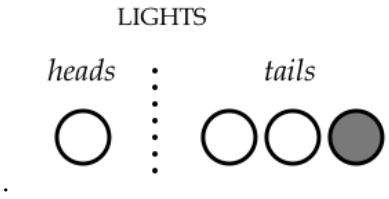
(a) Because there are twice as many predicaments with my CES in building 2 as there are in building 1.

(b) Because the predicaments with my CES in building 2 make up *twice as great a proportion of all the predicaments in the world* as do the predicamants with my CES in building 1.

Of course, these options are equivalent when the total number of predicaments is held fixed. (Comparing the numbers yields 1 vs. 2, while comparing the proportions yields 1/n vs. 2/n—where n is the total number of subjects in the world.) Since WS only distributes credences within worlds, it codifies each of these two ideas equally.

Crucially, however, these ideas come apart when we try to generalize on the random sample heuristic so that it applies across worlds. Consider, for example, this structurally similar case (based on examples from Bostrom): [34]

> (LIGHTS) The gods toss a fair coin. If *heads*, they create one subject that sees lights. If *tails*, they create three subjects, exactly two of whom see lights. (Their experiences are no more fine-grained than this.) Knowing the protocol, I wake and see lights.



In this case, the *number* of subjects who see lights is greater given *tails* (1 vs. 2). But the *proportion of* subjects who see lights—out of all subjects in the world—is greater given *heads* (1/1 vs. 2/3). (At least, this is true in the pure case where there are no other subjects in the world.)[35] In other words, if we want to treat LIGHTS analogously to BUILDINGS, we need to decide whether our intuitions in that case are driven by a comparison of numbers or a comparison of proportions.

5.1. Marbles and urns.

To motivate TYPICALITY, Bostrom often appeals to the reasonable-sounding claim that one should treat oneself as a random sample from all the subjects in the world. The heuristic that he has in mind is of the following sort. Finding yourself in existence with your CES is a bit like randomly select-

[34] Bostrom presents a number of ingenious cases in his 2001, 2002a, and 2002b.

[35] In fact, for the 'typicalist' it will also matter whether the gods themselves are in one's reference class (see §6). I will assume that they are not.

ing a predicament from all the predicaments in the world and discovering that it has this CES.

Think of one's predicament as akin to a marble selected from an urn. On that conception of things, LIGHTS is analogous to the following example—where seeing lights is analogous to being marked 'X':

> (MARBLES) A coin is tossed. If *heads*, the urn contains one marble, which is marked 'X'. If *tails*, the urn contains three marbles, two of which are marked 'X'.

Suppose I pick a marble at random from the urn, and find that it's marked 'X'. Clearly this is some evidence for *heads*— in fact, I should be 3/5 confident that *heads* came up. Why is this? Because the *proportion* of marbles in the urn that are marked 'X' is higher given *heads* than given *tails*. (As we will see, this is exactly what happens if we think of LIGHTS in terms of *proportions*.)

This seems like a very natural heuristic to offer in favor of *heads*, but it involves adherence to a certain model of the way in which we should treat our predicaments as having been selected. In fact, there is another way to imagine selecting the marble—one that provides an equally compelling analogy. The alternative approach is to think about this same case *from the perspective of the marble*. So rather than imagining yourself *selecting* a marble at random, imagine finding yourself in an urn after an uneventful marbley life. You know the set-up described above. You then notice that you are marked 'X'. What should your credences be about the coin toss? Well, a marble has to get into the urn in the first place. So you proceed as though that process involved a random selection among some pool of candidate marbles.[36] As a result, you consider yourself twice as likely to have found yourself in the urn to begin with if *tails* was tossed. But you are more likely to be marked 'X' *conditional* on being in the urn if *heads* was tossed. Taking both facts into consideration, you end up assigning 2/3 to tails. In effect, this is because the sheer *number* of marbles in the urn marked 'X' is higher given *tails* than it is given *heads*.

5.2. Weighted frequency. Suppose we prefer this second approach. At a first pass, we might try the following. For any hypothesis *h*, whether *de se* or not:

> All else equal, *h* deserves higher credence the greater the number of predicaments like mine exemplify *h*, assuming *h* is exemplified.[37]

---

[36] Things are easiest, of course, if there are finitely many.

[37] Bostrom considers a version of this principle restricted to *de dicto* hypotheses, which he calls the 'Self-Indication Assumption': see Bostrom 2002b: pp.66, 122-26; Bostrom and Ćirković 2003. Something similar is used by Bartha and Hitchcock 1999 to defuse the force of the Doomsday Puzzle. My FREQUENCY integrates the Self-

Here the *ceteris paribus* clause is crucial. We want to avoid the absurd results yielded by principles like:

> (ABSURD) For *any* two possible predicaments with my CES, I should be equally confident that I am in either of them.[38]

This treats all epistemically possible predicaments with my CES as being on a par, ignoring my (hypothetical) priors in the worlds where those predicaments live. As a result, not only does it favor *tails* in LIGHTS, it also yields much more absurd results. Suppose I know that the coin in LIGHTS is not fair, but had only a one-in-a-million chance of landing tails. Still, ABSURD would have me suspect that *tails* came up.

This problem is avoided if I weight the value assigned to each predicament with my CES by my prior probability in the world where the predicament lives. More generally, for any hypothesis H, I need to conduct a comparison between a prior *expected* number of predicaments with my CES that exemplify H, and a baseline prior expected number of predicaments with my CES. More carefully, define $n(\mathrm{H})$ for any hypothesis H as follows, where p is a credence function representing my hypothetical priors:

$$n(\mathrm{H}) = \sum_{W} p(\mathrm{W})\, n_{W}(\mathrm{H})$$

This takes, for every world, the number of predicaments that exemplify H, weights that number by the prior probability of the world, and sums the results. This yields a prior expected number of predicaments that exemplify H.[39] We can then state the updating rule very simply as:[40]

FREQUENCY: $$P_{E}\left(\mathrm{H}\right) = \frac{n(\mathrm{E}\,\&\,\mathrm{H})}{n(\mathrm{E})}$$

Indication Assumption with WS, while being precise about the *ceteris paribus* clause.

[38] See Elga (2004: 387) for a rejection of this sort of principle; Elga does not there consider the kind of modification represented by FREQUENCY.

[39] For the sake of simplicity, and to emphasize the connection with WS, I am once again setting aside the summation problem raised in the second part of fn. 9. If one were worried about this, one could avoid summing items as fine-grained as worlds to obtain the expected number of predicaments that exemplify *x*.

[40] Many thanks to Jacob Ross for helping me get clear on how best to formulate FREQUENCY.

Note that as formulated, FREQUENCY applies only when E represents one's *total* information, including apparent memories and so on, and 'p' is credence function representing one's hypothetical priors.[41]

The fully general version of this rule applies to *de se* as well as *de dicto* hypotheses—it tells us how to reason inward as well as outward. (If we simply want a rule for reasoning outward, we can restrict the rule to *de dicto* hypotheses and leave open the question of how to reason inward.) Accordingly, it converges with WS in cases like BUILDINGS: I should be 2/3 confident I'm in building 2. It also yields a 2/3 credence in *tails* for LIGHTS.[42] (Note that in a two-stage version of LIGHTS where everyone starts with their eyes shut, I will assign a credence of 3/4 in *tails* at the first stage, and then update to a 2/3 credence in *tails* when I see lights. This preserves the intuition that *I see lights* is evidence for *heads*, and can also be taken to reflect the idea that *I* was more likely to exist to begin with given *tails*.)

