Indifference and Bias in Self-Locating Probability

Sinan Dogramaci, Simon Goldstein, John Hawthorne, and Miriam Schoenfield

(Draft only)

Abstract: A popular indifference principle for self-locating probability faces skeptical challenges, most notably a skeptical challenge concerning Boltzmann Brains. We examine the prospects for various views that prescribe some degree of bias in place of indifference. While a bias view has attractive anti-skeptical features, it turns out that bias leads to puzzles about deference to chance, reflection, and other related principles.

1. Given *Center Indifference*, Skepticism threatens

Any person, at a time, in a world, is a *center*. For any center, there is a rational probability, given your evidence, that *you* occupy that center, i.e. that you are that person at that time in that world. This is a so-called self-locating probability. Our topic is the principle of self-locating probability known as *Center Indifference*, or CI. We will state it as a constraint on the rational unconditional probability assignment, or ur-prior, which we'll call 'Pr':

> *(Strong) Center Indifference* (CI): If c_1 and c_2 are centers belonging to the same world, $Pr(c_1) = Pr(c_2)$.[1]

CI is highly tempting. But it implies, given some internalist assumptions,[2] that certain bodies of evidence, perhaps even including our actual evidence, pose a skeptical threat. The threat is posed by any evidence that recommends, by some ordinary scientific method, some theory according to which there are many centers that have exactly that evidence and that most of them are massively hallucinating.[3]

[1] There is a weakening of Strong CI according to which indifference is only mandatory for pairs of *synchronic* centres. David Builes (2024) prefers something like the weaker synchronic version (on presentist grounds), but it is beyond the scope of this paper to evaluate his arguments.

[2] We assume that both intraworld and transworld neural duplication guarantee experiential and evidential duplication.

[3] Let us sketch another way that a skeptical threat might be posed even without intraworld duplication: suppose that at every world (according to the relevant theory) nearly every center is massively deluded and for each evidential type, the evidentially accessible centers are nearly all massively deluded. Here is a simple model. There are ten worlds and in each world ten people have experiences of types e1 to e10. The ten worlds vary in which of the ten people are not deluded—but always there are 9 deluded, 1 not deluded. At each world, each non-deluded person has

Disturbingly, we may actually have such evidence. Cosmologists take seriously the view that we live in a universe that's teeming with Boltzmann Brains or, as they're sometimes called, "BBs"—brains that exist only for a moment while having a random experience.[4] Suppose, to idealize away from matters of scientific sociology, that cosmologists used a standard scientific method to arrive at a high confidence in a theory according to which the vast majority of a finite number of centers with your evidence are BBs.[5]

As Elga (2025) points out, it follows from CI that in such a case there are two available responses—both of which entail a kind of skepticism. One option is to think that one is a pretty ordinary agent but that, contrary to what science claims, the world is *not* teeming with hallucinating duplicates. According to this option, we're skeptical about the scientific method. Alternatively, we might think that we ourselves are massively hallucinating, as are the majority of those with our evidence. Note that this doesn't mean that one will think that one is probably a Boltzmann Brain. Insofar as one takes onself to be massively hallucinating, one won't be very confident in the particularities of a BB cosmology. For example, it may be that most agents with one's evidence are hallucinating not because there are many hallucinating duplicates, but because there is one agent with that evidence (oneself) who happens to be a complete lunatic. Crucially though, if one accepts CI, one *can't* think that probably the cosmology according to which one has many hallucinating duplicates is correct and that one is an ordinary agent.

---

9 accessible deluded centers. CI forces the prior for each world to evenly distribute across the centers at that world. If the prior is indifferent over these ten worlds then, conditionalizing any body of evidence on that prior will result in a .9 credence that one is deluded.

[4] See Carroll (2021) for an influential presentation of the relevant context.

[5] A complicated web of issues is raised by infinities. One possible complication concerns having infinitely many centers within a world. One could argue that, if any BBs exist, then the relevant populations are infinite in a way that blocks getting any skeptical verdict out of CI. (Suppose for example that the Boltzmann Brain hypothesis was embedded in an inflationary cosmology that predicted infinitely many ordinary copies of evidence E and infinitely many Boltzmann copies of E.) And even if each world population was finite and Boltzmann dominant, matters are still complicated if there any uncountably many worlds. If there are uncountably many worlds compatible with E, then even if, at each world, the ratio of Boltzmann Brains with E to ordinary observers with E favors the former, there will still arguably be uncountably many epistemically possible centers that are ordinary and have E. (Nothing immediately follows from pointwise dominance about the proposition that one is a Boltzmann Brain, even given CI.) The question as to how to handle infinitary populations is an important topic, but the puzzles raised by finite populations are tricky enough, so we will simply bracket the issue here by considering cases in which only finitely many worlds exist, each with no more than finitely many centers. (We admit this is patently an extreme idealization since, even if only finitely many people ever exist, each one plausibly has infinitely many time slices, and since it is overwhelmingly natural to think that there are infinitely many nomologically possible worlds that vary with respect to fundamental continuous magnitudes.)

Another common source of worry is similar: our evidence might seem to indicate that the universe will contain a huge but finite number of minds living in artificial simulations. If according to that theory most centers with your evidence are living in a simulation, then CI doesn't allow you to think that you're an ordinary person with lots of simulation duplicates. Instead you must think that either the theory is wrong, or you yourself are the victim of a massive hallucination.[6]

Let us say E is bad news for a belief (or credence) forming method if, upon having E, you must either reject the method or think that you are massively hallucinating (which itself may involve a rejection of the method if that method recommends a specific version of the deception hypothesis). If CI is true, then evidence that ordinary methods take to be evidence for Boltzmann Brain cosmologies or simulation hypotheses are bad news for those methods.[7]

Our aim in this paper is to examine the prospects for a view that rejects *Center Indifference*.[8] We'll call the view that rejects *Center Indifference* '*Center*

[6] It's worth flagging that there's a weaker version of *Center Indifference* which some authors writing on this topic, like Elga (2000, 2004) and Lewis (2001), seem to have in mind, on which these sorts of challenges go differently. According to this principle (call it "Weak CI") for any agents at any time, if there are two evidentially matching centers in the same world, one should assign the same probability to each of those centers. (At the level of the ur-prior framework that we are about to develop, this would amount to the idea that any two worldmates with the same body of evidence are assigned the same ur-prior.) Weak CI is compatible with all kinds of bias at the level of priors. For example, it is compatible with assigning zero or negligible prior to being a brain in a vat in a world where the brain in a vat's experience is not perfectly duplicated by a human's. This weaker principle still makes skeptical trouble given certain Boltzmann Brain cosmologies. But even here there is scope for sneaky bias. One can conform to weak CI but still assign absurdly high probability to being an ordinary observer whose evidence is not perfectly duplicated. Even given standard cosmologies that predict a large number of Boltzmann Brains, there will be nomologically possible worlds where some ordinary observer lucks out and doesn't have their evidence *perfectly* duplicated. Thus Weak CI only directly makes the standard anticipated trouble for BB cosmology in the context of a very toy version of BB cosmology (which for simplicity we work with in the main text) where at *every* Boltzmann Brain world every experience of an ordinary observer has many perfect Boltzmann duplicates. In general, an insanely high prior that one's experience is not perfectly duplicated will leave one free to add in also sorts of additional bias in the presence of Weak CI. So in the context of thinking about bias and skepticism, it is more helpful to focus on our strong version. It also helps reorient the reader towards thinking of skeptical problems in terms of ur-priors which are completely blind to who you are, which is part of the aim of this paper.

[7] Pittard (forthcoming) gives a similar diagnosis of the skeptical threat posed by CI.

[8] Dogramaci and Schoenfield (2025) gave a series of semi-technical arguments that CI has absurd consequences. But we are convinced by Hughes (ms.) and Pittard (forthcoming) that two of their arguments have questionable assumptions and the remaining argument only derives from CI the kind of skeptical consequence we already expected and so the argument reveals no new reason to reject CI.

Bias'.[9] On this view, rather than be indifferent among centers, we should be biased against deceived ones. As we'll see, there is a set of problems that arise for *Center Bias* that mirror the problems in the Sleeping Beauty case: problems concerning deference to chance and violations of the *Reflection* principle. We'll explain why these structural parallels arise, and point to some ways in which the problems facing *Center Bias* are more pressing than the problems posed by Sleeping Beauty. But first, we want to say a bit more to motivate *Center Bias*.

2. Anti-Skepticism motivates *Center Bias*

We assume there is some correct reply to the traditional skeptical challenge, the kind of challenge that involves hypotheses like Descartes's deceiving demon hypothesis. Most of the standard internalist-friendly replies imply that I am apriori justified in believing my evidence won't be deceptive, at least not if it's ordinary evidence like the kind I'd get if I looked at my hands or drew a marble from an urn and checked its color. This is a kind of bias in favor of the hypothesis that ordinary evidence is not deceptive.

This bias is logically consistent with CI. When combined with CI, it becomes a bias in favor of the hypothesis that most agents (with ordinary evidence) are not deceived. But although combining the views creates no logical inconsistency, it arguably creates an explanatory gap.

