Here are two related questions about evidence:

(1) How should I update my credences when I get de se evidence—that is, evidence that does not simply characterize the world at large but tells me about my place in it?

(2) How should I reason when my experiential state might be non-unique—that is, when the world might contain more than one individual with all the experiences and apparent memories I have right now?

A principled answer to these questions would help us solve several puzzles that have been central to the recent literature in confirmation theory. (I have in mind, for example, the Dr Evil, Sleeping Beauty, and Doomsday puzzles, as well as the question whether finding out that the universe is ‘fine-tuned’ would count as evidence for the existence of many universes.)

This paper looks at three ways to answer these questions. Each can be thought of as replacing the standard conditioning rule with a new normative principle that constrains credences. Two of these principles already have champions—Nick Bostrom and Christopher Meacham respectively—but I will be arguing that a third rule fares better.

Let us begin, however, with a point of agreement among all three views—a limited indifference principle due to Adam Elga.

1. Parity

Consider the following scenario:

---

1 Many thanks to the many brilliant minds that have helped shape this paper through discussion or comments, including 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 whose input has been absolutely critical: John Hawthorne, Jacob Ross, and Charles Sebens.

2 This is also called ‘self-locating’ or ‘centered’ evidence; see Lewis 1979 for an introduction. De se hypotheses are often treated as sets of ‘centered’ possible worlds (world-time-subject triples), though I would prefer to treat them as functions from subjects-at-times to singular propositions about them. (Such a function is ‘true of’ a subject when it yields a truth taking that subject as argument.) For an excellent discussion of problems with the ‘centered worlds’ approach, see Ninan (ms).

3 See Bartha and Hitchcock 1999; Bostrom 2001, 2002a; Elga 2000, 2004; Leslie 1989, 1996. I will not be illustrating how each principle applies to each of these questions.
According to cosmological theory T, there are many unconnected universes, a tiny minority of which have an undetectable feature F. However, the vast majority of subjects in existence inhabit universes that are F.

Absent other relevant evidence, it seems we should be confident that, if theory T is true, our own universe is F. But what principle, if any, is at work here?

Here is one proposal. The likelihood of a given subject’s having exactly my experiences is pretty low. And since most of the subjects are in F-universes, I should treat it as more likely that my experiences occur at least once in an F-universe than in a non-F universe (assuming that having my experiences is independent from inhabiting an F universe.) But this idea does not explain our reaction to the following variant:

In addition, T predicts there are enough universes to guarantee that universes of both kinds contain beings with exactly our experiences. But the vast majority of such beings live in F-universes. Here again it seems that, if T is correct, we should think our universe is probably an F-universe. And this conditional credence should vary with T’s prediction about the proportion of subjects with our experiences that live in F-universes—at the limit T predicts that all subjects with our experiences live in F-universes and we should be certain that, conditional on T, we are among them!

To accommodate this kind of intuition, Elga (2004) introduces a limited indifference principle for cases involving experiential duplication, but it will require some jargon to state. First, take the sum of my current qualitative experiences and apparent memories, all internally individuated. Call this my complete qualitative state or ‘CQS’. Second, take a fully specific de dicto hypothesis: call that a world. And finally, take a fully specific situation within a world, in which a subject has a CQS at a time. Following Elga, call this a predicament. We can now state

Any two predicaments with one’s CQS in the same world deserve equal credence.6

---

4 Hypotheses involving enormous numbers of universes are very much a part of the contemporary scene in cosmology. For an overview, see Vilenkin 2006. See also Bostrom 2002b for an exposition of the philosophical issues raised by such theories.

5 Predicaments are maximal situations in which a subject can find herself, where this includes the actual world’s being thus-and-so. (As such, they are ‘world-bound’.)

6 The original principle uses indistinguishability of predicaments, but as Brian Weatherson points out, this yields some absurd results if being indistinguishable from is intransitive (2005, §4). Putting things in terms of sameness of CQS avoids these problems. Of course, if sameness of CQS is not ‘luminous’ (to use Timothy Williamson’s term), one will not always be in a position to know whether one is following the rule. But so it goes: our epistemic limitations render our decision-theoretic principles ‘mere’ idealizations.
Another way to implement the intuition behind this rule is with a constraint that applies directly to every hypothesis conditional on a world. Let \( e \) represent my current CQS, and \( P_e \) be my posterior credence function upon having \( e \). For any world \( w \), let \( N_w(e) \) be the number of predicaments that have \( e \) in \( w \). And let \( N_w(e \& h) \) be the number of predicaments in \( w \) that have \( e \) and also exemplify \( h \). (A predicament \( p \) exemplifies—or is an exemplar of—a hypothesis \( h \) just in case \( h \) is true of the subject of \( p \) at the time of \( p \).) Here, then, is the rule:

\[
\text{PARITY: } P_e(h \mid w) = \frac{N_w(e \& h)}{N_w(e)}
\]

Note the narrowness of this rule compared with what is ordinarily called ‘the principle of indifference’, viz. that, if one’s evidence no more supports one hypothesis than another, one should assign them equal credences.

PARITY has significant prima facie appeal for many, though it is not without its difficulties. In this paper I will set aside the debate over PARITY itself and consider three ambitious proposals for replacing conditionalization with a generalized version of PARITY, arguing that two of them are untenable but the third is a promising option for friends of PARITY.

---

7 One difference between the two principles is that PARITY involves the assumption that \( P_e(h \mid w) \) is defined for any world containing any predicaments with my CQS. (ELGA’S RULE is consistent with assigning a credence of 0 to every predicament in such a world and allowing \( P_e(h \mid w) \) to go undefined.)

8 Thus, for example, this principle is not subject in any obvious way to ‘cube factory’ worries of the sort raised by van Fraasen 1989. For some contemporary responses to such worries, see White 2008, Novack 2010.

9 See Weatherson’s 2005, which argues that ELGA’S RULE wrongly treats uncertain questions as risky ($\S$6), and also raises problems involving infinitary and unbounded cases. For example, suppose I know there are (countably) infinitely many subjects with my CQS, each tagged with a natural number. Then ELGA’S RULE and PARITY require me to assign equal credences to countably many de se hypotheses, causing trouble for countable additivity. (For one response to this kind of worry, see Ross 2010, $\S$5.)