5.3. Weighted typicality.

The other alternative is think of lights in terms of the *proportion* of marbles in the urn. Nick Bostrom, who favors this approach, calls it the 'Self-Sampling Assumption' and summarizes it like this:

---

[41] See the discussion about COIN in §4. Using FREQUENCY as stated on one's actual priors will produce continual shifting in favor of worlds containing more predicaments with my CES; however, for those who reject the relevant intuition about COIN, we can gerrymander a rule that allows us to update sequentially using only new evidence. Let 'E*' be my actual previous CES, including my apparent memories at that time. Then we can arrive at a new credence function, for all worlds W and W', using:

$$p_E(\mathrm{H}) = \frac{\sum_{W} p_{E^*}(\mathrm{W}) \frac{n_W(\mathrm{E}\,\&\,\mathrm{H})}{n_W(\mathrm{E}^*)}}{\sum_{W'} p_{E^*}(\mathrm{W}') \frac{n_{W'}(\mathrm{E})}{n_{W'}(\mathrm{E}^*)}}$$

This feels derivative on FREQUENCY. But given some idealizations, such as a prohibition on the possibility of memory loss or duplicates with false memories, updating incrementally using this principle is equivalent to updating at every point from one's hypothetical priors using FREQUENCY. (Many thanks to Charles Sebens for help formulating a diachronic version of FREQUENCY.)

[42] In BUILDINGS, there are three predicaments with my CES in the world, two of which are in building 2; this gives us 2/3. Meanwhile, in LIGHTS, I use the chance of each outcome of the coin to set its hypothetical prior, and then compare the number of predicaments with my CES in each world. This gives me a baseline expected number of 3/2, while n(E&*tails*) is 2/2, yielding a credence of 2/3 in *tails*.

> One should reason as if one were a random sample from the set of all subjects in one's reference class.[43]

Here Bostrom does not just mean that, holding fixed the *de dicto* facts, I should prefer *de se* hypotheses according to which my predicament is a more representative sample of all predicaments. (That would just be a way of stating WS.) He also means that, other things equal, I should prefer *worlds* in which my predicament is a more representative sample of all the predicaments. More generally, for any H, whether *de se* or not:

> Other things equal, H deserves higher credence the greater the proportion of predicaments (out of all predicaments) that are like mine and exemplify H, assuming H is exemplified.

Again, we need to explicate the 'other things equal' clause. This time we need to define the notion of a prior *expected proportion of* predicaments that exemplify *x*, out of all predicaments. This will be the prior probability-weighted sum of the relevant proportions at each world. Call this '*f*' for 'fraction:

$$f(\mathrm{H}) = \sum_{\mathrm{W} \in I} p(\mathrm{W}) \frac{n_{\mathrm{W}}(\mathrm{H})}{n_{\mathrm{W}}(all)}$$

(Here the summation is restricted to *I*, the set of inhabited worlds, since the fraction would go undefined for uninhabited worlds.) We can then compare the prior expected fraction of predicaments that exemplify E and H (out of all predicaments) with the baseline prior expected fraction of predicaments that exemplify E. This gives us our principle:[44]

TYPICALITY: $$p_{\mathrm{E}}(\mathrm{H}) = \frac{f(\mathrm{E} \,\&\, \mathrm{H})}{f(\mathrm{E})}$$

Like FREQUENCY, this principle is intended to apply to *de se* and *de dicto* hypotheses alike, and it assumes that one updates on one's total CES using hypothetical priors.

It is worth emphasizing that, like INVARIANCE, both FREQUENCY and TYPICALITY entail WS—as such, they converge on comparisons of *de se* hypotheses when the *de dicto* facts are held fixed. In fact, they converge on comparisons of hypotheses whenever both the number and proportion of observers that exemplify E are held fixed, such as in BUILDINGS. But the two principles

43 See Bostrom 2001, 2002a, 2002b.

44 Once again, thanks to Jacob Ross for helping me get clear on the formulation.

diverge in cases like LIGHTS: where FREQUENCY yielded a credence of 2/3 in *tails*, TYPICALITY yields a credence of 3/5 in *heads*.[45]

Finally, note that both principles create what I will call an *existential selection effect*: this is when *my* having this CES has evidential bearing on *de dicto* hypotheses, beyond entailing the *de dicto* fact that someone has this CES. In other words, both principles yield credences in *de dicto* hypotheses that diverge from the result of updating one's hypothetical priors with one's strongest *de dicto* evidence.

5.4 Sleeping Beauty.

Because of its prevalence in the literature, it is worth noting how our two principles (as well as invariance) treat this famous case involving potential diachronic duplication of one's evidential state:

> I will be put to sleep on Sunday, and a coin will be tossed. If it comes up heads, I will be woken on Monday morning only. If it comes up tails, I will be woken on Monday morning and on Tuesday morning as well—but between the two wakings, my memory of the first waking will be erased.

Let's start with my *de se* credences about the day of the week, conditional on *tails*. This feels very much like COIN. If *tails*, there is a time-slice of me on Monday, and a time-slice of me on Tuesday, both of whom wake with no memory of a previous waking. I might be in either situation, so it is tempting to split my credences between them—in effect, treating my current *de se* situation as a random sample from among those two time-slices.

But how should I weigh these against the time-slice that wakes given *heads*? Naturally, the invariantist resists any outward reasoning and takes the 'halfer' position that Beauty should assign equal credences to heads and tails. But if I treat myself as randomly sampled from the three predicaments with my CES—after all, each is just as likely to occur as any other—I'll assign 2/3 credence to *tails*. This makes the frequentist a solid 'thirder'.[46]

---

[45] My baseline expected proportion of predicaments with my CES is 3/4, while the proportion exemplifying *I am in building 2* is 2/4. So, as with FREQUENCY, the result is a 2/3 credence that I am in building 2. In LIGHTS, all the predicaments have my CES given *heads*, while only 2/3 have my CES given tails. (I will assume that there are no other subjects in the universe; this matters to TYPICALITY.) So factoring in the equal hypothetical priors on the two coin outcomes, I end up with 1/2 over 5/6, or a 3/5 credence in *heads*.

[46] At least, that is, if the frequentist treats two predicaments of the same observer with the same CES at different times in the same way that she treats two predicaments of different observers with the same CES at the same time.

For the typicalist, however, the example is actually underdescribed—we must know how many predicaments *unlike* Beauty's there are. At one extreme, Beauty is the only life form in the universe and she is only conscious during the two wakings given *tails* or the single waking given *heads*—then the typicalist will be a 'halfer'. After all, on either outcome, all predicaments have the same experience. But the typicalist will tend towards the 'thirder' position as the number of predicaments unlike Beauty's increases, because this will increase the difference in the typicality of her CES between the two outcomes. (For example, if there is one 'bystander'—a predicament unlike her waking CES—the ratio of predicaments like hers to those not like hers is 1/2 given heads and 1/3 given tails.)

6. Outward: FREQUENCY versus TYPICALITY

Let us now examine some considerations to which one might appeal in deciding between FREQUENCY and TYPICALITY.

6.1. Constraining conditional priors?

Bostrom is clear that his 'random sample' talk is metaphorical:

> There is no intimation of any physical randomization mechanism—some kind of stochastic time-traveling stork?—responsible for distributing observers in the world. [The Self-Sampling Assumption] should be read as a methodological prescription specifying certain types of conditional credences of the form P(I am such and such an observer | The non-indexical properties of the world are such and such). The phrase "as if one were a random sample" is simply shorthand for these recommendations.[47]

Presumably the relevant conditional credences will have to be *hypothetical* priors, since in many of the cases Bostrom is interested in, the subject was simply not around to have the relevant priors.

As we have formulated FREQUENCY and TYPICALITY, they do not operate on prior conditional credences about *de se* evidence, only on prior *de dicto* credences about worlds and the individuals they contain. But at an informal level, it is natural to invoke something like prior conditional credences in *de se* hypotheses. Thus in LIGHTS, the typicalist appears to reason as though her priors guaranteed that *she* would be created by the incubator regardless of the outcome of the coin toss. Meanwhile, the frequentist appears to reason as

---

[47] Bostrom 2003: 84. Or rather, as Bostrom himself stresses, the relevant credences would concern which predicament or 'observer-moment' I am in.

though her hypothetical priors treat her creation as more likely the more subjects are produced. Is either set of hypothetical credences more plausible? The first approach would seem right if I were some kind of a haecceity that was guaranteed to be embodied regardless of how many subjects the incubator produced. The second would seem right if I were a haecceity that had an equal chance of embodiment for every subject produced. But since presumably I am neither, we have another inconclusive heuristic.