Many proposed explanations of why one is apriori justified in believing one is not deceived will fail to explain why one is also apriori justified in believing others are not deceived. Here are some examples. First, there is the Reichenbachian idea that one is apriori entitled to trust that one is not deceived because otherwise one would have no hope of successfully inquiring into, deliberating about, explaining, or minimally understanding the world.[10] But this is no explanation of why one should trust that other people are undeceived. Second, some philosophers claim that one cannot trust apriori arguments that one might be deceived because

[9] The choice between *Center Indifference* and *Center Bias*, as well as its significance for cosmology and skepticism, is anticipated, at least very roughly, by Srednicki and Hartle (2010, 2013) who introduced the idea that cosmological theories must be evaluated in conjunction with one or another "xerographic distribution". Explained in our terms, the xerographic distribution determines the self-locating probabilities of evidentially equivalent centers. Their positive proposal seems to be a kind of permissive and case-by-case pragmatic approach to the choice. They also seem to endorse a version of the principle we will call "Evidence Irrelevance" below (as Dorr and Arntzenius (2017, fn.4) also noticed).

[10] Wright (2004) and Schechter and Enoch (2008) develop the idea.

those arguments undermine one's own reasoning.[11] But this again would not explain why one should doubt that others might be deceived. Third, some argue one should believe one is not deceived because one has pragmatic reasons to believe one is not deceived.[12] But one has no corresponding pragmatic reasons to believe that other people are not deceived. Finally, some philosophers think that it is literally impossible to believe that one is (massively) deceived, so one must be entitled to the only available option which is to believe that one is not so deceived.[13] But it is perfectly possible to believe that other people are massively deceived.

While these proposed explanations have their limitations and face objections, our point is that there's some plausibility to the thought that only apriori self-trust can be explained, not apriori trust in everybody. But if CI is true, then the only way one can believe one is undeceived is if one believes everyone is undeceived. If we lack an explanation of why everyone is undeceived or why it's justifiable to believe everyone is undeceived, then CI creates an explanatory gap.

Here's another odd consequence of CI. Suppose we begin to develop technology to envat people. Even if it looks extremely likely that the technology will succeed, the natural development of an anti-skeptical version of this approach will recommend being very confident that it won't: either humanity will suddenly end before the technology comes to fruition or that technology will hit a hitch. There may be alternative approaches available to the CI defender, but given that CI generates serious skeptical consequences for agents with evidence for certain cosmologies, and even traditional anti-skeptical views face explanatory challenges if CI is true, everyone should be interested in seeing what the best prospects are for a view that allows for bias over centers.

3. *Center Bias*

How should we formulate what we'll call '*Center Bias*'—the view that rejects CI? While some of what we say will apply to any kind of bias one might be interested in applying over centers, we'll focus on a rejection of CI that gives lower probabilities to centers that are massively deceived. We won't say much at this

[11] Wright (1991) and Rinard (2018) develop this idea.
[12] Rinard (2021) develops this idea.
[13] Greco (2012) develops this idea.

point about what being 'massively deceived' amounts to—it can be treated as something of a placeholder.[14]

We'll work with the highly simplifying idealization that being deceived and being undeceived is a binary distinction, even though deception obviously comes in degrees. (What we say can easily be adapted to a degree-theoretic approach.) The challenges for *Center Bias* that we want to focus on can be raised and, at least initially, treated using the binary idealization. For now, we will assume that whether a center is deceived or not supervenes on the history of the universe up to and including the time of that center. (We will return to this assumption later.)

We use an ur-prior conditionalization model to represent the rational credences on any body of evidence.[15] On this model, there is an ur-prior, Pr, which assigns an unconditional probability to every center. The probability of a world is the sum of probabilities of its centers. Any proposition can be represented as a set of possible centers (where non-indexical propositions like that expressed by 'There is a large elephant sometime in the history of the world' are such that if one center at a world belong to the set associated with that proposition, so does every center at that world). Conditional probabilities are given by the ratio formula in the usual way. The ur-prior defines the rational credences: rational credences are the probabilities assigned by the ur-prior, conditional on the proposition that constitutes whatever your present total evidence is. This is a "time-slice theory" of rationality: we do not calculate what you rationally believe now from what you believed earlier, only from your present evidence and the ur-prior.[16]

Here is the general view that we'll be focusing on:

> *Center Bias:* Where 'Pr' is the ur-prior, if c_1 and c_2 are centers belonging to the same world, then $Pr(c_1) > Pr(c_2)$ iff c_1 is undeceived and c_2 is deceived.[17]

[14] We note that the deception-theoretic gloss will not in any obvious way induce a bias at the level of priors against farmers in very large Boltzmann Bubbles who haven't thought or read about cosmology and the like, though it will potentially induce a bias against astronomers in similar bubbles, as well as those who pick up astronomy from the news, books, and so on. (And it's not even clear that it induces a bias at the level of priors against a Boltzmann Brain that is aware of a headache, has no perceptual appearance of an outside world, and who thinks it is a Boltzmann Brain.) Whether this is a feature or a bug is not a question we shall pursue here. A rather more physics-based bias that imposes a measure on normality and that then uses normality to speak to the Boltzmann Brain issue may well deliver rather different results.

[15] See Appendix B of Isaacs, Hawthorne, and Russell (2022).

[16] We realize that there are deep challenges to internalism lurking in a framework which gives us only an ur-prior and current evidence to work with. We won't address these challenges here.

[17] Note that there is no assumption here that c_1 and c_2 evidentially match. There is a restricted version of *Center Bias* that replaces 'centers' with 'evidentially matching centers'. (This would be a rejection of what we called 'Weak CI'

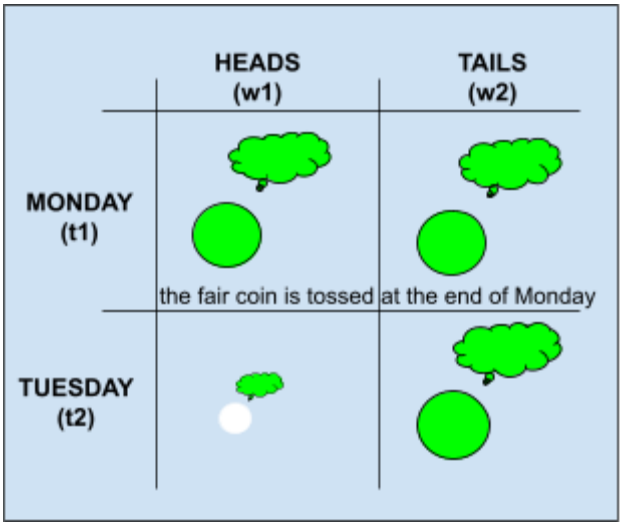
Let us represent the bias by a discount factor, which is a positive real number strictly less than 1, where the lower the number the greater the bias. If the ur-prior assigns probability x to an undeceived center c, the ur-prior will assign x multiplied by the discount factor to any deceived center in the same world as c. If the discount factor is .1 for example, then for any pair of deceived and undeceived centers in the same world, the probability assigned to the undeceived center will be ten times greater than the probability assigned to its deceived companion.

4. Does *Center Bias* obey *Deference to Future Chance*?

Here is a puzzle case that will help us start to examine *Center Bias*. On Monday, there is an undeceived center in a green room. At the end of Monday, this center dies, and a fair coin is tossed. If it lands Heads, then on Tuesday a deceived center is born in a white room bathed in green light. If it lands Tails, then on Tuesday an undeceived center is born in a green room. Assume nothing else happens and no one else exists.

Below is a diagram of the case. The deceived center in the lower left takes up less space to indicate that it receives less ur-prior probability than the other veridical centers, which all have equal size.

in note 6.) This restricted version of the principle would be enough to get the puzzles of the next section going. But we note that it would not by itself license assigning low prior to being a deceived Boltzmann Brain at a world if it had no evidentially matching undeceived companion. This fact makes the second, restricted, version far less powerful as an anti-skeptical tool: for example, if by the lights of the prior, most worlds had no ordinary individual with E but still had plenty of BBs with E, then the second view will, by itself do very little anti-skeptical work. By contrast, the view we're calling '*Center Bias*' can. However, even *Center Bias* will not discount the epistemic possibility of being a brain in a pure Boltzmann world with no ordinary observers (imagine the cosmological constant was set all wrong for ordinary observers to emerge, but that Boltzmann Brains still emerged by random fluctuations). That said, we will shortly develop models faithful to *Center Bias* that will help even here.



Let's separately state a number of assumptions that are initially individually plausible but lead to a puzzling consequence in our case.

(1) *Deference to Future Chance:* Pr(Heads|Monday) = Pr(Tails|Monday)[18]

After all, on the supposition that it's Monday, how could you be anything but 50% confident that a future flip of a fair coin will land Heads?

(2) *Indifference for Similar Veridical Centers:* Pr(Tails&Monday) = Pr(Tails&Tuesday)

(2) follows from *Center Bias*.

(3) *Conditional Certainty*:
Pr(Veridical-on-Monday or Deceived-on-Tuesday|Heads) = 1.
Pr(Veridical-on-Monday or Veridical-on-Tuesday|Tails) = 1.

We think (3) initially looks very plausible (but later we will reconsider it).

[18] This can be seen as an instance of a more general principle. Suppose some worlds share an initial history that includes a chancy process but not its outcome (and so they are allowed to diverge according to the outcome of the process). Then conditional on the proposition that one is in a certain subperiod of that history and outcome o has objective chance x, the ur-prior assigns probability x to outcome o.