Another problem arises even if we set aside issues of countable additivity. Suppose that rather than considering countably many predicaments in the same world, Dr Evil must divide my credences among infinitely many possible worlds. If he assigns the same infinitesimal credence (or zero credence) to each such world, he automatically satisfies ELGA’S RULE (and PARITY) no matter what credence he assigns to general hypotheses like ‘I am the duplicate’. As Weatherson notes, this problem can be avoided by reformulating the principle so that it governs multi-world hypotheses directly. For example:

\[(ER^*) \text{ For any hypotheses } x \text{ and } y \text{ such that in every world, } x \text{ has exactly } N \text{ times as many exemplars with one's CQS as } y \text{ does: } x \text{ deserves } N \text{ times the credence of } y.\]

(This generalizes on Weatherson’s suggestion, which compares only pairs of hypotheses with at most one exemplar per world.)
2. PARITY and de se updating

It is well known that de se beliefs dynamics cannot properly be modeled by a flat-footed application of the standard conditioning rule. Consider the following example from Frank Arntzenius (2003: 367). Jane is omniscient about the de dicto facts, watching a clock that she is certain is accurate. At first 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 conditioning. After all, standard conditioning 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 have already ruled out.

Notably, this problem goes away if Jane uses PARITY instead of conditioning to update on her de se evidence. PARITY tells her to conform her credence in it is 7am conditional on w to the expected fraction of predicaments with her CQS in w who in fact exemplify it is 7am. Since Jane is de dicto omniscient, she knows that w holds. And she knows that there is exactly one predicament in w with her CQS—that’s Nw(e)—and that one predicament also exemplifies the hypothesis that it’s 7am—that’s Nw(e&h). So she ends up certain that it is 7am.

The strategy of updating one’s de se credences with PARITY also satisfies the kind of intuition sketched above for UNIVERSES* even in cases where one acquires experiential duplicates over time. Consider an example based on Elga’s ‘Dr Evil’ case (2004):

---

10I am assuming a treatment of self-locating belief along the lines of Lewis 1979. For some recent discussions that offer proposals at varying levels of generality—some of which entail PARITY—see Halpern 2006, Titelbaum 2008, Meacham 2008, Meacham 2010, Schwarz forthcoming, Moss forthcoming.

11In our notation, the rule is P(h)=P(h|e). Assuming that the latter value is defined as P(e&h)/P(e) rather than treated as primitive, the result is undefined because P(e)=0. Those who prefer primitive conditional credences for de se hypotheses would still face the question of where these values should come from. However, the principles that follow, which I cast as alternative rules for updating, could instead be treated as proposals for constraining primitive conditional credences in de se hypotheses.

12This example involves a loss of certainty in de se hypotheses. (We could replace ‘the clock reads 6am’ with ‘the clock seems to read 6am.’) But suppose we hold that Jane should not be certain how things seem to her at a given time; while Jane thinks the clock seems to read 6am, she should leave open (for example) the possibility that she’s having a cognitive hallucination and in fact the clock seems to read 7am. But even then, 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, forthcoming, §2. (There may also be other possibilities that Jane should not be ruling out; the point only requires that she finds it likely that, if the clock seems to read 7am, she’s having a cognitive hallucination.)
You are told that at a future time $t_2$, the gods will create someone with exactly the same CQS that you have at that time, including apparent memories. After a short time, they will destroy the duplicate, while you live on indefinitely.

At $t_1$, you should be certain that you are the original and indeed that you will never be the duplicate. But many intuit that when you arrive at $t_2$ you should lose your certainty that you are not the duplicate and start to worry. Again, simply trying to condition on your de se evidence that it is now $t_2$ won’t work, because your prior credence in its being $t_2$ is zero. But using PARITY to arrive at your de se credences gets the intuitive result. This is because your new credence that you are the original conditional on any world will equal the fraction of predicaments with your CQS in that world that are the original. And in every world consistent with your de dicto evidence, this value is $\frac{1}{2}$.

These results are suggestive, at least for those who agree with the intuitions that PARITY was formulated to capture. However, we do not yet have complete solution to the problem of de se updating, because PARITY only yields credences in de se hypotheses conditional on worlds. In order to arrive at unconditional de se credences, subjects who get de se and de dicto evidence together must rely on an independent way of arriving at new credences in worlds. For example, suppose you know that the gods will only engage in the duplication prank if they are feeling angry, and that this in turn has a high objective chance if it’s stormy out. You then learn simultaneously that it is $t_2$ and that it’s stormy outside. This time, working out your de se credences (e.g. that you are about to die) requires integrating PARITY with a method of arriving at credences for de dicto hypotheses (e.g. that the gods are angry).

3. Three principles

PARITY tells you, in effect, to believe de se hypotheses to the degree that they are exemplified by more predicaments with exactly your CQS. (I will sometimes call these ‘predicaments like yours’.) But it does not deliver a verdict for hypotheses that disagree about the de dicto facts. For example, consider these variations on examples due to Bostrom.

---

13 Yielding an unconditional credence of $\frac{1}{2}$ for being the original technically requires an assumption about how the space of worlds is partitioned, or else the use of use ER* in place of PARITY. (See the second part of fn. 9 above.)

14 There are alternative solutions to the problem of de se updating that self-consciously avoid the results yielded by PARITY, such as the ‘shifted conditioning’ recommended by Schwarz (forthcoming). I won’t try to adjudicate between these alternatives here.

15 Bostrom presents a number of ingenious cases in his 2001, 2002a, and 2002b.
(BUILDINGS) In building 1, the gods create one subject that sees lights. In building 2, the gods create three subjects, exactly two of which see lights. (Their experiences are no more fine-grained than this.) You wake and see lights.

(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 which see lights. (The subjects’ experiences are no more fine-grained than this.) You wake and see lights.