If we were to take this heuristic seriously, though, it would indicate something strange about the typicalist approach. Consider LIGHTS and suppose that, unbeknownst to me, the coin came up *tails*. In that case, there is another subject who sees lights and is wondering how the coin toss came up. If we are both typicalists, then we will both reason as though we would have observed *something* either way, but of course we can't both be right![48] Or consider the following case:

> (TWINS) If *heads*, two people with my CES are made; if *tails*, four people with my CES are made. In addition, everyone expects to meet exactly one other person. (The meetings are randomly arranged in case of *tails*.)

Suppose I have equal hypothetical priors for *I have this CES* conditional on each outcome, and thus I assign equal posterior credences to *heads* and to *tails*. When I meet my match—call him 'Phil'—this clearly should do nothing to change my credences. But what should my hypothetical priors be in *Phil's having this CES* conditional on either outcome? Presumably these should also be equal—it would be odd to reason as though *I* was equally likely to be produced on either hypothesis but *Phil* was not. But then the prior probability of meeting Phil should be much higher given *heads* than given *tails*—since I was certain to meet him given that heads came up and we were both produced. So meeting Phil should be evidence for *heads*!

The typicalist ought to resist this line of reasoning. She could, for instance, back off from the idea that TYPICALITY can be characterized as constraining hypothetical conditional priors involving *de se* beliefs, and stick with TYPICALITY as an updating rule that operates only on *de dicto* priors. Or she can deny that we should update using conditional *de re* credences—except insofar as *de se* credences are themselves *de re*. But while these considerations about Phil are far from conclusive, I do find them suggestive.

---

[48] There are counterpart theorists who would deny this, holding that both thoughts could be true if they invoked a loose enough counterpart relation. Such a view would certainly complicate the idea that there are rational constraints on credences that can be cashed out in terms of hypothetical credences of the sort described in the text; after all, whether I consider it certain that I would exist on either outcome will end up turning on which counterpart relation is operative.

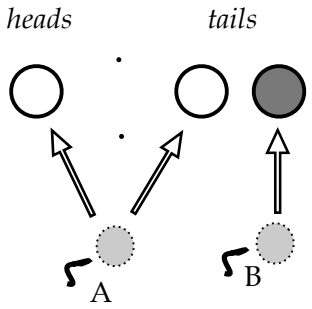
6.2 Guaranteed existence

There is, then, something intuitive about FREQUENCY's preference for worlds containing a greater raw number of predicaments with one's CES. A more surprising benefit is that FREQUENCY, rather than TYPICALITY, properly handles certain cases where one's guaranteed existence is actually built into the protocol.

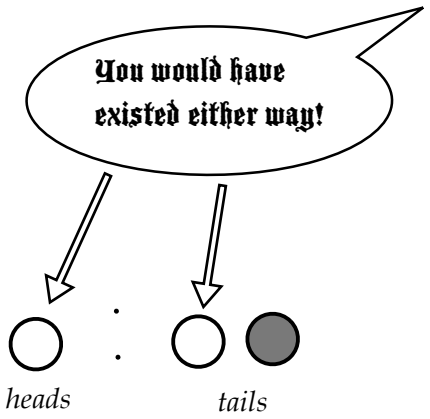
Consider an incubator case where the number of subjects differs between *heads* and *tails.* In such a case not everyone can learn that they would have existed on either outcome of the coin toss. If everyone is told they would exist either way, this testimony would be undermined by the subjects' knowledge of the protocol; however, there are a variety of cases in which *some* individuals get evidence that they would have existed either way. For example, consider:

So for example, suppose that again you find yourself seeing lights, but this time it's built into the protocol that seeing lights entails that you would have existed either way.

> (GAMETES) In the beginning there are two sperm-egg pairs, A and B. If *heads*, only pair A will be incubated. If *tails*, both A and B will be incubated. But only the subject resulting from pair A will see lights. I wake as a result of this process and see lights.



Suppose, further, that being the developed result of the A-pair is necessary and sufficient for being me. Or consider a case where seeing *lights* is replaced with God telling me that I would have existed regardless of the outcome of the coin toss:



Now, TYPICALITY instructs me to assign a higher credence to *heads* than to *tails* in both of these cases, because if *heads* came up, my CES is more representative of all predicaments. But in fact, it seems obvious that I should assign equal credences to *heads* and *tails* in these cases. After all, I am certain that I would have existed and had these very experiences on either outcome. Surely that is the sort of case, if ever there were one, where my *de se* evidence should have no effect on my *de dicto* priors. And this is precisely what FREQUENCY recommends.

In short, if we actually build a guarantee of existence into the protocol, rather than taking it for granted as TYPICALITY intuitively does, we often end up with cases that FREQUENCY gets right and TYPICALITY gets wrong.

6.3 The problem of subjecthood

TYPICALITY requires me to compare the set of predicaments like mine with the set of *all* predicaments—what Bostrom calls 'the reference class'. But what must something be like to count as a subject? For example, do dogs count? How about turtles? In order to implement the principle, we will often need an answer to this question.

This issue does not arise for FREQUENCY, because it concerns only the number of predicaments *with my CES*. For that reason, we don't need to decide whether anything counts as a subject—all that matters is whether it has my CES. To illustrate this point, consider the following case:

> (DOG) If *heads*, the incubator produces a creature with my CES; if *tails*, it produces a creature with my CES and a dog with a doggy CES.

For the frequentist, this is easy. The outcomes deserve equal credences. There is no need to decide whether dogs count as subjects, and it doesn't matter what the creature with my CES is like 'from the outside'. The typicalist, on

the other hand, needs to decide whether dogs are sufficiently subject-like. If she includes the dog in her reference class, she will prefer *heads*; if not, she will assign equal credences to the two hypotheses.

Is there any non-arbitrary way for her to decide? Consider a sorities series of millions of cases like DOG, except that in each case the dog is replaced with a more human-like creature until at last it's another human being. At some point in this series, the typicalist must stop recommending a credence of 1/2 in *heads* and start recommending a credence of 2/3, because TYPICALITY does not allow for intermediate credences. Of course, we could build a cut-off point into the principle, but the result would seem too arbitrary to have a very good claim to constraining rational credences.

Bostrom's suggests that there is 'a subjective factor in the choice of reference class'—the principle need not single out a 'uniquely correct credence function' (2002a: 182). In other words, the rule *requires* me to implement TYPICALITY with some reference class or other, but gives me leeway about which reference class to choose; however, even this doesn't avoid the problem. Presumably it would not be rationally acceptable, for example, to include plankton or tomato plants in my reference class. So we face a new question: what are the boundaries on acceptable choices for a reference class?[49]

The typicalist might view this objection as unfair because it trades on the fact that the term 'subject' or 'predicament' is not fully precise. After all, she might say, confirmation theory typically operates in an idealized setting where one's hypotheses and credences are fully precise. And in that setting, for example, questions like 'what credence should we give to the claim that *x* is bald when *x* is a borderline case of 'bald'?' simply do not arise. However, there is an important asymmetry here. The problem of subjecthood does *not* go away even if I imagine formulating hypotheses with a great many precise predicates rather than vague predicates like 'subject'. Even in such a setting I would still have to decide which class of objects to include in my reference class. The problem is not solved by switching to a setting where all the terms are fully precise; it is as pressing as ever.