(4) *Bias Against Deceived Centers:* Pr(Heads&Monday) > Pr(Heads&Tuesday)

(4) also follows from *Center Bias*. But now we get a strange consequence:

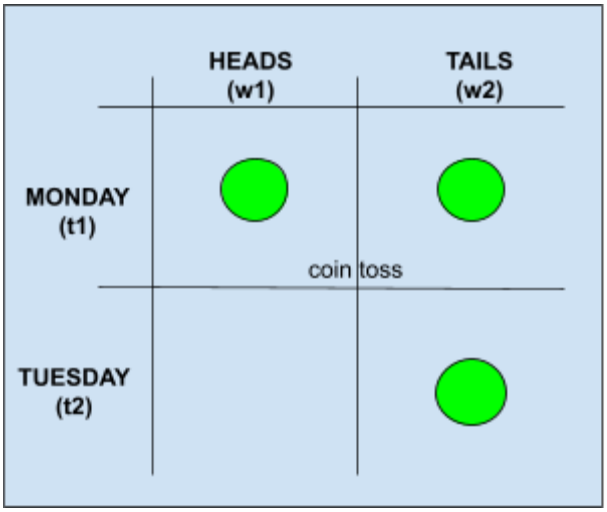
(5) *Violation of Ur-Prior Chance Deference:* Pr(Heads) < Pr(Tails)

The diagram makes it obvious that (5) follows from (1) - (4), since a world's probability is the sum of its centers' probabilities.

This is strange. According to (5), the ur-priors for two worlds that divide according to the outcomes of a fair coin toss (and which we can imagine are otherwise deterministic) are unequal.

If we modify the case so that Pr(Heads&Tuesday) = 0, as in the diagram below, we'd have a case that resembles Elga's (2000) Sleeping Beauty Problem.[19] Elga argues for (1) and (2) as applied to Sleeping Beauty. (3) would be substituted by Pr(Monday|Heads) = Pr(Monday or Tuesday|Tails) = 1. (4) would remain true because Pr(Heads & Tuesday) = 0 and Pr(Heads & Monday) > 0, and then (5) follows: Pr(Heads) < Pr(Tails). (Note that while Elga defends the thirder view, he doesn't give a theory on which conditionalizing the ur-prior yields the thirder's credences.)

[19] The astute reader will notice that the full strength version of bias does not even require that the evidential situation matches across Monday and Tuesday. In the first case make the Monday rooms red and the Tuesday rooms green. Then the argument goes through. (And then a denial of *Deference to Future Chance* is especially catastrophic, since the proposition that one is red on Monday is something that an agent can actually get as their total evidence). If *Deference to Future Chance* is thus violated, then when an agent conditionalizes on that total evidence they will not be 50/50 on a coin flip that is known to be in the future and objectively fair. Similarly, adjust the Beauty case so the things are red on Monday. The validity of the argument is untouched. We chose the uniform versions in part so that the arguments will still both go through straightforwardly for those tempted to a weaker version of the Bias view with 'evidentially matching centers' replacing 'centers.' Note even in that case though, a requirement that one give more than zero to all centers will make trouble in the adjusted Beauty Case. That regularity assumption, combined with *Conditional Certainty* and *Deference to Future Chance* will entail a violation of *Ur-prior Chance Deference*.



We'll next propose a way of developing *Center Bias* that allows the ur-prior to assign Pr(Heads) = Pr(Tails). We'll also show how the proponent of the thirder view of Sleeping Beauty can use an ur-prior conditionalization framework to say something similar.

5. The Bias View with *Deference to Future Chance*: the SI Rule

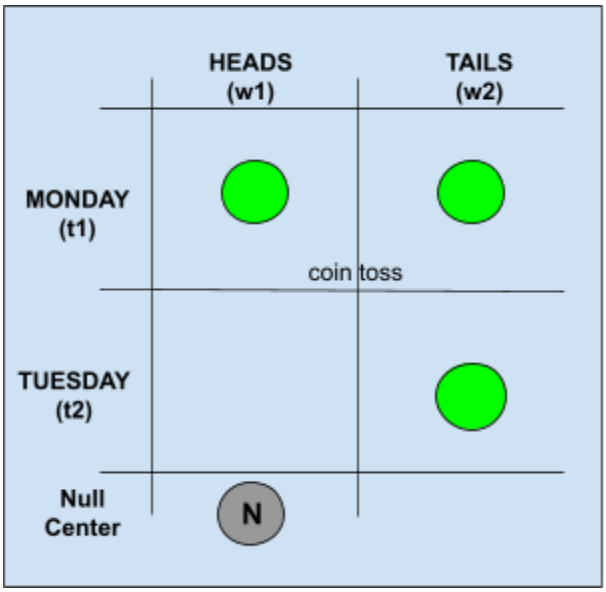
In the ur-prior conditionalization model, the thirder view of Sleeping Beauty is an instance of the so-called *Self-Indication* rule for assigning priors, or SI as we'll call it. As we shall develop it, the SI rule has the general feature that "I exist" confirms possible worlds in proportion to the size of their populations.

For "I exist" to confirm a proposition it needs to rule something out. Following an innovation from Isaacs, Hawthorne, and Russell (2002, appendix B), we say that it rules out the *null center*.[20] This means that centers are not only the occupied times and places in a world. Now we add that each world can also have an unoccupied center, the null center. You assign a null center your ur-prior probability that its world is actual and you do not exist.[21] This allows us to model the natural thought that not only is there a chance that the world is like this or like

[20] They also mention a more Williamsonian version with multiple null centers where the probability assigned to null is distributed over those centers. But we will not need to introduce this complication here.
[21] Readers who believe they exist necessarily can replace "do not exist" with "do not exist concretely" throughout.

that, and a chance that I am here now or there then, but there is (or was) also a chance (epistemically speaking) that I am never anywhere.[22]

This allows us to now model the Sleeping Beauty case as follows:

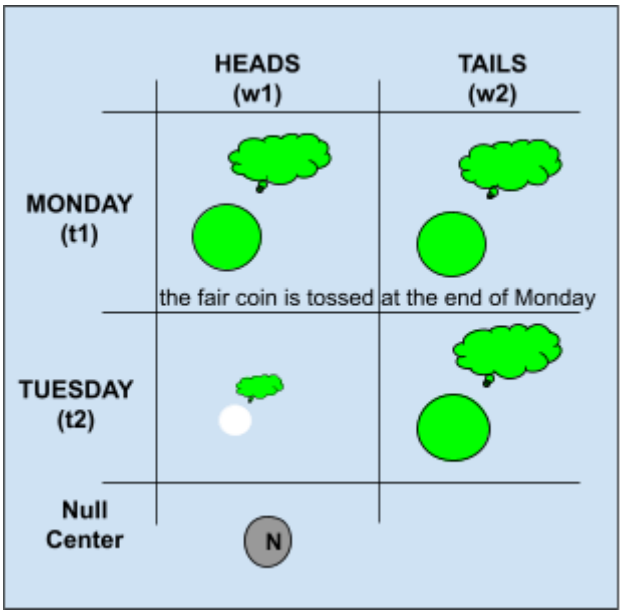


SI generally assigns relatively more probability to the null center in worlds with relatively smaller populations, which produces the effect that I'm more likely to exist if there are more centers. In the above picture of Sleeping Beauty, we can get the thirder result in a simple way, by adding probability to the null center only in the Heads world and making it equal to the Monday center in the Heads world (so each of the 4 centers in the diagram gets probability ¼). In general, what's needed to deliver the thirder view using ur-prior conditionalization and *Ur-Prior Chance Deference* is that the null center gets higher probability given Heads than given Tails.[23]

Similarly, we can model our puzzle case as follows:

[22] Another approach would be to have a distinct null center for each space-time location. As Chris Meacham reminds us (p.c., and see Meacham (forthcoming)), Lewis (1979), loosely following Quine (1969), defined centers simply as trios of <time, place, world>, which automatically generates countless unoccupied centers. This has the potential to lead to a very different kind of treatment of null, but we won't explore it at length here, especially as the idealization to finitely many centers at each world looks very strange in that context. But as a very toy model imagine a spacetime grid of two locations and run the simple sleeping beauty example in that setting. Then we can recover *Deference to Future Chance* in a parallel way to in the main text. (If grid size depends on coin tosses, that is a different matter!)

[23] For example, we could get the thirder result if each of the green centers in the diagram got probability ⅙ in the ur-prior, instead of ¼. The null center in the Tails world then gets probability ⅙, and the null center in the Heads world gets probability 2/6.



By assigning the null center in the Heads world less probability than the null center in the Tails world, we can solve the puzzle the same way. Since the null center makes claim (3) from above false, we can now endorse all the principles we intuitively want to endorse: Center Bias, Deference to Future Chance, and Ur-Prior Deference to Chance.

What is the general form of the SI rule? We'll use "concrete center" to unambiguously describe the centers other than the null center. Isaacs, Hawthorne, and Russell (2022, app. B) proposed the following:

> *Self-Indication with Indifference* (SII)*:*
>
> Let N be the maximum number of concrete centers in any world.
> For any concrete center c in world w, $\Pr(c \mid w) = 1 / N$. Whatever remains of the probability assigned to a world, once the concrete centers have received their probabilities, goes to null.

SII puts a larger fraction of a world's ur-prior into null as the world's concrete population decreases, which means that I am less likely to exist given a low population world than given a high population world. SII also entails CI (or more

precisely, it gives equal ur-prior to every *concrete* center in a world—which may not equal the probability assigned to the null center in that world).[24]

To combine the idea of SI with *Center Bias*, we want a rule that implies that how likely I am to exist depends not only on the *quantity* of centers but their *quality*. All else equal, I'm less likely to exist in a world the more deceived centers there are.

How should we formulate the biased version of the SI rule? Here we can avail ourselves of the discount factor presented earlier.