(In both cases it helps to assume the world contains no other subjects than the ones created here.) Note the structural similarity:

Let’s start with BUILDINGS. Because PARITY tells you to treat each of the lights-seeing predicaments as equally likely in every world consistent with your evidence, it will lead you to suspect you are in building 2. But PARITY is silent on your credences for heads and tails in LIGHTS; it only constrains the way you distribute credences within worlds.

Here are three competing intuitions about LIGHTS that are consistent with PARITY:

“The chances dictate equal credences to heads and tails, and these should be unaffected when you see lights. After all—someone was bound to see lights on either outcome!”

“But you don’t just know that someone sees lights—you know that you see lights! That was guaranteed given heads, but you might have been in the dark given tails. Since that is your only evidence beyond the coin toss, heads probably came up.”

“Actually, seeing lights was not guaranteed by heads—you might not have existed at all. Your most specific evidence is that you exist and see lights, which seems more likely given tails. So LIGHTS should be treated like BUILDINGS after all.”

---

16 In fact, for the proportionalist it will also matter whether the gods themselves are in one’s reference class (see §8). I will assume that they are not.

17 That is, following the principle that I should proportion my subjective credences to my expectation of the objective chances; see Lewis 1980. See the next section about whether my existence might count as ‘inadmissible evidence’.
Let us spell out each of these intuitions as a suggestion about how PARITY should be integrated with one’s knowledge of the chances induced by the coin toss. (For the moment we can treat that knowledge as encoded by ‘priors’, even though strictly speaking the subject in LIGHTS has no temporally prior beliefs; see below.)

The first intuition is captured by the invariantist strategy suggested in Meacham (2008). The idea is to first arrive at one’s de dicto credences using whatever de dicto priors and evidence one has at one’s disposal, and then distribute the resulting credence value for each world equally over that world’s predicaments with your CQS. This means purely de se evidence never affects your credences in de dicto hypotheses: so your credences in the outcomes of the coin toss in LIGHTS should simply match their objective chances.

The proportionalist strategy is favored by Nick Bostrom. The basic thought is that you should proceed as though your experiences likely to be representative. This means that, other things equal, you should assign higher credences to hypotheses in which predicaments with your CQS constitute a higher proportion of all predicaments in the world. (More carefully: take the fraction of all predicaments in each world that have your CQS, weight this value by your prior credence in that world, and then normalize over the results.) The result for LIGHTS is a higher credence in tails, since all the predicaments see lights given tails.

Finally, according to the frequentist, you should begin by assigning equal credences to all possible predicaments with your CQS. But then you should weight the result for each predicament by your prior credence in the world in which it occurs, and normalize the results. The result is a general inclination towards worlds that contain more predicaments with my CQS, modulated by my de dicto priors. As a result, the frequentist recommends that the subject in LIGHTS assign a higher credence to heads.

Since all three strategies entail PARITY, they all agree about BUILDINGS: you should be \( \frac{2}{3} \) confident that you’re in building 2. But the strategies do diverge when considering de dicto hypotheses that differ in the number of predicaments with your CQS, or the proportion of such predicaments out of all the predicaments in the world. Very roughly, the frequentist is inclined towards de dicto hypotheses on which there are more predicaments like hers (other things

---

18 This is, at least, how a proponent of PARITY would implement what Meacham calls ‘compartmentalized conditioning’. See also Halpern 2004 for a very similar approach that does not apply to possible cases of simultaneous duplication.
20 This is a version of the principle Bostrom calls ‘The Self-Indication Assumption’; see Bostrom 2002b: pp.122-26, Bostrom and Ćirković 2003. Something similar is used by Bartha and Hitchcock 1999 to defuse the force of the Doomsday Puzzle.
equal). The proportionalist is inclined towards *de dicto* hypotheses on which there are fewer predicaments unlike hers. And the invariantist thinks that neither factor should affect her *de dicto* credences.

Before turning to a more detailed investigation and critique of these strategies, it is worth to make two points about the prior credences to which they appeal.

(i) *Absent priors and old evidence.* In LIGHTS, you have just come into existence, and so lack any temporally prior credences. (Likewise for the poor duplicate in DUPLICATE.) But the proportionalist and frequentist want you to proceed as though you had priors for the objective chances of the coin toss. How can you update on your very first experience (seeing lights) when you lack any beliefs from a time before you had that experience?

The answer is that this is an instance of a more general problem. Consider:

(VASECTOMY) Prior to meeting my mother, your father flipped a coin. If 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 qualitative evidence. At the time of the toss, the objective chance of tails was $\frac{1}{2}$. But that should no longer be my credence in tails: the Principal Principle does not apply because, to use Lewis’s lingo, you have inadmissible evidence—namely that you exist. To arrive at a credence for tails, I would like to integrate the background chances with this evidence, but it’s knowledge of the chances with my how can knowledge of the chances be integrated with this evidence? This is inadmissible old evidence. Your credence that you exist conditional on heads (or tails for that matter) is one. And you can’t update on something you’ve always known.

This is, of course, a version of the problem of old evidence, for which various solutions have been proposed. In what follows I will assume that in cases where

---

21 For the Principal Principle, see Lewis 1980. 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.)

22 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.

23 For some discussions of the problem, see Earman 1992, ch. 5; Glymour 1980, ch. 3; Howson and Urbach 1989, 272-75; Joyce, ch. 6.
subjects have no priors, they can still use ‘hypothetical priors’—roughly, a credence function that somehow disregards or brackets the relevant old evidence. (And, as we shall see, it is important that in VASECTOMY my very existence is old evidence I need to exploit a low hypothetical prior credence on my existence conditional on my father having had a vasectomy.)

(ii) ‘Shangri-la’ cases. One might think that use of hypothetical priors need only stand in for real priors in cases where the latter are unavailable for use, as when one 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, you get created at $t_1$ and live happily ever after. If tails, you get created at $t_2$ with false memories of being at $t_1$ and then destroyed. Your CQS at $t_2$ would be the same either way.