[49] Similar problems seem to arise if we try to make the principle somehow graded. For example, suppose we say that the type of epistemic norm in play is one that comes in degrees—so that, the closer the mental faculties of a creature are to those of a normal adult human being, the less reasonable it is to treat the creature as outside one's reference class. But this raises the question: what do *fully* reasonable subjects consider to be their reference class? Neither does it help to say that borderline creatures can count as *fractions* of subjects, so that the less aware and intelligent a creature is, for example, the smaller a fraction it deserves. For now we must decide when creatures stop counting as full subjects, when they stop counting as any fraction, and which fractions correspond to apes and antelopes.

It seems the typicalist will have to fall back on the claim that in a wide variety of cases, it's vague which credences are rationally acceptable for even an ideal subject to have. This is not a decisive problem, especially if it turns out that vagueness in epistemic normativity is unavoidable; however, the range of borderline cases seems especially significant for the typicalist—so much so that it is entirely unclear how to apply the rule any time two hypotheses differ on the number of animals that exist. Perhaps, as Bostrom hopes, the problem of subjecthood is an enigma that will yet be made clear by further reflection or argument (2002a: 205). But avoiding this thorny issue altogether is a *prima facie* benefit of FREQUENCY.

6.4 The prediction problem

Setting aside the question of what counts as a subject, TYPICALITY also faces a dilemma about whether *future* subjects should be treated as members of the reference class. Consider this case, from Bostrom (2001, p. 367):

> (EDEN) Adam and Eve are the only subjects in the universe, and know that if they have children, the world will fill up with their descendants; and if not, there will be no other subjects. They toss a coin and take an unbreakable vow to have children only if it comes up *tails*.

We can suppose that none of Adam and Eve's descendants will have exactly their experiences. If they include any future descendants in their reference class when considering the outcomes of the coin toss, TYPICALITY will cause them to be very confident that *heads* will come up! After all, each should reason that the proportion of subjects with his or her CES will be *much* higher if they have no descendants. As a result, their credences will hugely diverge from what they know to be the objective chance of the outcome. Moreover, as Bostrom himself points out, they could rationally predict nearly any event by tying it to a firm intention about whether or not to have children—for example, if they are hungry they could agree to have children only if a wounded deer enters their cave. They would then be nearly certain of an easy dinner—a crazy result.

Crucially, there is no analogous problem for the frequentist. Admittedly, FREQUENCY can be exploited to make Adam and Eve favor one outcome from a future coin toss. But this can only be done in such a way that they are no longer certain that the coin toss is in the future. As a result, FREQUENCY will not cause them to diverge from what they take to be the current objective chance of a given outcome. To illustrate this point, consider the following variant on the story:

> (EVE) Eve is alone in the world at t1. The gods are about to toss a coin. If *heads*, they will do nothing. If *tails*, they will produce many subjects at t_2 and give them all the very CES that Eve had at t_1.

Since the expected number of observers with her CES is much higher given *tails*, FREQUENCY requires Eve to predict that the coin comes up *tails*. But there is a crucial disanalogy here. If Eve is a frequentist, she will suspect that *it is already t2* and that she is one of the many individuals produced by the incubator and given false perceptions and memories after the coin toss came up *tails*. As a result, she should suspect that she is not making a *prediction* about the coin toss at all—even though in fact she is. In fact, her additional credence in *tails* all stems from epistemic possibilities in which *tails* has already occurred and therefore has an objective chance of 1. Accordingly, her credence in *tails* will still match her expectation of its objective chance; she does not violate the Principal Principle.

In short, there is nothing counterintuitive with the frequentist's treatment of EVE, and this is in sharp contrast with the typicalist's treatment of EDEN. In the latter case, Adam and Eve have no doubt about whether the coin toss is in the future, or about the veracity of their memories. Requiring that they have near-certainty in *heads* just seems crazy. At times, Bostrom seems willing—even eager—to bite this bullet. But he also claims that the typicalist could avoid the prediction problem by *excluding* future predicaments from her reference class.[50] I argue in Appendix 2 that this leads to equally awful results.

6.5. Doomsday.

A famous instance of the prediction problem is the Doomsday puzzle.

> Suppose there are only two possibilities. On one, we survive another million years, and the complete history of the universe contains 200 trillion people. On the other, we go extinct very soon due to a great calamity, and there will only have been 200 billion people. To make things simpler, suppose that Fate tossed a coin at the beginning of the world to decide between these histories.

Now, consider the fact that we find ourselves among the first 200 billion people. Does this, all by itself, support either doom hypothesis? Some say that it does.[51] Think about it this way. Given *Doom Late*, a very small fraction of observers would be among the first 200 billion people. But given *Doom Early*, they all would! If we treat ourselves as a random sample from among all observers in history, then discovering that we are among the first 200 billion is far more likely given *Doom Early* than given *Doom Late*.

Scary! But only one of our three ways of fleshing out the random sample heuristic yields this result. Once again, the invariantist will be unmoved by the

[50] See Bostrom 2001, p. 381; 2002a, chs. 9 and 10.

[51] See Bostrom 2001, 2002a; Leslie 1989, 1996.

outward reasoning, since it goes beyond entailment links. If Fate tossed a coin, one's credences should remain 50/50.

Meanwhile the argument that we should think doom is near appeals specifically to typicalist reasoning—the proportion of individuals that have my CES is a thousand times higher if there are 200 billion total than if there are 200 trillion total. The typicalist should be very worried; but the frequentist should not be, as Bartha and Hitchcock have shown.[52] We can break down her reasoning into three stages. First, there is the coin toss—this sets the hypothetical priors. Next, there is the fact of one's existence, prior to 'opening one's eyes'. For the frequentist, this gives her evidence in favor of Doom Late—in effect, the greater the total number of individuals, the more likely she is to exist in the first place! And finally, she takes her evidence into account and finds that she is among the first 200 billion. There are just as many people having this sort of experience on either hypothesis, so her credences revert to 50/50.

A very similar result differentiates TYPICALITY and FREQUENCY when it comes the the question whether the 'fine-tuning' of the universe is evidence for the existence of many universes: I've saved that discussion for Appendix 2.

6.6. The 'presumption' problem

Given all of this, why does Bostrom prefer TYPICALITY to FREQUENCY? The main reason he offers involves the following case:

> (PRESUMPTION) 'It is the year 2100 and physicists have narrowed down the search for a theory of everything to only two remaining plausible candidate theories, T_1 and T_2… According to T_1 the world is very, very big but finite, and there are a total of a trillion trillion subjects in the cosmos. According to T_2, the world is very, very, *very* big but finite, and there are a trillion trillion trillion subjects. The super-duper symmetry considerations are indifferent between these two theories. Physicists are preparing a simple experiment that will falsify one of the theories. Enter the presumptuous philosopher: "Hey guys, it is completely unnecessary for you to do the experiment, because I can already show to you that T_2 is about a trillion times more likely to be true than T1!" '.

The example can be strengthened by fixing some additional background. (For instance, it helps to stipulate that the expected number of predicaments like mine increases with the total expected number of subjects, and perhaps that

[52] Bartha and Hitchcock 1999; see also Kopf, Krtous, and Page 1994/2012.

whether T_1 or T_2 obtains turns on some random occurrence early in the Big Bang that had an objective chance of .5.)[53]

The result is counterintuitive, but is it a sufficient reason to prefer TYPICALITY over FREQUENCY? If I am a frequentist, I will (all else equal) prefer theories where there are *more* predicaments like mine. And if I am a typicalist, I will (all else equal) prefer theories where there are *fewer* predicaments *unlike* mine. Both results can be made to seem extreme when we are considering very large numbers. After all, consider:

> (PRESUMPTION3) As in PRESUMPTION except the relevant theories are T_3, which says there are a trillion non-green subjects in the universe and a trillion trillion green subjects, and T_4, which says there are a trillion of each.

The typicalist, having noticed that she's non-green, will declare it completely unnecessary to test these theories empirically, because T_4 is a trillion times more likely than T_3. This seems pretty presumptuous as well.