Self-Indication with Bias (SIB):

Let N be the maximum number of concrete centers in any world.
For any concrete center c in world w that is undeceived, $\Pr(c|w) = 1/N$. For any concrete center c in world w that is deceived, $\Pr(c|w) = (1/N) \times$ (the discount factor). Whatever remains of the probability assigned to a world once the concrete centers have received their probabilities goes to null.

Assume, for example, that the discount factor is .1 and the prior assigns equal probability to worlds w_1 and w_2 and there are two deceived centers at w_1 and one undeceived center at w_2. Then the probability assigned to the center at w_2 will be ten times the probability assigned to each center at w_1, so the ur-prior will be such that Pr(I exist (concretely) in w_1| I exist (concretely) in w_1 or w_2) = 5/6.

It is clear enough how this will apply to our puzzle case. Assume the discount factor is .1 and that there is *Ur-Prior Chance Deference* so that the prior probability is split between the two worlds. Then the two undeceived centers on Monday get the same probability, which will be the probability of that world divided by N, which is 2 in our very simple case.[25] The evenhanded treatment of the Monday centers means that *Future Chance Deference* is also respected. Meanwhile the probability given the undeceived Tuesday center is given by the probability of that world divided by 2 (ie. ¼) and the deceived Tuesday center will

[24] Although SSI entails that null gets assigned zero probability in the most populous world, that is not crucial to the view. What is crucial is just that the more concrete centers there are in a world, the smaller the probability assigned to null in that world. (See also note 23).

[25] In general, to get the unconditional ur-prior of a center, Pr(c), from the conditional ur-prior, Pr(c|w), which SII and SIB specify, use the ratio formula to find Pr(c&w). When c belongs to w this will be Pr(c), and when c does not belong to w this will be 0.

have its probability discounted according to the discount factor. Some probability will inevitably go to null. It is the *Conditional Certainty* premise that fails, claim (3) from above. (In the toy model, if the discount factor is .1, nothing goes to null in the world without deceit, 9/40 probability goes to null in the world with deceit). Meanwhile, for the Sleeping Beauty version, we again get the result that both *Future Chance Deference* and *Ur-Prior Chance Deference* are preserved and *Conditional Certainty* fails.

So, the discount factor model is one example of how the Bias View can be developed so that it defers to future chance. SI is not the only rule that Isaacs, Hawthorne, and Russell (2022) describe. Let's see what happens if we adjust the *Self-Sampling* rule (SS) to incorporate a discount factor. First, let's look at the basic SS rule.

6. The Bias View without *Deference to Future Chance*: the SS rule

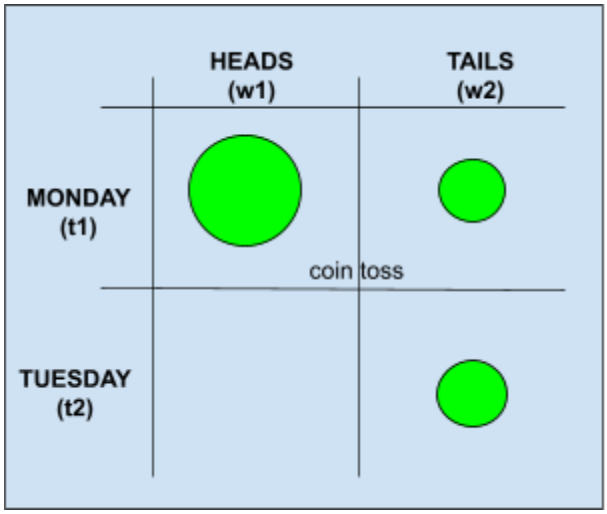
Here is how Isaacs, Hawthorne, and Russell (2022) state the general form of the SS rule. (The key feature of their rule is that the prior is certain of 'I exist' conditional on 'Someone exists'.)

Self-Sampling with Indifference (SSI)*:*

Let n_w be the number of concrete centers in world w.
For any concrete center c in world w, $\Pr(c|w) = 1/n_w$.
If a world w has 0 concrete centers, then $\Pr(\text{null}|w) = 1$.[26]

In the *very simplest* versions of Sleeping Beauty, *Self-Sampling with Indifference* recommends the so-called the *halfer* view. Assuming our SSI advocate respects *Ur-Prior Chance Deference*, then they will treat our toy Sleeping Beauty model like this:

[26] Again there are all sorts of other ways of achieving the same effect among posteriors. Suppose you didn't quite want to go in for certainty of your existence conditional on someone existing but instead you want the more cautious view that (when defined—ignore worlds with zero concrete centers here) the conditional probability of your existing given that someone exists is insensitive to the number of concrete centers. Then one could for example assign half the probability of every world with concrete centers to a null center and split what is left across the concrete centers. So perhaps the generalized slogan for the SS approach is best put as "so long as there are concrete centers, differences in the number of concrete centers makes no difference to the prior likelihood of my existing at a world".



The view assigns Heads and Tails equal ur-prior, and since there is no null center it assigns all the probability for Heads to Monday and since Monday and Tuesday in w_2 are on a par, it assigns the same probability to Monday-and-w_2 and Tuesday-and-w_2. However, the view denies the claim that we labeled (1) above:

(1) *Deference to Future Chance:* Pr(Heads|Monday) = Pr(Tails|Monday)

The halfer view instead says that

(6) *Violation of Deference to Future Chance:* Pr(Heads|Monday) > Pr(Tails|Monday).

The SS rule does not have the distinctive property, which SI has, that "I exist" confirms possible worlds in proportion to the size of their populations. Instead it has the Doomsday property: evidence E favors worlds where a relatively smaller proportion of people in that world have E. The Doomsday effect is vivid in a slightly enhanced version of Sleeping Beauty where there is a Sunday center that experiences red prior to the Monday routine. To simplify, assume there are no other worlds or centers. Then, in the two center world (the Heads world), assuming *Ur-Prior Chance Deference*, each center gets one quarter, and in the three center world (the Tails world) each center gets one sixth. If you wake up and experience green, and conditionalize, you will be 4/7 in Tails. (In this way probabilities in Sleeping Beauty are extremely sensitive to population size outside the experiment.

As the non-greens increase in equal proportion, the probabilities for Sleeping Beauty will move from the half distribution to distributions closer and closer to the thirder distribution.) And what if you wake up on Sunday and experience red? Then you should be 3/5 in Heads. So *even before the experiment begins* you have biased probabilities about the outcome of a fair coin flip. A stark illustration of the Doomsday phenomenon and another failure of *Future Chance Deferenc*e.

SSI entails *CI*, so we again need to modify it if we want to make it consistent with the *Center Bias*. Here's how we can do that:

Self-Sampling with Bias (SSB):

Apply the following procedure:

- For any concrete center *c* in world *w* that is undeceived, assign it 1.
- For any concrete center *c* in world *w* that is deceived, assign it the discount factor.
- Renormalize so that the numbers assigned to undeceived and deceived centers in w add up to Pr(w).
- If a world *w* has 0 concrete centers, then Pr(null|w) = 1.[27]

A simple example: suppose the ur-prior assigns w a probability of 1/3 and there are two centers in w: one deceived one, one non-deceived, and the discount factor is .1. Then the procedure requires us to find numbers to assign to the non-deceived and deceived so that they (a) add up to 1/3 and (b) are in the ratio 10-1. The procedure recommends assigning 10/33 to the non-deceived and 1/33 to the deceived center. It should be obvious enough that this will not improve things as far as deference to chance goes. Return to our first puzzle. Assuming a discount factor of .1 and *Ur-Prior Chance Deference*, the rule recommends a prior probability of 20/44 in Monday&Heads, 2/44 in Tuesday&Heads, 11/44 in Monday&Tails and 11/44 in Tuesday&Tails. So the probability of Heads conditional on Monday is 20/31, a failure of *Deference to Future Chance*.[28]

[27] There are also hybrid views according to which, while the sheer numbers of centers makes no difference to the likelihood that one exists, the proportion of deceived and non-deceived centers does. We won't explore such views here.

[28] One way to make this vivid is to imagine that in the Monday rooms one experiences red and in the Tuesday rooms one experiences green. Then, if you wake up and see red, you should be 20/31 confident that a future coin will land Heads.

One final observation. In our experience conversing with Self Samplers, defenders of the view sometimes try to soften the blow of failing to defer to future chance, by claiming that, in *realistic* cases, when the numbers of observers is sufficiently large (in contrast to the toy Sleeping Beauty case), they can still approximate *Deference to Future Chance*. (Here is not the place to evaluate how satisfactory that is.) To extend that defensive strategy in the presence of bias amongst centers, one would have to say that in *realistic* cases, there will not be interestingly significant asymmetries in delusion in the future that turn on chance events.

This completes our survey of SI and SS. Readers will be aware that there is another kind of rule discussed in the literature—the *Compartmentalized Conditionalization* rule—that Hawthorne, Isaacs and Russell also model. That rule turns out to be rather uninteresting in the context of bias and so we postpone discussion of it to an appendix.

7. In General, Which Rules Permit *Deference to Future Chance*?

So, we have looked at two ways of developing *Center Bias*. We can develop it using either SI or SS. And we've learned that an SS ur-prior that defers to chance will result in violations of *Deference to Future Chance*.

In this section we present some results about which rules for assigning ur-priors to centers can respect *Ur-Prior Chance Deference* and *Deference to Future Chance*. (Readers uninterested in this topic can skip to the next section without losing the main thread.)

We can partition the space of options based on whether they accept or reject each of the following two claims:

> *Existence Irrelevance:* For any qualitative proposition Q, the ur-prior, Pr, is such that Pr(Q|I exist) = Pr(Q|somebody exists).[29]

[29] Here we can think of a qualitative proposition as represented by a set of centers with the following profile: if a center from a world w is included in the set then every center from every world that qualitatively matches w (including w itself) is included in the set.