The structural resemblance with DUPLICATION can be illustrated as follows. (Circles filled with the same color represent predicaments with the same CQS.)

```
DUPlication    COIN

heads       tails

 t2    t2
   ↓    ↓
 t1    t1
```

Assume heads comes up in COIN. In that case, you start off certain that you are not going to die, but many intuit that you should start to worry you are going to die when you reach $t_2$—just as with DUPLICATION. 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 making good on this intuition requires not only (i) supplementing PARITY with some way of taking into account the chances of heads and tails; but also (ii) giving up the rule that your de dicto credences should only change as a result of conditioning on evidence. This is because you start off certain that heads came up but then start to doubt this when you reach $t_2$; conditioning will not allow you to lose certainty in a de dicto hypothesis. Moreover, your failure to adhere the conditioning rule is not explained by any involuntary loss of information—as when one gets hit over the head and develops amnesia. By
hypothesis, you proceed normally through time with no adverse cognitive events. You do become worried that your memories are not veridical, but if you updated only by conditioning, that wouldn’t happen. The epistemic possibility of false memories is a symptom of your violation of the rule, not its source. In short conditionalization not only fails to model the dynamics of your de se beliefs; it also fails to model the dynamics of your de dicto beliefs.

Those who are convinced by this case (as I am) can replace the appeal to temporally prior credences in their updating rule with an appeal to hypothetical priors, which encode the subject’s epistemic norms as applied to propositions in the absence of any evidence. On this approach, one’s current credences are always 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 you are uncertain whether you have any actual priors. The heart of the difficulty with coin is that you only become uncertain whether you have any priors as a result of abandoning your actual priors.

In my exposition of the three principles below, I will take this approach for granted. But for those unmoved by the relevant intuition about COIN who nevertheless accept PARITY, there are ways of accommodating all three principles so that they require the use of genuine priors whenever they are available.

4. INVARIANCE

According to invariantism, my de dicto credences should be arrived at by simply conditioning on my de dicto priors using de dicto evidence. Thus, for purposes of comparing de dicto hypotheses, my evidence is equivalent to its strongest de dicto entailment. In LIGHTS, for example, for the purposes of comparing heads and tails, my evidence is treated as equivalent to someone sees lights.

After setting my de dicto credences, I then apply PARITY within each world. This allows subjects to update one’s credences in de se hypotheses even when they do not learn any de dicto facts. Thus Jane can update when she sees the clock read ‘7am’; and I can conclude that I’m probably in building 2 in BUILDINGS.

24 Neither is this like a case where a new de dicto credence impinges on one’s cognition with the force of evidence, as in Jeffrey conditioning—Jane arrives at her new credences by reasoning. (If the coin is weighted, arriving at her credence that she just came into existence will involve some calculation!)


26 See fn. 40. 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 PARITY. In that case they will be interested in the discussion that follows only insofar as it contains arguments against certain ways of generalizing on PARITY!
Meanwhile, allowing hypothetical rather than actual priors, invariantism allows diachronic shifts that violate standard conditioning in cases like COIN.

We can state this procedure more formally. Let $w$ be a world-specifying hypothesis; let $P_e$ be my credence function when $e$ is my CQS; let $P$ be my hypothetical prior credence function; and let $e'$ be the strongest *de dicto* hypothesis entailed by $e$—namely that some predicament has this CQS at some time. For each $w$, one first achieves a posterior credence for $w$ by conditioning on $e'$: $P_e(w) = P(w | e')$.

Meanwhile, for any hypothesis $h$, PARITY yields the posterior probability given any world $w$. Summing these values for every $w$ gives us a posterior for $h$. In short, the rule is:27

\[
\text{INVARANCE: } P_e(h) = \sum_w P_e(w) \frac{N_w(e \& h)}{N_w(e)}
\]

That is, take the fraction of predicaments in a given world that exemplify $h$, out of all predicaments in that world with my QES. Then weight the fraction by the probability of that world, obtained by *de dicto* updating on priors. One’s credence in $h$ should be the sum of the resulting values for every world.

At first blush, this is a natural way to generalize PARITY. But it has some very counterintuitive consequences. These can be split into two kinds of cases: those in which I intuitively do not get any relevant evidence, but INVARANCE tells me I do; and those in which I intuitively do get relevant evidence, but INVARANCE tells me I do not.

(i) **Not getting evidence when I should.** First, consider a variant on LIGHTS that involves fewer subjects but proceeds in two steps:

**(TWO STEP)** The protocol is: heads, one subject, lights on; tails, two subjects, lights on for only one of them. Every subject wakes up with their eyes shut. I wake up, open my eyes and see lights.

\[\begin{tabular}{c|c}
\hline
& heads & tails \\
\hline
t2 & \includegraphics[height=1cm]{heads.png} & \includegraphics[height=1cm]{tails.png} \\
\hline
t1 & \includegraphics[height=1cm]{heads.png} & \includegraphics[height=1cm]{tails.png} \\
\hline
\end{tabular}\]

27 This way of putting the principle treats $P_e(w)$ as undefined in cases where one assigns a credence of zero to any worlds containing predicaments with my CQS. One could instead use a less fine-grained partition of worlds, and a restatement of parity along the lines of (ER*) from fn. 9.
At t1 the invariantist requires that I assign equal credence to heads and tails. Then I open my eyes and see that the lights are on. I learn from this that I am not the person in the dark room. But since this is merely de se evidence, INVARIANCE blocks any credence from seeping from one world to another. As a result, it requires me to retain equal credences in heads and tails. But it seems clear that I do get evidence for heads when I open my eyes—one tails was consistent with finding them off.

In addition, the invariantist will have 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 which there is an independent 50% objective chance of seeing lights.  

This time learning that the lights are on actually does rule out a de dicto hypothesis, namely that coin toss came up tails and both subjects saw darkness. So this time INVARIANCE dictates that seeing lights is evidence for heads. And this is the right result, but it seems bizarre to treat this case differently from LIGHTS. In both cases, I would say that it's twice as likely that I will see lights given heads as given tails. But this difference in my intuitive conditional de se priors has no effect when I open my eyes and see lights.