Which type of presumption is worse? I seem to be able to get into both frames of mind, each one governed by one of the two models for treating one's evidence as a random sample and illustrated by one of the two marble metaphors discussed in §5. But there is no denying that both results are unsettling as the numbers get arbitrarily high.

It might seem that TYPICALITY is better off when we think about such cases. After all, the result for TYPICALITY is that very bizarre and profligate worlds get all but *ruled out*, whereas the result for FREQUENCY is that very bizarre and profligate worlds get all but *ruled in*. Keep in mind, however, that what makes these worlds bizarre (if they *are* bizarre) is the sheer number of experiencing individuals—and the fact that these are bizarre worlds should be operative at the stage of hypothetical priors, where Occhamistic inclinations are properly in play. The idea that my purely *de se* evidence should shift my credences *further* in favor of certain simple worlds—in particular, those that are simple with respect to the number of individuals unlike me—is arguably just as odd as the idea that my evidence should shift my credences in favor of certain complex worlds— those that maximize the number of evidential states like mine.

---

[53] In addition, it may help to control for any prior bias in favor of hypotheses that are more ontologically parsimonious, which might balance out the effect of FREQUENCY. To this end, we could treat T_2 as a hypothesis on which the universe contains the same total number of objects, but still has many more *subjects*. See Bostrom and Ćirković 2003.

7. Outward: the breaking point

As usual, infinity ruins everything. Here there are a few separate problems, the first of which is an infinite variant of the presumption problem.

(i) Suppose the hypothesis that there are infinitely many individuals with my CES has nonzero probability at the stage of hypothetical priors. Then when I attempt to update on my CES, FREQUENCY collapses because its values become undefined. And this problem is worse than the additivity problem faced by WS. We can't simple become permissive about credences in such worlds, or allow our credences about them to go undefined or imprecise. Because unconditional *de dicto* credences are at issue, there is no obvious way to quarantine the weirdness so that it does not infect all of our other credences.

In fact, it's hard to see how the principle could be fixed without (a) insisting on priors of zero for the relevant worlds; (b) imposing an implausible exception for infinitary cases; or (c) accepting that we should be certain that there are infinitely many predicaments with our CES. None of these seem like very good options. To require a credence of zero in the relevant worlds seems quite harsh given that working cosmologists take such possibilities seriously.[54] And simply excluding infinitary worlds from the principle appears completely ad hoc. After all, for all finite numbers, the amount of evidence we get for the duplication hypothesis increases as the number of duplicates in each world goes up. If, instead, we tried to impose a mere 'cap' for how much evidence we get for such an infinite world, there would be some finite world for which we get *more* evidence! Any choice of where the cap should go would seem completely arbitrary.

How much better off is TYPICALITY on this score? The analogous problem for TYPICALITY involves worlds with finitely many individuals like me and infinitely many unlike me. Regardless of how probable this is at the stage of hypothetical priors—as long as it is not certain—I will become certain that it is false when I update on my evidence. Some will intuit that this is not quite as bad as the problem for FREQUENCY, but they are both pretty bad. Admittedly, this result involves *ruling out* an ontologically profligate world with certainty rather than *ruling one in*—but again, we should already have adjusted for any bias against ontological profligacy at the level of our *priors*. The problem—both for TYPICALITY and for FREQUENCY—is the *certainty* about these infinitary worlds engendered by ordinary evidence.

---

[54] Of course, if we allow primitive conditional probabilities, assigning H a prior of zero doesn't *rule it out* in the sense that one could never get evidence for it.

(ii) A second problem involves trying to make comparisons *between* worlds in which there are infinitely many individuals (both with my CES and without).[55] This is particularly salient due to the rise of 'big-world' cosmologies, both of the 'multiple universe' variety associated with inflationary theories, and of the 'many worlds' variety associated with the Everettian interpretation of quantum mechanics. If there are infinitely many individuals, this appears to break down the notion of proportion that is required by both principles. But the hope is that we needn't give up on the ability to distinguish between such hypotheses using our *de se* evidence.[56]

A natural solution is to appeal to some measure on the space of individuals that can impose a proxy for talk of proportions. For example, drawing from a trick discussed earlier, we can imagine starting with arbitrarily small spheres and increasing their volume until they are arbitrarily large—all the while seeing what proportions of individuals in the sphere have my CES. If this process tends to converge on a proportion of 1/2, then we treat that whole world as having a proportion of 1/2. This might work for TYPICALITY but it is not clear that anything of this sort could be done for FREQUENCY. (One could try preferring hypotheses that are *denser* with respect to the distribution of individuals like me, but that seems implausible.) On this issue, TYPICALITY appears to have the upper hand.

(iii) A third problem is this. There are cases where FREQUENCY generates undefined credences even without any positive credence in infinitary worlds. Suppose my hypothetical prior in the hypothesis that there are N individuals with my CES is proportional to $1/N^2$. As it stands, this should not be problematic— it might be a natural way to implement Ockhamistic tendencies while ensuring that my credences converge to 1. Unfortunately, though, when I take into account my evidence, FREQUENCY will instruct me to sum 1/N for every N in order to arrive at n(E). But this sum does not converge, yielding an infinite denominator and plenty of undefined credences.[57] Perhaps we have to live with the fact that for some sets of priors, FREQUENCY will offer no useful guidance. But it seems strangely arbitrary that it would go undefined for a natural prior distribution like the one just sketched.

---

[55] Thanks to Cian Dorr for pushing me on this problem, and suggesting the solution for TYPICALITY that I discuss.

[56] See, e.g., Smeenk 2014, Adlam 2014.

[57] Thanks to Cian Dorr for this point.

8. Conclusion

We have examined formalizations of the random sample heuristic applied to both inward and outward reasoning. In particular, we considered three principles that can replace the standard updating rule in a way that handles both *de se* and *de dicto* evidence. The most conservative of these blocks any outward reasoning beyond entailment links. But that principle—INVARIANCE—yields highly counterintuitive results in a variety of cases: withholding evidence where intuitively one should get it; granting evidence where intuitively one should not; yielding different credences in cases that seem evidentially the same; and leading to egregious violations of Reflection.

Unfortunately, the two principles that allow outward reasoning beyond entailment links have problems as well: they both face versions of the 'presumptuous philosopher' problem, and they both have trouble with infinity. (Here TYPICALITY seemed a bit better off than FREQUENCY, especially in that it's better suited to exploit natural measures like distance.) But TYPICALITY faces a battery of problems of its own: (i) it generates a suggestive asymmetry between how one treats hypothetical credences about one's own existence and other people's existence; (ii) it requires making seemingly arbitrary decisions about what sorts of creature count as subjects; and (iii) it faces the prediction problem and a resulting violation of the Reflection principle.

Given all this, what should we say about outward reasoning beyond entailments? We can't simply live reject it entirely—that is what INVARIANCE counsels. But neither have we found a fool-proof way to live *with* it. So should we turn to a thorough-going permissivism about outward reasoning beyond entailments? That is far from obvious: we would avoid the claim that any one set of bad consequences is obligatory, but we would be forced to admit that *all three* sets of bad results are permissible! (However, if the inadequacies of all three principles do lead us to accept complete permissivism about outward reasoning beyond entailments, we finally have our answer to the various puzzles in confirmation theory that involve such reasoning: anything goes!)

As for those of us who suspect there *are* constraints, these results should give us pause. In the absence of coherent general rule, our intuitions about what a subject should do in this or that case will hardly be conclusive. After all, there may be no way to generalize on those intuitions without running into highly counter-intuitive results. In addition, sometimes our intuitions are too imprecise to motivate the conclusion we seek. For example, random-sample heuristics have frequently been used to motivate the conclusion that doom is near or that fine-tuning is evidence for many universes—but it turns out only TYPICALITY, not FREQUENCY, yield those results. (See Appendix 1.) And it is at best far from clear that TYPICALITY is a better way to spell out the random sample heuristic.