Evidence Irrelevance: For any qualitative proposition Q and any total experiential state *e*,[30] the ur-prior, Pr, is such that Pr(Q|I have *e*) = Pr(Q|somebody has *e*).

Here's what the views we've surveyed say:

Self-Indication rejects both *Existence Irrelevance* and *Evidence Irrelevance*.

Self-Sampling accepts *Existence Irrelevance* and rejects *Evidence Irrelevance*.[31]

Compartmentalized Conditionalization rejects *Existence Irrelevance* and accepts *Evidence Irrelevance*. (This combination is distinctive of *Compartmentalized Conditionalization*, which, as promised, we shall discuss further in an appendix.).

The remaining option is to accept both *Existence Irrelevance* and *Evidence Irrelevance*. But, as it turns out:

Result 1: There is no prior compatible with both *Regularity, Existence Irrelevance* and *Evidence Irrelevance*.[32]

We can also show that:

---

[30] For simplicity we here assume that when you are in total experiential state *e*, your total evidence is that you are in *e*. The principle can easily be adapted to alternative conceptions of total evidence. See note 2.

[31] Let's look at *Existence Irrelevance* first. According to *Self-Indication*, I'm more likely to exist in more populated worlds. So Pr(the world is highly populated|I exist) does not equal Pr(the world is highly populated|somebody exists). In contrast, the simple version of *Self-Sampling* recommends an ur-prior according to which I exist in all and only the worlds in which somebody exists. So the propositions "I exist" and "somebody exists" will be treated as equivalent. Now let's look at *Evidence Irrelevance*. Both *Self-Indication* and *Self-Sampling* reject *Evidence Irrelevance*. To see why, consider just two worlds, each with three centers. In w_1, there are two centers that experience green and one that experiences red. In w_2, there are two centers that experience red and one that experiences green. According to both *Self-Indication* and *Self-Sampling* $Pr(w_1|\text{I experience green}) > Pr(w_1) = Pr(w_1|\text{somebody experiences green})$.

[32] Proof: Suppose there are two worlds, one with a red center, and another with a red center and a blue center. *Existence Irrelevance* entails that null gets the same probability in each world. Regularity requires the blue center to get some probability. But that means that the ratio of probabilities assigned to the two red centers will not be the same as the ratio assigned to the two worlds. And that in turn means that the probability of the second world conditional on "I am red" will not be the same as conditional on "the I am such that someone is red". Result 1 implies that on at least some ways of interpreting the "Relevance Limiting Thesis", this thesis is incompatible with ur-prior conditionalization.

Result 2: Any ur-prior that respects *Ur-Prior Chance Deference, Deference to Future Chance*, and *Regularity*, violates both *Existence Irrelevance* and *Evidence Irrelevance*.[33]

It turns out, then, that once we endorse *Ur-Prior Chance Deference*, then (setting aside departures from Regularity) the only way to defer to future chance is to treat indexical information about *me* very differently from how we'd treat qualitative information about a mere *somebody*.

8. Some deeper challenges to the bias approach

So far we've spelled out some options for formalizing approaches to self-locating probabilities in ways that allow for bias over centers, and we've shown how those options interact with additional constraints one might want to impose (like *Deference to Future Chance*). In the course of laying out these options, we've assumed an ur-prior that is biased against centers that we've called "deceived," but we've intentionally said very little about what counts as deception because "deceived" was primarily acting as a placeholder for whatever kind of bias you might want to impose on centers. All we assumed is that whichever centers are categorized as "deceived" are discounted in the ur-prior. But in this section we want to show that an epistemologist interested in applying *Center Bias* to address certain skeptical concerns—like Boltzmann Brain skepticism—must be very careful in how they think about which centers get the "deceived" status/get discounted in the ur-prior.

Here's a warm-up example to illustrate. Suppose you're sitting in a room that you're almost certain has white walls, and you're blindfolded. You're about to

[33] Proof: Our earlier version of the Sleeping Beauty problem shows why *Deference to Future Chance* requires denying *Existence Irrelevance*. For, our argument showed that, assuming *Ur-Prior Chance Deference*, the only way to maintain *Deference to Future Chance* is to assign the null center different probabilities in the Heads world and the Tails world. But this requires taking your existence to be relevant to the probability of a qualitative proposition. To see why we must also deny *Evidence Irrelevance*, consider the following pair of worlds: in both worlds there will be a green center on Monday, and on Monday night a coin flip will determine what happens on Tuesday and Wednesday. In the Heads world, there's a green center on Tuesday and a red center on Wednesday. In the Tails world, there are red centers on both Tuesday and Wednesday. If *Evidence Irrelevance* holds then Pr(Heads|I experience green) = Pr(Heads|somebody experiences green) = Pr(Heads) = 0.5 (because of *Ur-Prior Chance Deference*). But note that in the Heads world, the 0.5 probability must be split between the Monday and Tuesday green centers, and in the Tails world it all goes to the Monday center. From this it follows, assuming Regularity, that Pr(Heads|Monday) < Pr(Tails|Monday) which contradicts *Deference to Future Chance*.

press a button labeled “red light” that you’re sure will illuminate the wall with red lighting. You press the button, open your eyes and, as you expected, the wall (which is indeed white) appears red. Do you count as “deceived” because your visual experience does not accurately represent the color of the wall? (Similarly, we might ask, does Neo deliberately re-entering the Matrix in the later parts of the movie count as “deceived”?) If the answer is yes, then, according to *Self Indication with Bias*, the center sitting in front of a white wall but having a visual experience as of a red wall after pushing the “red light” button will be discounted. Depending on the degree of discount this might mean that the person who is certain they are shining red lighting on the wall ends up very confident that the wall is red. If this result is to be avoided, then “deception” can’t be interpreted as a matter of mere divergence between the content of perceptual experience and reality.

In response to this case, the following proposal may look tempting: *a center is **deceived** to the extent that the rational credences, given its evidence, are inaccurate* (as measured by some standard scoring rule). For example, it’s natural to think that the brains in vats who have no inkling of being envatted should count as “deceived” because they are rational in being highly confident in the contents of their experiences and those rational credences are highly inaccurate. In contrast, the person in a redly lit white room, who rationally thinks they are in such a room, shouldn’t count as deceived because, although their experiences don’t match reality, their rational credences do.

The problem with this approach is that it’s circular. The *Center Bias* theorist is giving us an account of rational credence, and part of that account is the claim that certain centers—the “deceived ones”—should be discounted in the ur-prior. If we ask which centers count as “deceived,” the response can’t be noncircularly stated in terms of the credences that are rational given the center’s evidence—for the account is meant to be telling us which credences those are.

Perhaps then centers should be discounted to the extent that, *in the absence of discounting*, the rational credences in external world propositions given that evidence are inaccurate. But if “without discounting” means “with Center Indifference”, then this approach will yield the result that Boltzmann Brains with

evidence like ours won't be discounted very much, which means *Center Bias* won't be able to help with Boltzmann skepticism.[34]

None of this is enough to show that there's no epistemological account of deception that can avoid the consequence that you should increase your confidence that the wall is red after you press the button that says "red light", while also delivering the verdicts needed to avoid Boltzmann Brain skepticism. But providing such an account will be challenging. There may be radically different ways to sort all the centers in the desired way. For example, consider a physics-first way of defining the centers we should be biased in favor of: we could just declare that centers that live in the earlier epoch of the universe are all far likelier than any of the later centers, which we heavily discount. That would imply that you're unlikely to be a Boltzmann Brain. This is a very simplistic brute force solution to the problem, and we mention it not to recommend it, but to demonstrate one kind of avenue for developing such a solution.[35] We don't have a general and explanatorily satisfying proposal that we're ready to present and defend for how to do that.

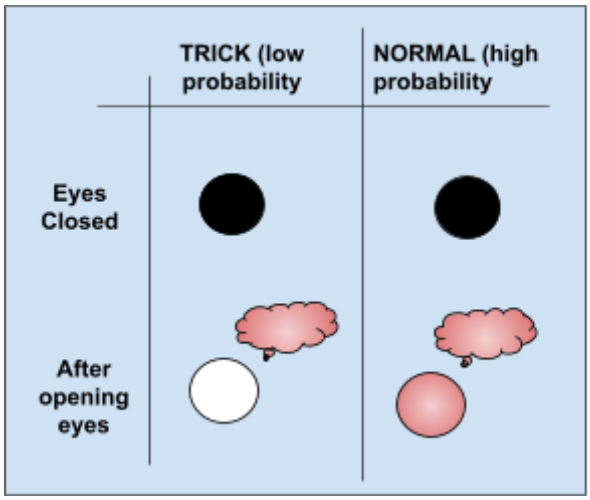
In what follows we'll describe some cases which seem to show that the defender of *Center Bias* will violate *Reflection* and, also, be susceptible to a kind of epistemic self-manipulation. One possible strategy for defending *Center Bias* against these cases would be to deny that some of the centers in the puzzle cases that we are going to label as "deceived" really ought to count as deceived. But given the difficulty of providing such an account, it's very unclear whether such a strategy can succeed.

Let's start with *Reflection*. *Reflection* says that your credences at a given time should match the expectation of your rational credences in the future. It has the consequence that if you're certain that in the future your rational credence will be c, it ought to be c now.