(ii) Getting evidence when I should not. There are also cases in which the invariantist says that one does get evidence, but intuitively one does not. For example:

(CHANCES2) As in CHANCES except that, however the coin lands, for every subject there is a 50% objective chance that the lights will be on.

---

28 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 ommitted].
29 I am not here endorsing the particular credence assignment recommended by INVARIANCE, namely 4/7 in heads.
30 This is a variant of an example used by Cian Dorr in connection with Sleeping Beauty, in Dorr (m.s.).
In this case, the invariantist will once again recommend equal credences for heads and tails at t1. Now suppose I see lights at t2. This intuitively provides me with no evidence at all for heads or tails: it was equally probable on either outcome. But INVARIANCE contradicts 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. To put it differently, there are two distinct possible heads worlds and four distinct tails worlds. Each of the heads worlds initially has a credence of 1/4 and each of the tails worlds has a credence of 1/8. In seeing that the lights are on, I am able to rule out one of the heads worlds and one of the tails worlds—the one where both subjects fail to see lights. 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. In that case I will have equal credences in the outcomes 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.

A 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 van Fraassen’s Reflection Principle, which

---

31 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, 1 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’.)

32 I have stated INVARIANCE in a way that entails PARITY, but the problems I discuss here apply equally to the more general approach of Meacham 2008.
requires us to conform to our future expected credences.\footnote{The principle 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, pg 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 such a principle will be violated by the invariantist.} Admittedly there are plenty of cases where one simply cannot avoid violations of Reflection, but in such cases one is worried that one will have memory loss, lose track of time, have one’s brain tampered with, and so on.\footnote{See for example Arntzenius 2004.} The invariantist has no such excuse.

Given these problems, I will set aside INVARIANCE in what follows, and focus on the other two principles introduced in §3.

5. FREQUENCY vs. PROPORTION

Consider a world \( w \) and two de se hypotheses X and Y. Suppose that in \( w \), X has \( n \) times more exemplars like mine than Y does. In that case, PARITY will ensure that, conditional on \( w \), X is assigned \( n \) times as much credence as Y. But why is this? Here are two options:

(a) Because the number of predicaments in \( w \) that have my CQS and exemplify X is \( n \) times as great as the number of predicaments that have my CQS and exemplify Y.

(b) Because the proportion of predicaments (out of all predicaments in \( w \)) that have my CQS and exemplify X is \( n \) times as great as the proportion of predicaments that have my CQS and exemplify Y.

Since \( w \) holds fixed the number of predicaments like mine as well as the total number of predicaments, these amount to the same thing. (Consider BUILDINGS and the hypotheses I’m in building 1 vs. I’m in building 2. Comparing the number of predicaments like mine that exemplify each hypothesis yields 1 vs. 2; whereas comparing the proportion of predicaments with my CQS that exemplify each hypothesis, out of all predicaments in the world, yields 1/4 vs. 2/4.) For this reason, PARITY codifies each of these intuitions equally.

Crucially, however, these two ideas come apart when we try to generalize PARITY so that it applies across worlds. For example, recall LIGHTS. (Heads: one subject, lights on. Tails: three subjects, lights on for two of them.) In that case, the number of subjects who see lights is greater given tails, but the proportion of subjects who see lights is greater given heads. So if we adopting (a) as our model will incline us towards tails in this case, while adopting (b) will incline us towards heads. Let us examine these two possibilities more carefully.
(i) Frequency. Suppose we want to generalize on (a). At a first pass, we might try the following. For any hypothesis $h$, whether de se or not:

Other things equal, $h$ deserves higher credence the greater the number of predicaments like mine exemplify $h$, assuming $h$ is exemplified.\(^{35}\)

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

\[ \text{(ABSURD) For any two possible predicaments with my CQS, I should be equally confident that I am in either of them.} \]

This treats all epistemically possible predicaments with my CQS 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 has 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 CQS by my prior probability in the world where the predicament lives. More generally, for any hypothesis $x$, I need to compare a prior expected number of predicaments with my CQS that exemplify $x$, with a baseline prior expected number of predicaments with my CQS. More carefully, define $N(x)$ for any hypothesis $x$ as follows.

\[
N(x) = \sum_w P(w) N_w(x)
\]

This takes, for every world, the number of predicaments that exemplify $x$, weights that number by the prior probability of the world, and sums the results. This yields a prior expected number of predicaments that exemplify $x$.\(^{37}\) We can then state the updating rule very simply as:\(^{38}\)

\[ \text{(FREQUENCY)} \]

---

\(^{35}\) 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. FREQUENCY integrates the Self-Indication Assumption with PARITY, while being precise about the ceteris paribus clause.

\(^{36}\) See Elga (2004: 387) for a rejection of this sort of principle; Elga does not there consider the kind of modification represented by FREQUENCY.

\(^{37}\) For the sake of simplicity, and to emphasize the connection with PARITY, 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 $x$.

\(^{38}\) Many thanks to Jacob Ross and Charles Sebens for helping me get clear on how best to formulate FREQUENCY.
FREQUENCY: \[ P_e(b) = \frac{N(e \& b)}{N(e)} \]

This gives the right result in BUILDINGS: I should be 2/3 confident I’m in building 2. It also yields a 2/3 credence in tails for LIGHTS. II
And 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. (Note that as formulated, FREQUENCY applies only when e represents one’s total information, including apparent memories and so on, and ‘P’ is a hypothetical prior credence function.)

(ii) Weighted proportion. The other alternative is to generalize on PARITY as characterized by (b). Nick Bostrom, who favors this approach, summarizes it like this:

One should reason as if one were a random sample from the set of all subjects in one’s reference class.

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 PARITY.) He also means that, other things equal, I should prefer worlds in which

39 In BUILDINGS, the total number of predicaments with my CQS in the world is three, and two of them are in building 2: so FREQUENCY yields the desired result that I should be
2/3 confident I’m in building 2. 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 CQS 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.