Appendices

1. Fine-tuning

As is well known, a number of philosophers and physicists have claimed that the alleged fine-tuning of the universe is evidence for the existence of many universes.[58] By 'the alleged fine-tuning' I mean the claim that various constants in our physical theory (e.g. the mass of the proton, the strength of the weak electromagnetic force, the strength of gravity) could easily have varied very slightly, and that if any of them had done so, the universe would not have been hospitable to life.

The idea is that, if we learn that the chance of a given universe producing life is a lot lower than we used to think, we should increase our credence in the hypothesis that there are many universes. Typically the approach is to treat the fact that *life exists* as evidence (presumably *old* evidence) that is assessed first against a background theory according to which life is likely to arise in a given universe, and then against a background theory according to which life is *unlikely* to arise in a given universe. And whatever credence for one ended up with for the multiple universe hypothesis after the first assessment, one's credence in that hypothesis should be significantly higher after the second.

Since the evidence that *life exists* (or the more specific evidence that life of such-and-such a kind exists) is *de dicto*, this argument works even for the invariantist. But notice what happens when we are frequentists. To make things simple, assume the Big Bang either produced one universe or a trillion of them through an 'inflationary' expansion. Call these outcomes *One* and *Many*, and suppose that at the outset each had a 50% objective chance of obtaining. I start out in the naïve state, thinking that the chance of a given universe producing life—indeed, its chance of producing my CES, which is really what counts for the frequentist—is pretty high. For the moment let's use n for this value. (I'm assuming for simplicity that there is at most one subject with my CES per universe.) Now, in this naïve state, I assess *One* and *Many* as follows.

The expected number of predicaments with my CES that exemplify *Many* will be my prior in *Many* (½) times the expected number of predicaments with my CES given *Many*, which is a trillion over n. Meanwhile the expected number of predicaments with my CES that exemplify *One* will be my prior in *One* (½) times the expected number of predicaments with my CES given *one*, which is $1/n$. Plugging all of this into FREQUENCY yields a very high credence in *Many* (see the figure below). This in itself should perhaps be unsurprising,

[58] e.g. Leslie 1988; van Inwagen 1993; Parfit 1998; Smart 1989.

since we have encountered cases like PRESUMPTION where FREQUENCY yields a very high credence in a hypothesis that predicts many instances of my CES.

$$p_E(\mathit{Many}) = p_E(\mathrm{H}) = \frac{n(\mathrm{E}\,\&\,\mathrm{H})}{n(\mathrm{E})} = \frac{\frac{10^{12}}{2n}}{\frac{10^{12}+1}{2n}} = \frac{10^{12}}{10^{12}+1}$$

$$\frac{1}{2}\times\frac{10^{12}}{n} \qquad \left(\frac{1}{2}\times\frac{10^{12}}{n}\right)+\left(\frac{1}{2}\times\frac{1}{n}\right)$$

Now, when we learn that the universe is fine-tuned, we assign a very low value for the chance of a given universe producing my CES—namely '*n*'—and reassess the evidence in light of that fact. But note that '*n*' cancelled out of the equation above. As a result, learning the fine-tuning evidence makes no difference to my credence in *Many*. In effect, I start off preferring *Many* because it makes my existence a trillion times more likely than *One* does. But this ratio does not change if I learn that the chance of my CES being produced in a given universe is lower than I thought—both sides of the ratio drop by the same factor.

If this result seems strange, consider this analogy, inspired by one that Bartha and Hitchcock use while defending a very similar result applied to the Doomsday argument.[59] Suppose you go to the mailbox and find an envelope that reads:

> This envelope either contains $1 or $1 million! We flipped a coin. If *heads*, we randomly chose 1 person from the phonebook and sent them a $1 million check. If *tails*, we randomly chose a million people from the phonebook and sent them all $1.

Suppose you trust what's written on the envelope. Now you could proceed by updating on either of the following two bits of evidence:

E_1: Someone got a notice in the mail.

E_2: *I* got a notice in the mail.

[59] Bartha and Hitchcock 1999 apply something like FREQUENCY to defang the doomsday argument. See also Dieks, 1992. For an early discussion of the Doomsday argument, see Leslie 1996.

If you simply conditionalize on E_1, you will assign equal credences to the outcomes of the coin toss, which means that there's an even chance you are holding a check for a million dollars! But this is clearly not the right response to this case: the fact that you got a letter at all is far more likely given *tails.*

According to the frequentist, this is a good analogy for the comparison of *Many* and *One* in the naive state. Now suppose you learn, to your surprise, that in the last week the mail service has been extremely unreliable. The chance of you getting any letter at all was 1/n for some high 'n'. Should this make any difference to your view about whether you're holding a million dollars? If you were simply using E_2, then you would consider this a great deal of evidence for *tails*— after all, the fact that someone got one of the envelopes is now much more likely given *tails*.

But clearly this is the wrong way to reason about the coin toss. The unreliability of the post office in fact gives you no new evidence for *tails*. It simply has the effect of lowering, by the same factor, the chance of your getting an envelope given either outcome. The ratio between the resulting values has not changed. And this, according to the frequentist, is a good analogy for learning that the chance of life in a given universe is extremely low.

2: Typicality and future subjects

In response to the prediction problem discussed in §6, Bostrom has suggested that future subjects be excluded from the typicalist's reference class. It is not obvious how to apply this idea to every case, such as one in which two *de se* hypotheses disagree about whether a given predicament is in the future. But things are more straightforward in the Adam and Eve case. At a minimum, the idea seems to include this:

> (EXCLUSION) If one is certain that a predicament obtains in the future—if it obtains at all— it is not numbered among the total number of predicaments in a world, i.e. n_W(all), when applying TYPICALITY.

But this is not much help, for a few reasons.[60]

2.1 Creation, execution, reflection.

Suppose I know that the incubator is about to toss a coin. If *heads* is tossed, it will do nothing. If *tails* is tossed, then at t_2 it will create a subject that is unlike me at any time. Either way, I will get no qualitative evidence about the

[60] I will set aside worries involving the relativity of simultaneity.

outcome. A typicalist can save the intuition that I should have equal credences in *heads* and *tails* by invoking EXCLUSION, which tells me that at t_1 I should exclude the subject that may be produced at t_2 from the value of n(all) in the tails world.

However, at t_2 that predicament would no longer be future relative to me, so as a result I will suddenly suspect that *heads* came up. And this shift in credences will occur despite the apparent lack of relevant evidence—in fact, at t_1 I could have *predicted* that I would shift my credences about the coin toss at t_2. More generally, whatever I think the objective chances are about someone successfully procreating, in the absence of evidence about their success I should revise my expectations downward around the time that I would expect the new being to count as a subject.

Meanwhile, there is an inverse effect involving chances of death. Suppose there are only two subjects (me and someone evidentially unlike me) and a coin has been tossed: *heads*, the other subject is suddenly executed at t_2; *tails*, nothing happens. (I will get no qualitative evidence one way or the other.) Assuming my current CES never gets repeated, *heads* involves fewer predicaments that are unlike mine, so the original typicalist suspects that heads will be tossed *at the outset*. (In fact, this is a simple version of the 'Doomsday argument' for typicalists.)[61] Meanwhile, following EXCLUSION, I will make no such prediction, but I will suddenly and predictably begin to suspect at t_2 that the other subject has been executed.

Perhaps these predictable shifts to pessimism are better than pessimism from the outset, since the latter involves violating the Principal Principle. In fact, while the exclusionist admittedly violates van Fraassen's Reflection Principle in such cases,[62] so does the frequentist in cases like EVE; so it might be tempting to treat these results as equally bad, but that would be a mistake. All friends of WS must admit that some violations of Reflection are acceptable: for example, witness the violation in DUPLICATE. But what's special about cases like DUPLICATE is that the subject comes to doubt the veracity of her memories and thus is not sure that she is violating the principle at all. And this is precisely what happens in EVE. These cases are intuitively among a range of cases where a violation of Reflection is acceptable, such as cases where one has reason to believe one has become cognitively impaired, brainwashed, or lost one's memory. Nothing of this sort at all is going on in an ordinary case where one knows that a couple is attempting to reproduce and their chances of success are N, or that there is a chance of N that someone will be executed.