In a *Center Bias* context, *Reflection* violations can crop up in quite ordinary cases. Suppose that before you open your eyes you know you'll have an experience as of a red wall.[36] You're almost certain it will be veridical but you think there is a miniscule chance that the wall has trick lighting and your experience will be

[34] Assuming CI, agents with evidence like ours will be largely agnostic about external world matters, and so their credences about such matters won't be particularly inaccurate (see our discussion of Elga (2025) in section 1 above).
[35] See Dorr and Arntzenius (2017) for a more sophisticated physics-first approach.
[36] We could easily modify the case to be one in which you're uncertain what kind of experience you'll have when you open your eyes by including additional worlds. We stick to a minimal set of worlds with only red experiences for simplicity.

non-veridical. Call the two worlds under consideration NORMAL and TRICK. The diagram below represents these two worlds. (The upcoming two diagrams are intended to be non-committal about how probabilities should be assigned to centers within a world – the size of the centers in the diagram is not indicative of their probability.)



Let "Pr" be the ur-prior as usual. If we think that the only center in either of these worlds that is deceived is the center that hallucinates a red wall after opening their eyes, then the following will be true given *Center Bias*: Pr(TRICK|Red Experience) < Pr(TRICK|Eyes Closed Experience). This is because, assuming *Center Bias*, neither of the Eyes Closed centers will be discounted (neither is deceived), but one of the centers having a red experience will be. This means that regardless of how confident one should be in TRICK when one's eyes are closed, one should be even less confident in TRICK once one experiences red. (This is true regardless of whether one accepts *Self Indication with Bias* or *Self Sampling with Bias*).

In this case, the adherent to *Center Bias* knows that, upon opening their eyes, their credence in TRICK will rationally be lower than it is currently, which constitutes a violation of *Reflection*.

The thirder in Sleeping Beauty (the view recommended by *Self Indication*, with and without bias) also faces a *Reflection* violation.[37] But the technology

[37] *Self Sampling* (with and without bias) recommends that Beauty assign credence of 0.5 to heads when she wakes and so doesn't face the well-known *Reflection* violation that *Self Indication* faces. Instead, *Self Sampling* violates *Deference to Future Chance*.

required to generate *Reflection* violations with indifference across centers involves compromised memory, whereas in this case, there is a *Reflection* violation even in the absence of the possibility of compromised memory.[38]

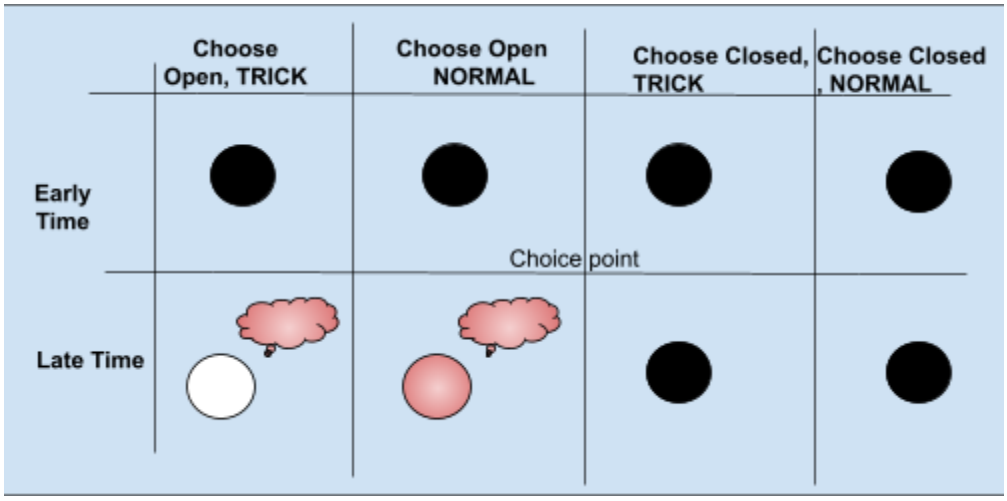
Alongside the danger of violating *Reflection*, there is related danger of allowing a problematic kind of epistemic self-manipulation. Roger White (ms.) has shown that the thirder view in Sleeping Beauty (already well-known for violating *Reflection*) permits such self-manipulation.[39] *Center Bias*, unfortunately, seems to also permit it, and here too, without any memory loss.

Take the setup for the *Reflection* violation but now suppose the agent has a choice to keep their eyes closed for the entirety of their lives (and so not risk being deceived). In this version of the case we have four worlds, distinguished both by whether the lighting in the room is normal lighting or trick lighting, and by whether the agent chooses to open their eyes. (See diagram below.) If the only center in this story that is deceived is the center whose eyes are open in the trick-lighting world, then for the same reasons described in the discussion of *Reflection*, it will turn out that by choosing to open your eyes you can increase your confidence that the lighting is normal.[40]

[38] In the standard Sleeping Beauty setup, Beauty wakes twice in the Tails world. In a variant, instead she has a duplicate created in the Tails world. In the variant, Beauty's own memory is not tampered with, but the duplicate's memories (in particular the memory of existing on Sunday) are false. Both cases produce a violation of *Reflection* on the thirder view, and both involve tampering with a person's mnemonic faculties—either erasing true memories or installing false memories.

[39] Bernhard Salow (2018) shows that externalist views that reject requirements on transparency of evidence are committed to the rationality of what he calls "intentionally biased inquiry." An intentionally biased inquiry is one in which, as a result of performing the inquiry, your credence could go up, but it definitely won't go down. The kind of self-manipulation case we'll describe here is a particularly strong version of intentionally biased inquiry. (By performing the inquiry your credence is certain to go up).

[40] Making a choice that will predictably change your credences isn't always problematic. Suppose for example that Buridan's ass is curious about whether he's the kind of donkey that will go left or the kind that will go right. He can find that out by going left. The Buridan's ass case is not problematic because, all along, the ass thinks that the question of whether he's the kind of donkey that goes left is probabilistically *dependent* on the decision about which direction to go. What's problematic about the case under discussion in the text is that before I decide whether to open my eyes, I regard the decision of whether to open my eyes as probabilistically *independent* of the kind of lighting in the room. But then I make the choice to open my eyes and increase my probability that the room has normal lighting.

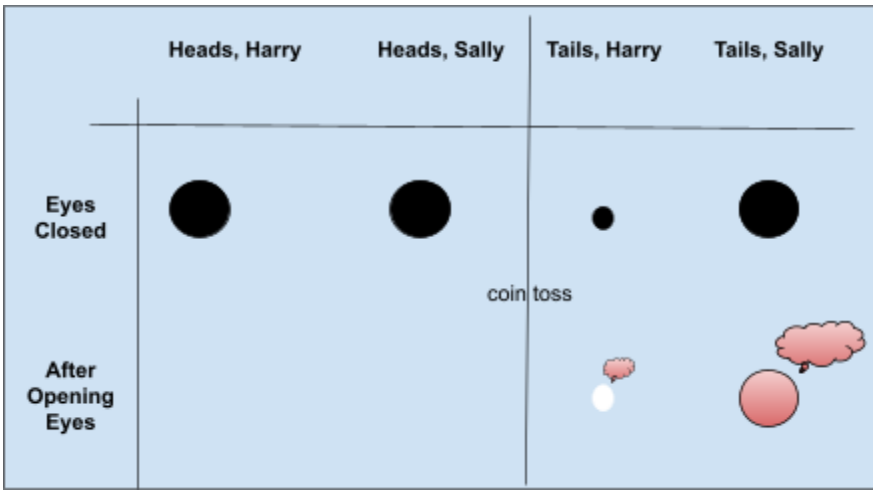


One way to address the *Reflection* violation is to make *Center Bias* more sensitive to facts about personal identity. Here's one way to do this. We have been talking as if the prior selects *centers* in a biased way. But it is quite natural to think that the priors, insofar as they exercise bias, select *agents* in a biased way—dividing them into epistemological misfits and non-misfits according to various events in their lives—but do not exercise further bias concerning *when* in the life of the agent they are: conditional on being a certain misfit, the prior is indifferent about when in the life of the misfit one is centered.

Suppose you like this "agent-first" rather than "center-first" approach to bias. Then you can avoid the *Reflection* problem by saying that the center in TRICK is discounted even before opening their eyes (the thought being "I'm unlikely to be somebody who will be deceived in the future"). If we discount *both* of the centers in TRICK, as the "agent-first' view recommends, there won't be any difference in how one compares TRICK and NORMAL across time, and so we will avoid the violation of *Reflection.* However, this route violates *Deference to Future Chance*.

To see this, imagine a slightly more complicated case with two agents. Harry and Sally (who are evidential duplicates) will wake up with their eyes closed. A coin toss will determine, whether, in a moment, Harry and Sally both die or both live (see diagram below). If they live, Harry will have a hallucinatory experience as of a red room and Sally will have a veridical experience of a red room. On the agent-first view, both the Harry centers count as misfits in the Tails world (in

virtue of one of Harry's time slices having a hallucinatory experience). But in the Heads world, no centers are misfits because no centers are deceived in that world. So there are four centers at the early time, two in each world, and just one of them gets a discount by the agent-first procedure. Now consider the moment when Harry and Sally wake up with their eyes closed. Assume SIB, where in the context of SIB we can interpret a deceived center in the technical sense as a center that belongs to a worm that's sometimes deceived in the primary sense. If Harry and Sally now conditionalize, they will not defer to the objective chance of death and so we have a violation of *Deference to Future Chance*.[41]



(Using a slightly more complex case, we can even show an inconsistency between the agent-first bias view and *Deference to Future Chance*, even without assuming either SIB or *Ur-Prior Deference to Chance*. The case is in appendix C.)