40 See the discussion about COIN in §2. Using FREQUENCY as stated on one’s actual priors will produce continual shifting in favor of worlds containing more predicaments with my CQS. 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 CQS, including my apparent memories at that time. Then we can arrive at a new credence function using:

\[ P_e(b) = \frac{\sum_w P_e(w) N_w(e \& b)}{\sum_w P_e(w) N_w(e)} = \frac{N_w(e \& b)}{N_w(e)} \]

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 ur-priors using FREQUENCY. (Many thanks to Charles Sebens for this diachronic version of FREQUENCY.)
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) would be 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’:

\[
F(x) = \sum_{w \in I} P(w) \frac{N_w(x)}{N_w(all)}
\]

(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:

\[
\text{PROPORTION: } P_e(h) = \frac{F(e \& h)}{F(e)}
\]

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 CQS using hypothetical priors.

It is worth emphasizing that, like INVARIANCE, FREQUENCY and PROPORTION both entail PARITY—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 BUILDINGS. But the two principles diverge in cases like LIGHTS: where FREQUENCY yielded a credence of 2/3 in tails, PROPORTION yields a credence of 3/5 in heads.\(^{42}\)

Finally, note that both principles create what I will call an existential selection effect: this is when my having this CQS has evidential bearing on de dicto hypotheses, beyond entailing the de dicto fact that someone has this CQS. In

\(^{41}\) Once again, thanks to Jacob Ross here.

\(^{42}\) My baseline expected proportion of predicaments with my CQS 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 CQS given \( heads \), while only 2/3 have my CQS given tails. (I will assume that there are no other subjects in the universe; this matters to PROPORTION.) 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 \( heads \).
other words, both principles yield credences in \textit{de dicto} hypotheses that diverge from the result of updating one’s hypothetical priors with one’s strongest \textit{de dicto} evidence.

6. Putative heuristics

Let us now examine some potential reasons for preferring one to the other.

(i) Marbles and urns. To motivate PROPORTION, Bostrom often appeals to the reasonable-sounding claim that one should treat oneself as a random sample from all the subjects in the world. And he has in mind the following sort of heuristic. Finding yourself in existence with your QES is a bit like randomly selecting a predicament from all the predicaments in the world and discovering that it has this QES.

Think of one’s predicament as like a marble selected out of an urn. On that conception of things, LIGHTS is analogous to the following example:

\textit{(Marbles)} A coin is tossed. If \textit{heads}, the urn contains one marble, which is marked ‘X’. If \textit{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 \textit{heads}— in fact, I should be 3/5 confident that \textit{heads} came up. This seems like a very natural heuristic to offer in favor of PROPORTION over FREQUENCY. But I think 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 for FREQUENCY.

The second approach is to think about this same case \textit{from the perspective of the marble}. So rather than imagining yourself \textit{selecting} a marble at random, imagine finding yourself in an urn after an uneventful marbly life. You know the setup 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.\textsuperscript{43} As a result, you take yourself to be twice as likely to have found yourself in the urn to begin with if \textit{tails} was tossed. But you are more likely to be marked ‘X’ \textit{conditional} on being in the urn if \textit{heads} was tossed. Taking both facts into consideration, you end up assigning 2/3 to tails. And of course, this is exactly how the frequentist wants to treat LIGHTS.

In short, the proportionalist’s urn analogy builds in the idea that one was guaranteed to experience something or other, as in the case where one is

\textsuperscript{43} Things are easiest, of course, if there are finitely many.
randomly selecting a marble out of an urn. But the frequentist's analogy takes into account that one might not have existed at all.

(ii) 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. SSA should be read as a methodological prescription specifying certain types of conditional credences of the form $P(\text{I am such and such an observer | The non-indexical properties of the world are such and such})$. The phrase “as if one were a random sample” is simply shorthand for these recommendations.\(^{44}\)

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

As I have formulated FREQUENCY and PROPORTION, they do not operate on prior conditional credences about \textit{de se} evidence, only on prior \textit{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 \textit{de se} hypotheses. Thus in LIGHTS, the proportionalist appears to reason as though her priors guaranteed that \textit{she} would be created by the incubator regardless of the outcome of the coin toss. Meanwhile, the frequentist appears to reason as 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.

There is, however, something a bit strange about the proportionalist approach. Consider LIGHTS and suppose that, unbeknownst to me, the coin came up \textit{tails}. In that case, there is another subject who sees lights and is wondering how the coin toss came up. If we are both proportionalists, then we will both

\(^{44}\) Bostrom 2003: 84. Or rather, as Bostrom himself stresses, the relevant credences would concern which predicament or 'observer-moment' I am in.
reason as though we would have observed *something* either way, but of course we can't both be right!\(^{45}\) Or consider the following case:

(TWINS) If *heads*, two people with my CQS are made; if *tails*, four people with my CQS 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 CQS* conditional on each outcome, and as a result 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 CQS* 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 proportionalist ought to resist this line of reasoning. She could, for instance, back off from the idea that *proportion* can be characterized as constraining hypothetical conditional priors involving *de se* beliefs, and stick with *proportion* as an updating rule that operates only on *de dicto* priors. And while I grant that these considerations are far from conclusive, I do find them suggestive.

7. Guaranteed existence?

There is, then, something intuitive about *frequency*'s preference for worlds containing a greater raw number of predicaments with one’s CQS. A more surprising benefit is that *frequency*, rather than *proportion*, 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:

\(^{45}\) 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 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.
So for example, suppose that again you find yourself seeing lights. But this time you know that seeing lights is evidence 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, PROPORTION instructs me to assign a higher credence to heads than to tails in both of these cases, because if heads came up, my CQS 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 PROPORTION intuitively does, we often end up with cases that FREQUENCY gets right and PROPORTION gets wrong.

8. The problem of subjecthood

PROPORTION 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 CQS. For that reason, we don’t need to decide whether anything counts as a subject—all that matters is whether it has my CQS. To illustrate this point, consider the following case:

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

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 CQS is like ‘from the outside’. The proportionalist, 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 sorites 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 proportionalist must stop recommending a credence of 1/2 in heads and start recommending a credence of 2/3, because PROPORTION 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 PROPORTION 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?46

The proportionalist might complain that this objection is 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.