---

[61] See fn. 60 below.

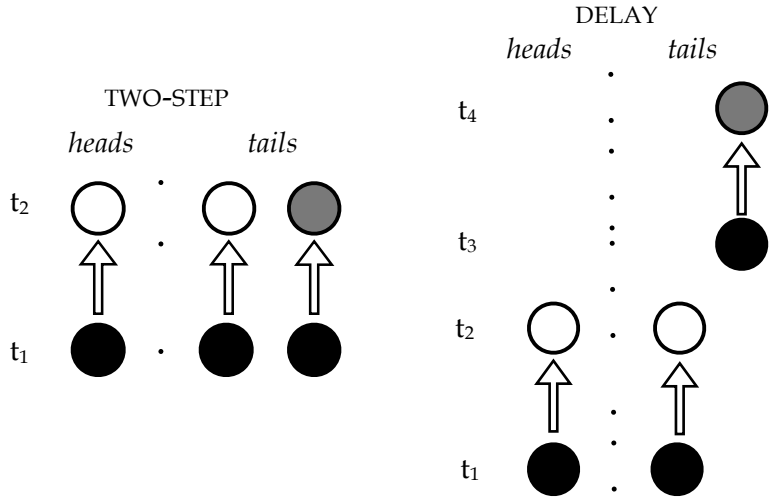
[62] See fn. 33 above.

But EXCLUSION requires a violation of Reflection in precisely such a case, even if one is perfectly self-aware about the violation.

2.2 Phantom evidence.

It is not only in creation and execution cases that EXCLUSION yields bizarre shifts in credences. Consider:

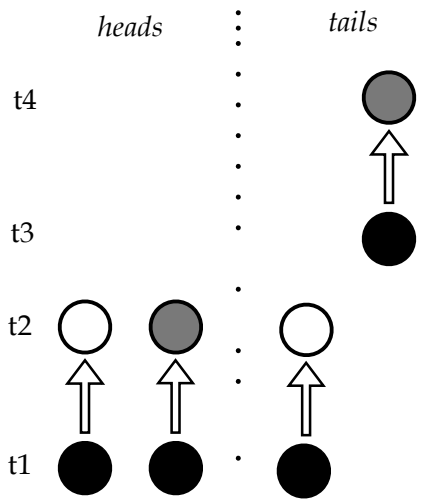
> (DELAY) Just like TWO-STEP except that if *tails*, the subject for whom the lights are off is not produced until after the other is gone.



Suppose I am a subject in DELAY. I awake with my eyes closed. At this point, the original way of applying TYPICALITY tells me I should assign equal credences to *heads* and *tails*; however, EXCLUSION tells me not to include the predicament at t_2 in the value of n(all) for *heads*, and not to include the predicament at t_4 in the value of n(all) for *tails*. (Given *tails*, I cannot assume that the predicament at t_2 is in the future.) As a result, I will assign a higher credence to *heads* at the outset: a credence of 3/5. And when I open my eyes and see that the lights are on, I will shift to having equal credences in *heads* and *tails*, because I must exclude the subjects at t_3 and t_4. This is just bizarre. Intuitively the case should be just like TWO-STEP, where the typicalist (with or without EXCLUSION) starts off with equal credences and then gets evidence for heads. But in DELAY, the exclusionist starts off preferring *heads* and then gets evidence for *tails*!

It might be tempting to revise EXCLUSION so that one can also exclude predicaments that one is certain obtain in the *past*, if they obtain at all. This at least will avoid the result that I get evidence for *tails* when I see lights in DELAY. But, to begin with, it yields the result that I get no evidence one way or

the other, which is still wrong. Moreover, consider this variant of DELAY in which there are two subjects on each hypothesis:



Intuitively, in this case no one gets any evidence that bears on the coin toss. But our new version of EXCLUSION has me treat seeing lights as evidence for *tails*. (At the outset I exclude all eyes-open predicaments; and after seeing lights I exclude all predicaments that are not at t_2.) In fact, I get evidence for *tails* even if the lights are off, so while I start off with equal credences in the coin toss, I can be sure that I will end up with a higher credence in *tails* no matter what happens: another egregious violation of Reflection.

What these cases show is that typicalists cannot plausibly treat location in time as importantly different from location in space for purposes of calculating n_W(all).

3: The Many Brains Problem

Here is an alleged problem for principles like FREQUENCY, due to Tim Maudlin and discussed by Chris Meacham:

> Consider the hypothesis that you're a brain in a vat…. Your current credence in this possibility… is presumably very low. Now consider the proposition that you're in a world where brains in vats are constantly being constructed in states subjectively indistinguishable from your own. Let your credence in this proposition be $0 < p < 1$, and your credence that there will be no multiplication of doxastic alternatives be $1 - p$.

The worry is that principles like FREQUENCY will lead all of us to increase our credence in this strange hypothesis—and in fact, according to Meacham, our credence should converge to 1. Note that this kind of argument does not involve a subject who has special reasons to be worried about being duplicated. The concern is that any ordinary person should eventually come to believe she is being duplicated, as long as she begins with a non-zero credence in the duplication hypothesis at issue. (Interestingly, if this argument were any good, TYPICALITY and INVARIANCE would face a similar—though admittedly weaker— kind of argument.[63])

Meacham provides a proof of his claim that ordinary people will eventually converge on the strange hypothesis, but his proof assumes for simplicity that "there are only two worlds under consideration, one normal world and one brain-duplicating world; it's easy to see how the result generalizes to multiple worlds" (266). But is this easy to see? After all, while it is epistemically possible that there are brains in vats with my CES constantly being *produced*, it is also epistemically possible that there are brains in vats with my CES constantly being *destroyed*.

Take the hypothesis that since my birth, brains in vats mirroring my successive experiential states have been produced at a rate of 1 per minute—call that H_1. (When I have lived N minutes, H_1 postulates N experiential duplicates of me.) Now consider the hypothesis that at my birth there were N brains in vats with my CES, set to be destroyed at a rate of 1 per minute—call that H_2. If I use FREQUENCY and I began with equal credences in those two hypotheses, I should find that every minute, some of my credence in H_2 leaks over to H_1. But it does not follow that my overall credence in brains-in-vat hypotheses has grown at all. Admittedly, when I reach N minutes of age, I rule out this particular 'destruction' hypothesis for good— but I have plenty more destruction hypotheses where that came from, not to mention hypotheses where the number of brains grows until it reaches *n* and then shrinks thereafter![64]

---

[63] After all, supposing that in the rest of my epistemic space, the ratio of predicaments with my CES out of all predicaments is 1 in a trillion trillion, this 'strange' hypothesis will eventually come to have a trillion trillion times its original credence (though given normal credence assignments my credences won't converge to 1 if I use TYPICALITY). And likewise for INVARIANCE: let the hypothesis be that at every interval, a brain is produced for every one of the possible continuations of my own CES from a moment before. As I have more experiences, I will rule out plenty of normal worlds, but never rule out any worlds consistent with that hypothesis, and the credences I assign to this strange hypothesis will continue to grow. See the next fn. for the problem with all of these arguments.

[64] A similar point can be made against the analogous arguments I sketched in the

Of course, an agent who worries about 'production-hypotheses' but grants zero credence to all 'destruction-hypotheses' may indeed have the problem Meacham raises. But why would any ordinary person have that kind of credence distribution?

Works Cited:

Adlam, Emily, 2014. "The Problem of Confirmation in the Everett Interpretation". *Studies in History and Philosophy of Science* Part B: 47:21-32.