Here's one last puzzle case for *Center Bias* stemming from deception about the future. Suppose Eve knows there is a 0.000001% chance she will spontaneously vaporize. Now suppose we thought that in the world in which she *does* vaporize, her pre-vaporized self counts as "deceived." Then, according to either *Self Indication with Bias* or *Self Sampling with Bias*, that pre-vaporized

[41] Can you also avoid susceptibility to self-manipulation by adopting the agent-first view? Not exactly. But if you adopt the agent first view, the kind of self-manipulation you are subject to is not problematic. For if the center in the world with trick lighting who chooses to open their eyes is discounted at the early time, when their eyes are still closed, then the agent at the early time doesn't regard the choice of whether to open her eyes as probabilistically independent of what kind of lighting is in the room. See also note 39.

center is discounted, which means Eve won't be able to defer to the future chance of her vaporization.

As we mentioned at the start of this section, one strategy the advocate of *Center Bias* might pursue in response to these cases is to argue that, while a Boltzmann Brain or a classic brain in a vat must be discounted, the various deluded centers in these puzzle stories should not be likewise discounted. But, as we also mentioned already, we ourselves do not see how to state any principled and non-circular theory that delivers those verdicts.

9. Conclusion

The existing rules in the literature for updating on self-locating evidence all presuppose *Center Indifference*. Yet *Center Indifference* generates significant skeptical pressure: first, because familiar anti-skeptical strategies more naturally justify self-trust than trust in all of one's duplicates, and second, because some of our best scientific evidence suggests that the actual world may contain predominantly deceived agents. In this paper, we have explored two ways of developing a *Center Bias* alternative that aims to relieve this pressure—one extending *Self-Indication* and the other extending *Self-Sampling*. Once probabilities are allowed to vary across centers within a world, however, familiar tensions re-emerge: the trade-offs between *Deference to Future Chance*, *Reflection*, and related principles that arise when the number of centers vary, as they do in the Sleeping Beauty problem, reappear in the biased setting. As we pointed out, these challenges arise for *Center Bias* even without memory being compromised. We do not attempt here to resolve these tensions or to adjudicate among the competing principles. Our aim has instead been to clarify both the promise and the cost of taking bias across centers seriously, and to map the landscape of options that any satisfactory alternative to *Center Indifference* must confront.

Bibliography

- Builes, 2024, "Center Indifference and Skepticism", *Nous*
- Carroll, 2021, "Why Boltzmann Brains are Bad", in *Current Controversies in Philosophy of Science*, Routledge
- Dogramaci & Schoenfield, 2025, "Why I Am Not a Boltzmann Brain", *Philosophical Review*

- Dorr & Arntzenius, 2017, “Self-Locating Priors and Cosmological Measures”, in *The Philosophy of Cosmology*, Cambridge
- Elga, 2000, “Self-Locating Belief and the Sleeping Beauty Problem”, *Analysis*
- Elga, 2004, “Defeating Dr. Evil with Self-Locating Belief”, *PPR*
- Elga, 2025, “Boltzmann Brains and Cognitive Instability”, *PPR*
- Greco, 2012, “The Impossibility of Skepticism”, *Philosophical Review*
- Hughes, MS, “Boltzmann Brains and Self-Locating Evidence”
- Isaacs, Hawthorne, and Russell, 2022, “Multiple Universes and Self-Locating Evidence”, *Philosophical Review*
- Lewis, 1979, “Attitudes *De Dicto* and *De Se*”, *Philosophical Review*
- Lewis, 2001, “Sleeping Beauty: Reply to Elga”, *Analysis*
- Loewer, 2001, “Determinism and Chance”, *Studies in the History and Philosophy of Science*
- Meacham, forthcoming, “Fine-Tuning, Multiple Universes, and Self-Locating Beliefs”
- Pittard, forthcoming, “Advancing the Boltzmann Brain Skeptical Challenge”
- Quine, 1969, “Propositional Objects”, in *Ontological Relativity*, Columbia
- Rinard, 2018, “Reasoning One’s Way Out of Skepticism”, in McCain & Poston, *The Mystery of Skepticism*, Brill
- Rinard, 2021, “Pragmatic Skepticism”, *PPR*
- Salow, 2018, “The Externalist’s Guide to Fishing for Compliments”, *Mind*
- Schechter & Enoch, “How are Basic Belief-Forming Methods Justified?”, *PPR*
- Srednicki & Hartle, 2010, “Science in a Very Large Universe”, *Physical Review D*
- Srednicki & Hartle, 2013, “The Xerographic Distribution: Scientific Reasoning in a Large Universe”, *Journal of Physics: Conference Series*
- White, ms, “How to Persuade Yourself of (Almost) Anything”
- Wright, 1991, “Scepticism and Dreaming”, *Mind*
- Wright, 2004, “Warrant for Nothing”, *Proceedings of the Aristotelian Society*

Appendix A: The *CC* rule

The third rule that Isaacs, Hawthorne and Russell (2022) discuss is the *Compartmentalized Conditionalization* rule, or CC. As it is developed in the literature, CC makes the following distinctive claim. Think of one's evidence as a proposition of the form *I am F*, where F is a property of centres. CC makes the following distinctive claim:

> *CC's claim:* For any F that is someone's possible total evidence, Pr(Q|I am F) = Pr(Q|somebody is F).

As Isaacs et al. hint at in passing, this view is simply not available in a setting where epistemically accessibility is not S5. This point is for the most part not noticed in the literature and is worth driving home. Suppose for example, that there are intransitivities of indiscriminability among phenomenal states S1, S2, and S3. S1 is indiscriminable from S2, S2 is indiscriminable from S3, but S1 is discriminable from S3. Consider now the following simple model: in w_1 there is one center in state S2. In w_2 there are two centers, in states S1 and S3. And to keep things simple assume there are no other worlds. Given the discriminability facts, it is altogether natural to think that the agent who is in S1 has as evidence the proposition that they are in S1 or S2, that the agent who is in S2 has as evidence the proposition that they are in S1 or S2 or S3 , and that the agent who is in S3 has as evidence the proposition that there are in S2 or S3. Now the proposition that someone is in S1 or S2 is something that the priors are already certain of. Similarly for the proposition that someone is in S2 or S3. Assume a regularity principle according to which each center gets non zero probability from the ur-priors.

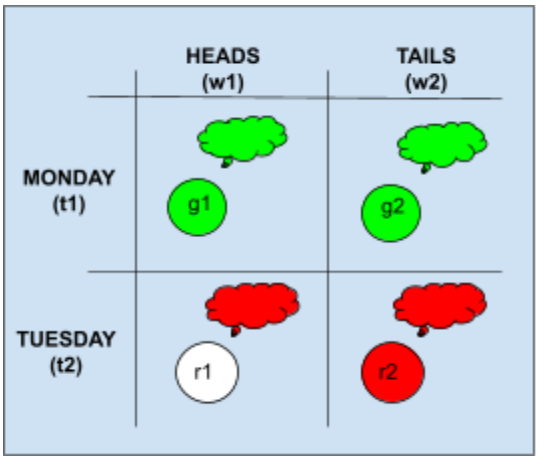
Suppose first that the ur-prior assigns the same amount to null for each world. Then the agent who is in S1 will rule out a concrete center in w_2 but no concrete center in w_1 and thus, if that agent is conditionalizing on the ur-priors, they are guaranteed to have a different credal ratio between w_1 and w_2 than the one in the priors. This contravenes the distinctive claim associated with CC. Suppose instead that the priors assign different priors to null for the two worlds. Then the agent who is in S2 will rule out null and nothing else and so end up with a different credal ratio between the two worlds than the one encoded in the prior. Thus,

assuming regularity, and even given a free hand with the priors, there is absolutely no way to conditionalize on those priors and respect the main idea. This point obviously generalizes to more realistic settings in which there is S5 evidential accessibility failure.

Given that S5 failures for evidential accessibility seem pretty much inevitable, the main idea of CC is arguably a non-starter. That said, if one makes the idealizing assumption of S5 accessibility, one can, in the context of CI, write down a rule that vindicates the main idea. (The basic trick is to increase null as a function of the shortfall in evidential *types* at a world from the maximal number of evidential types).

If one likes bias, matters are even worse. One can take the rule just gestured at and find ways to tweak it using a discount factor. The exercise is a bit complex. But the motivation for such an exercise is rather unclear since once the bias view is imposed there is no way—even in the presence of S5 accessibility—that the resulting rule will vindicate the standard motivating idea of CC (with or without *Ur-Prior Chance Deference*).

Look at the following diagram.



Let the prior assign any ratio it likes between Heads and Tails other than a ratio that puts all the probability into one world (we again assume Regularity).

If the prior ratio between being green at w_1 and being green at w_2 does not match the prior ratio between w_1 and w_2, CC's claim is violated.[42] For the ur-prior is already certain in this case that *somebody* will experience green. So if the ratio between the green centers isn't the same as the ratio between the worlds, learning

[42] This is true regardless of how you choose to assign probability to null centers.

"*I* see green" will change your credence in the worlds. Now note that *Center Bias* forces your credence in Tails-and-Red to be the same as your credence in Tails-and-Green, but your credence in Heads-and-Red to be lower than your credence in Heads-and-Green. But what that means is that if the prior ratio between the greens match the prior ratio between the worlds, the prior ratio between the reds will *not* match the prior ratio between the worlds. (If the ratio between w_1 and w_2 is the ratio between g1 and g2, and r2 = g2, while r1 does not equal g1, then elementary algebra tells you that the ratio of r1 to r2 is not that of g1 to g2 and hence not that of w_1 to w_2.) That means that if the prior conditionalizes on being red, that will yield a new posterior ratio between w_1 and w_2 that is different from the prior ratio, whereas if the prior conditionalizes on being such that someone is red, nothing happens. What we learn is that on incredibly weak assumptions, bias makes the main CC idea a complete non-starter.