It seems the proportionalist will have to fall back on the claim that in a wide variety of cases, it’s vague what credences it is 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. But the range of borderline cases seems especially significant for the proportionalist— 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 surely avoiding this thorny issue altogether is a prima facie benefit of FREQUENCY.

9. The prediction problem

Setting aside the question of what counts as a subject, PROPORTION 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

---

46 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 as 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 it 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.
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, PROPORTION 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 CQS 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 this variant on the story:

(EVE) Eve is alone in the world at t1. The incubator is about to toss a coin: if heads, it will do nothing. If tails, it will produce many subjects at t2 and give them all the very CQS that Eve had at t1.

Since the expected number of observers with her CQS 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 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. And as a result 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 proportionalist’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.
10. Excluding future subjects

At times, Bostrom seems willing to bite this bullet. But he also claims that the proportionalist could avoid the prediction problem by excluding future predicaments from her reference class.\footnote{See Bostrom 2001, pg. 381; 2002a, chs. 9 and 10.} It is not obvious how to apply this idea to every case, such as one in which two \textit{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:

\begin{quote}
\text{(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. $N_w(\text{all})$, when applying PROPORTION.
\end{quote}

But this is not much help, for a few reasons.\footnote{I will set aside worries involving the relativity of simultaneity.}

\begin{quote}
(i) Creation, execution, reflection. Suppose I know that the incubator is about to toss a coin. If \textit{heads} is tossed, it will do nothing. If \textit{tails} is tossed, then at t2 it will create a subject that is unlike me at any time. Either way, I will get no qualitative evidence about the outcome. A proportionalist can save the intuition that I should have equal credences in \textit{heads} and \textit{tails} by invoking EXCLUSION, which tells me that at t1 I should exclude the subject that may be produced at t2 from the value of $N(\text{all})$ in the tails world.

However, at t2 that predicament would no longer be future relative to me, so as a result I will suddenly suspect that \textit{heads} came up. And this shift in credences will occur despite the apparent lack of relevant evidence—in fact, at t1 I could have \textit{predicted} that I would shift my credences about the coin toss at t2. 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: \textit{heads}, the other subject is suddenly executed at t2; \textit{tails}, nothing happens. (I will get no qualitative evidence one way or the other.) Assuming my current CQS never gets repeated, \textit{heads} involves fewer predicaments that are unlike mine, so the original proportionalist suspects that heads will be tossed \textit{at the outset}. (In fact, this is a simple version of the ‘Doomsday argument’ for proportionalists.)\footnote{See fn. 60 below.}

Meanwhile, following EXCLUSION I will make no such prediction, but I will suddenly and predictably begin to suspect at t2 that the other subject has been executed.
\end{quote}
Perhaps these predictable shifts to pessimism are better than pessimism from the outset, since the latter involves violating the Principal Principle. In fact, while exclusionist admittedly violates van Fraassen’s Reflection Principle in such cases, 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 PARITY must admit that some violations of Reflection are acceptable: 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 so is not sure she is violating the principle at all. And this is precisely what happens in EVE. These 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 when knows that a couple is attempting to reproduce and their chances of success are $n$, or that there is an $n$ chance that someone will be executed. But EXCLUSION requires a violation of Reflection in precisely such a case, even if one is perfectly self-aware about the violation.

(ii) 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 PROPORTION tells me I should assign equal credences to heads and tails. But EXCLUSION tells me not to include the predicament at $t_2$ in the value of $N(\text{all})$ for heads, and not to include the predicament at $t_4$ in the value of $N(\text{all})$ for tails. (Given tails, I cannot assume the predicament

---

50 See fn. 33 above.
at t2 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 t3 and t4, so I end up. This is just bizarre. Intuitively the case should be just like TWO STEP, where the proportionalist (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 t2.) 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 proportionalists cannot plausibly treat location time as importantly different from location in space for purposes of calculating $N_w$(all).
11. ‘Presumption’ either way

We have so far been examining a number of problems for PROPORTION; but why has FREQUENCY been so neglected? Bostrom’s main objection to FREQUENCY is that it yields a counterintuitive result in cases like the following:

(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, T₁ and T₂… According to T₁ the world is very, very big but finite, and there are a total of a trillion trillion subjects in the cosmos. According to T₂, 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 T₂ 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 whether T₁ or T₂ obtains turns on some random occurrence early in the Big Bang that had an objective chance of .5.)

The result does seem counterintuitive. But consider a version of this example that arises simply due to PARITY:

(PRESUMPTION2) There are two very different regions in the cosmos. Region A contains a trillion trillion subjects. Region B contains a trillion trillion trillion subjects. Physicists are preparing a simple experiment that will establish which region we are in. Enter the presumptuous philosopher: “It’s unnecessary to do the experiment, because it is a trillion times more likely that we are in region B!”

This is just an exaggerated version of UNIVERSES, but it does elicit a sense of presumptuousness, at least to some degree. It is therefore worth identifying some possible sources of this intuition—however minor—that have nothing to do with PARITY in particular.

(i) In a realistic case the background elements of the story—that there are a trillion times more subjects in B, and so on—would not be known with anything like certainty, because they would be based on a physical theory that would

51 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 T₂ 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.
be at least somewhat tenuously held. But lack of certainty in the background facts makes it more difficult to rule out A.

(ii) The philosopher in the story assumes that an experiment is only worthwhile if it has a fighting chance of changing our credences in some significant way. But evidence can be evidentially significant in other ways; most importantly, by increasing the resilience of our credences with respect to various kinds of additional data that we might encounter. And resilience itself is a worthy aim of empirical endeavor.