Arntzenius, Frank. 2003. "Some Problems for Conditionalizing and Reflection.' *The Journal of Philosophy* 100(7):356-370

Bartha, P., and C. Hitchcock. 1999. "No One Knows the Date or the Hour: An Unorthodox Application of Rev. Bayes's Theorm." *Philosophy of Science* 66: S329-S53.

---

previous footnote against TYPICALITY and INVARIANCE. Indeed, Meacham makes this kind of point when defending INVARIANCE from just such an argument:

> [I]f we placed no restrictions on which strange worlds were allowed, then the experience of eating chocolate ice cream would eliminate lots of strange worlds as well as lots of normal worlds. Whether your credence in strange worlds increases relative to your credence in normal worlds depends on which strange and normal worlds... [your] priors and evidence lead [you] to believe could be [yours]. And it's reasonable to think that if you have doxastic worlds like ours, your credence in strange worlds will not gain on your credence in normal worlds.

The point here is that a normal person would not *only* consider skeptical scenarios in which brains are being produced with every possible subsequent CES I might have, but also scenarios in which only *some* of those brains are being produced. Since some of those scenarios don't involve brains that go on to experience chocolate ice cream, I can rule them out in the ordinary way, and it is not clear that the epistemic space devoted to skeptical scenarios as a whole increases. But Meacham fails to notice that a structurally similar point can be made in defense of frequency:

> The [argument against INVARIANCE]… entails that people with certain idiosyncratic doxastic set-ups will come to believe something counter- intuitive. The [argument against accounts like FREQUENCY], on the other hand… entails that people like us should come to believe that we live in a strange world. So the skeptical arguments considered weigh more heavily against [FREQUENCY] than they do against the account I favor. (pp. 264-265)

Berker, Selim. 2008. "Luminosity Regained." *Philosopher's Imprint* 8(2).

Bostrom, Nick. 2001. "The Doomsday Argument, Adam & Eve, Un++, and Quantum Joe." *Synthese* 127: 359-87.

———. 2002a. *Anthropic Bias: Observational Selection Effects in Science and Philosophy*. London: Routledge.

———. 2002b "Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation." *Journal of Philosophy* 99(12): 607-23.

Bostrom, Nick, and Milan Ćirković. 2003. "The Doomsday Argument and the Self-Indication Assumption: Reply to Olum." *Philosophical Quarterly* 53: 83-91.

Earman, John. 1992. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, Mass.: MIT Press.

Dieks, D. 1992. "Doomsday - Or: The Dangers of Statistics." *Philosophical Quarterly* 42: 778-84.

Dorr, Cian. upublished ms. "A Challenge for Halfers".

Dyson, L., M. Kleban and L Susskind, "Disturbing Implications of a Cosmological Constant" Journal of High Energy Physics 0210:011,2002.

Lisa Dyson, Matthew Kleban, Leonard Susskind

Elga, Adam. 2000. "Self-Locating Belief and the Sleeping Beauty Problem." *Analysis* 60:143–147

———. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69:383–396.

Garriga, J. and A. Vilenkin. 2008. "Prediction and Explanation in the Multiverse" *Physical Review D* 77, 043526.

Gott, Richard J. 1993. "Implications of the Copernican principle for our future prospects". *Nature* 363, 315-319. doi:10.1038/363315a0

Glymour, Clark. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Guth, Alan. 2000. "Inflationary Models and Connections to Particle Physics" arXiv:astro-ph/0002188v1

Halpern, J. 2006. "Sleeping Beauty reconsidered: conditionalizing and reflection in asynchronous systems." In T. Gendler & J. Hawthorne (Eds.), Oxford studies in epistemology (Vol. 1, pp. 111–142). Oxford University Press.

Hartle, J. and M. Srednicki, 2007. "Are We Typical?", *Physical Review D* 75, 123523. arXiv:0704.2630

Howson, Colin, and Peter Urbach. 1989. *Scientific Reasoning: The Bayesian Approach*. New York: Open Court.

Joyce, James. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

———. 2005. "How Probabilities Reflect Evidence." *Philosophical Perspectives 19*: 153-178.

Leslie, John. 1989. *Universes*. London: Routledge.

———. 1996. *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge.

Lewis, David. K. 1979. "Attitudes *De Dicto* and *De Se*." *Philosophical Review*, vol. 88: 513–43.

———. 1980 "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability, Vol 2*, edited by Richard C. Jeffrey. Berkeley: University of California Press.

Linde, Andre. 2007. "Sinks in the Landscape, Boltzmann Brains, and the Cosmological Constant Problem". http://arxiv.org/pdf/hep-th/0611043v3.pdf

Meacham, Christopher J.G. 2008. "Sleeping Beauty and the Dynamics of De Se Beliefs." *Philosophical Studies*, 138: 245-269.

———. 2010. "Unravelling the tangled web: Continuity, internalism, non-uniqueness and self-locating beliefs." In T. S. Gendler & J. Hawthorne (Eds.), Oxford studies in epistemology (Vol. 3, pp. 86–125). Oxford University Press.

Moss, Sarah. 2012. "Updating as Communication" *Philosophy and Phenomenological Research* 85(2): 225–48.

Ninan, Dilip. unpublished ms. "Self-location and Other-location".

Novak, Greg. 2010. "A Defense of the Principle of Indifference", *Journal of Philosophical Logic* 39:655–678

Olum, Ken. 2002. "The Doomsday Argument and the Number of Possible Observers." *Philosophical Quarterly* 52: 164-84.

Page, Don. 1996. "Is our Universe Likely to Decay within 20 Billion Years?" Int. J. Mod. Phys. D5 583. arXiv:hep-th/0610079v1

Parfit, Derek 1998. "Why Anything? Why This?", London Review of Books, Jan 22, pp. 24–27.

Pust, Joel 2007. "Cartesian Knowledge and Confirmation". *Journal of Philosophy* 104 (6):269-289.

Ross, Jacob. 2010. "Sleeping Beauty, Countable Additivity, and Rational Dilemmas" *Philosophical Review*, 119 (4): 411-447

———. 2012. "All Roads Lead to Violations of Countable Additivity." *Philosophical Studies* 161 (3):381-390.

Schwarz, Wolfgang. 2012. "Changing Minds in a Changing World". *Philosophical Studies* 159(2): 219-239

Schwarz, Wolfgang. 2015. "Belief Update Across Fission", *British Journal for the Philosophy of Science* 66 (3):659-682

Skyrms, Brian. 1980. *Causal Necessity*. London, Yale University Press

Smart, J. J. C. 1989. *Our Place in the Universe: A Metaphysical Discussion*. Oxford: Blackwell.

Smeenk, Chris. 2014. "Predictability Crisis in Early Universe Cosmology" *Studies in History and Philosophy of Science Part B:* 46 (1):122-133.

Tegmark, Max. 2004. "What Does Inflation Really Predict?" arXiv:astro-ph/0410281v2

Titelbaum, Michael. 2008. "The relevance of self-locating beliefs." *The Philosophical Review*, 117, 555–606.

van Fraassen, Bas. 1984. "Belief and the Will", *Journal of Philosophy* 81, 235-256.

———. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.

van Inwagen, Peter. 1993. *Metaphysics*. Colorado: Westview Press.

Vilenkin, Alex. 2006. *Many Worlds in One: The Search for Other Universes*. New York: Hill and Wang.

———. 2011. "The Principle of Mediocrity". arXiv:1108.4990v1

Weatherson, Brian. 2005. "Should we Respond to Evil with Indifference?" *Philosophy and Phenomenal Research* 70: 613-35.

White, Roger. 2008. "Evidential symmetry and mushy credence" In T. Szabo Gendler, & J. Hawthorne (Eds.), Oxford studies in epistemology (Vol. 3). New York: Oxford University Press

Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.