(iii) When a non-ideal subject as a matter of fact succeeds in reasoning in an ideal way to some conclusion, we may expect that subject to have uncertainty about the perfect rationality of their reasoning. But the effect of this second-order doubt will often be to temper the results of their first-order deliberations. In particular, an ordinary proponent of PARITY should arguably not be certain that PARITY is the best way to reason. And this by itself may preclude using PARITY to reach near-certainty that we are in region B. (Of course, if we hedge our bets in this way, we should by our own lights suspect that we are not assigning exactly the correct credence to being in region B.) The fact that we forgive—even expect—this kind of second-order humility in non-ideal subjects should not cause us to give up the view that the relevant inferences are in some sense ideal. This suggests there might be a kind of over-arching sense of rationality according to which a proponent of PARITY who is not entirely secure in that principle ought not to entirely dismiss the value of undertaking the experiment. (After all, the experiment might very well make her more certain of the correctness of PARITY!) There are tricky issues in the neighborhood about the possibility of conflicting types of epistemic normativity, but the issue is worth flagging.

Taken together, these three considerations may somewhat mitigate our intuitions of presumptuousness when it comes to PARITY. However, many will still have stronger intuitions against the philosopher in PRECONDITION1 than in PRECONDITION2, and this appears to count as a cost. But is it a reason to prefer PROPORTION over FREQUENCY? If I am a frequentist, I will (other things equal) prefer theories where there are more predicaments like mine. And if I am a proportionalist, I will (other things 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:

(PRECONDITION3) As in PRECONDITION except the relevant theories are T3, which says there are a trillion non-green subjects in the universe.

53 Consider, by way of analogy, a subject who reasons deductively to a conclusion from a set of premises in which she is certain, perfectly using rational inference rules through a sequence of logical transformations. Nevertheless, she may harbor doubts either about the rationality of the inference rules, or about her own success in applying them.
and a trillion trillion green subjects; and \( T_4 \), which says there are a trillion of each.

The proportionalist, 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—the frequentist’s or the proportionalist’s? 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 section 6. Others may find the frequentist’s presumption more egregious, and for them, this should count in favor of PROPORTION. Even for them however, this should not outweigh the combined force of the arguments I have marshaled against PROPORTION.

12. The ‘many brains’ problem

Consider the following 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, PROPORTION and INVARIANCE would face a similar—though admittedly weaker—kind of argument.)

---

54 After all, supposing that in the rest of my epistemic space, the ratio of predicaments with my CQS 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 PROPORTION). 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 CQS 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, the credences I assign to this strange hypothesis will continue to grow. See the next fn. for the problem with all of these arguments.
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 CQS constantly being produced, it is also epistemically possible that there are brains in vats with my CQS 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 CQS, 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 old, 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.\(^{55}\)

\(^{55}\) A similar point can be made against the analogous arguments I sketched in the previous footnote against PROPORTION and INVARIANCE. Indeed, Meacham makes this kind of point when defending INVARIANCE from just such an argument:

> If 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 CQS 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. (p. 264-265)
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?

13. An application: fine-tuning

Let me conclude with an intriguing application of FREQUENCY. 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. By ‘the alleged fine-tuning’ I mean the claim that various constants in our physical theory (such as the mass of the proton, the strength of the weak electromagnetic force, the strength of gravity, and so on) 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.

Now 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 naive state, thinking that the chance of a given universe producing life—indeed, its chance of producing my QES, 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 QES per universe.) Now in this naive state I assess One and Many as follows.

The expected number of predicaments with my QES that exemplify Many will be my prior in Many ($\frac{1}{2}$) times the expected number of predicaments with my QES given Many, which is a trillion over $n$. Meanwhile the expected number of predicaments with my QES that exemplify One will be my prior in One ($\frac{1}{2}$) times the expected number of predicaments with my QES 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 not be surprising, since we have

---

56 e.g. Leslie 1988; van Inwagen 1993; Parfit 1998; Smart 1989.
encountered cases like PREMPTION where FREQUENCY yields a very high cre-
dence in a hypothesis that predicts many instances of my CQS.

\[
P_e(\text{many}) = P_e(h) = \frac{N(e \& h)}{N(e)} = \frac{\frac{10^{12}}{2} \times \frac{1}{n}}{\frac{10^{12}}{2n} + 1} = \frac{\frac{10^{12}}{2n}}{\frac{10^{12}}{2n} + 1}
\]

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 QES—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 differ-
ce 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 CQS being produced in a given
universe is lower than I thought—both sides of the ratio drop by the same fac-
tor.

This is the existential selection effect in action. If it seems strange, consider
this analogy, inspired by one that Bartha and Hitchcock use while defending a
very similar result applied to the Doomsday argument. 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 mailed 1 person a $1 million check. If tails, we mailed $1 to a
million people randomly chosen from the phonebook.

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

e1: Someone got a notice in the mail.
e2: I got a notice in the mail.

If you simply conditionalize on \( e_1 \), you will assign equal credences to the out-
comes 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 the letter at all is far more likely given tails.

\[\text{Bartha and Hitchcock 1999 applies something like FREQUENCY to defang the dooms-
day argument. See also Dieks, 1992. For an early discussion of the Doomsday argument,}
\text{see Leslie 1996.} \]
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.

14. Conclusion

We have considered three ways of generalizing *parity* and replacing standard conditionalization that converge in cases where the number and proportion of predicaments with my CQS are held fixed. Of the three, *invariance* is the least attractive. It 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.

Meanwhile, both *proportion* and *frequency* face versions of the ‘presumptuous philosopher’ problem. But only *proportion* has the following problems in addition: (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 or else a combination of the asymmetry problem and some variety of the new subjects problem. Of the three principles, then, I tentatively conclude that *frequency* is our best bet.

If this is right, there are fairly straightforward consequences for some famous existence lotteries. I have explored a consequence for the question whether the alleged fine tuning of the universe is evidence for many universes. In addition, *frequency* yields the standard ‘thirder’ result for Sleeping Beauty, and significantly defuses the counter-intuitive force of the Doomsday puzzle. But I will leave a detailed discussion of these results for another occasion.

58 See the previous fn.
Works Cited:


Dorr, Cian. unpublished ms. “A Challenge for Halfers”.


Ninan, Dilip. unpublished ms. “Self-location and Other-location”.


Schwarz, Wolfgang. ms. “Belief Dynamics Across Fission”