The reader can by all means pretend evidential accessibility has an S5 structure, take the "CC rule" in Isaacs et al. and tweak it for bias. But since the resulting views will be utterly divorced from the standard motivations for CC, the reader will have to find a new rationale for this complex exercise. Note that there is a terminological choice here. If you take it to be definitive of *Compartmentalized Conditionalization* that "CC's Claim" comes out true, you will reject CC given bias. If you instead take Isaacs et al's rule to be definitive then (setting aside S5 issues) you will say that CC doesn't vindicate what it was designed to vindicate and perhaps invite readers to think of another use of it. Whatever one's terminological choice, the "CC rule" is out there, with or without bias tweaks, and there is the question whether it can be put to any useful work. (We doubt it.)

Appendix B: Self-Sampling, Self-Indication and Ur-Prior Chance Deference

In our presentation, the SI rule that made the conditional probabilities on "I exist" rather different to those conditional on "someone exists" was naturally paired with a thirder view of Sleeping Beauty and *Deference to Future Chance* and the SS rule was paired with a population sensitivity for Sleeping Beauty that takes a bit of getting used to, coupled with a violation of *Deference to Future Chance*. But those marriages are on the assumption of *Ur-Prior Chance Deference*. Give up *Ur-Prior Chance Deference* and one can perfectly well be a thirder and deploy the SS rule. Let's provide a feel for how to do it. Take an SI distribution. Now take the

"Cartesianization" of that distribution by conditionalizing on "I exist". (If you care about giving agentless worlds their day in the priors, take the "quasi-Cartesianization" generated by the proposition expressed by "I exist if anyone does". We shall not fuss about agentless worlds in what follows.) Now use that Cartesianization as the ur-priors for an SS view. The deliverances will be exactly as thirder friendly and *Deference to Future Chance* friendly as the original SI view.

Are these thirder friendly SS priors hard to characterize directly rather than via a detour through SI friendly priors? Not necessarily. Take a SI prior that was indifferent over worlds and imposed SI. Then that corresponds to a SS thirder friendly prior that weighted worlds according to population size, so that a population with two agents gets double the probability of a world with one. How about with *Center Bias* imposed? Let's call the *effective agent ratio* between worlds the ratio between agents with the discount factor applied (so that, for example, if the discount factor is 0.1, a world with two nondeceived agents has an effective agent ratio of 20-1 vis-a-vis a world with one deceived agent. If the SI ur-prior was indifferent over worlds, the new rule is that one weights worlds according to the effective agent ratios.

The resulting priors may initially look shocking, given how out of line they are with the chances. But they will license *Deference to Future Chance* just as much as the corresponding SI setup. And one may think it is *Deference to Future Chance* that matters—not *Ur-Prior Chance Deference*—since it is *Deference to Future Chance* that will, once agents have conditionalized, get them to have sensible views about future coin flips and so on.

The example shows that one has to be very careful about the words 'Self Indication' in the ur-prior framework. If the key definitional constraint is along the lines of 'being a thirder' then one should say it is not essential to Self Indication that 'I exist' has incremental weight vis-a-vis 'Someone exists'. Meanwhile if we (as we are in effect doing here) use 'SI' with the key definitional constraint that it is essential to SI that 'I exist' has incremental weight, one should not think that being a thirder somehow requires the SI view.

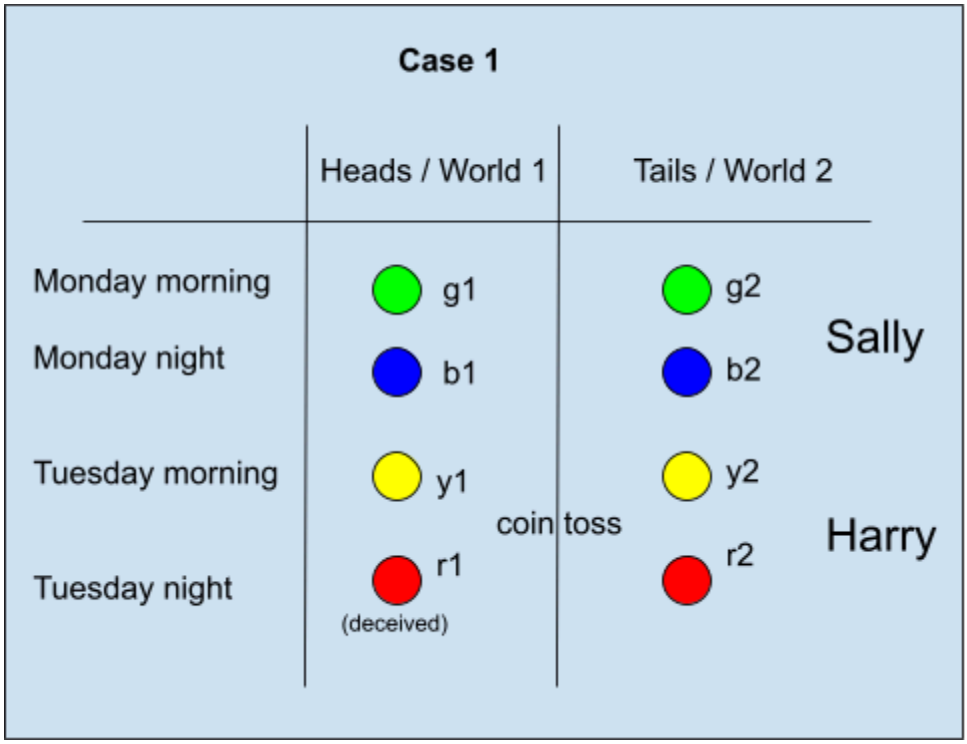
Suppose one is a thirder. Is there anything to choose between an SI prior and and SS prior that delivers thirding results? The matter is rather delicate. One might think it is just a notational convenience. One might instead think it is somehow

really important to have ur-priors that defer to chance. And the preference for an SI prior will go much deeper than that if one is a Williamsonian who wants to, say, model the prior likelihoods of counterfactual properties varying in various ways across situations where one is non-concrete and who wants to reflect on how unlikely it is for you to exist even in a world of six billion people given the mass of possible people. That said, an attitude of notational convenience will certainly be tempting to many. We cannot pursue the matter further here.

Appendix C: Agent-First Bias conflicts with *Deference to Future Chance*

In the main text we relied on SIB to raise a problem for the Agent-first Bias view. Here we give a just slightly more involved case to show that *Deference to Future Chance* is sufficient on its own to make trouble for the view.

Case 1: On Monday just Sally exists. Monday morning she veridically sees a green room. Monday afternoon she veridically sees a blue room. On Tuesday just Harry exists. Tuesday morning Harry veridically sees yellow, and Tuesday night Harry sees red. A fair coin is tossed on Tuesday afternoon to decide whether Harry's red experience will be veridical or not.



(Sally has two colors in her life just to make it vivid that her life is exactly as long as Harry's so that we can apply *Agent Bias*.)

If *Deference to Future Chance* is true,[43] then

(a) b1 = b2, and

(b) y1 = y2.

If *Agent-First Bias* is true, then

(c) r2 = y2 = b2, and

(d) r1 = y1 < b1.

But these can't all be true. (To easily see the inconsistency, take the first three claims, (a) - (c), and chain together the equalities to prove that y1 = b1 which conflicts with (d), the final claim.)

(One way to get an intuitive feel for the source of the trouble is this: we tossed the coin in the *middle* of Tuesday, which means Harry's earlier morning-time yellow centers have to be *equal* in order to defer to that coin, but,

[43] See footnote 18 above for the general statement of *Deference To Future Chance* that we apply here.

thanks to Sally's salutary presence, *Agent Bias* forces Harry's Heads centers to be less than his Tails centers.)

Appendix D: Coin Flips and Existence

Dmitri Gallow proposed to us the following principle which looks very tempting, at least initially:

> **Existence Can Be Independent of Coin Flips** If a fair coin flip makes absolutely no difference to who exists or for how long they exist, then the ur-prior probability of *I exist* is not affected by which way the coin happens to land.

$$\Pr(\textit{I exist} \mid \textit{Heads}) = \Pr(\textit{I exist} \mid \textit{Tails})$$

Abbreviate that as ECB.

Can this principle, ECB, be reconciled with the Center Bias view, CB?

ECB is consistent with the SSB view (the second version of the CB view we developed, using self-sampling). But as we saw, SSB is inconsistent with *Deference to Future Chance*.

Can ECB be reconciled with the conjunction of CB and *Deference to Future Chance*? No, the three are inconsistent. In fact, we'll show something stronger and slightly more interesting. Notice first that our main principle CB is equivalent to the conjunction of two claims: an indifference claim, CBI, and a bias claim, CBB, as stated below. We'll get inconsistency from just three claims: ECB, *Deference to Future Chance*, and CBB. (It obviously follows that ECB, *Deference to Future Chance*, and CB are inconsistent.)

> CBI: Where 'Pr' is the ur-prior, if c_1 and c_2 are centers belonging to the same world, then $\Pr(c_1) = \Pr(c_2)$ if c_1 and c_2 are both deceived or are both undeceived.

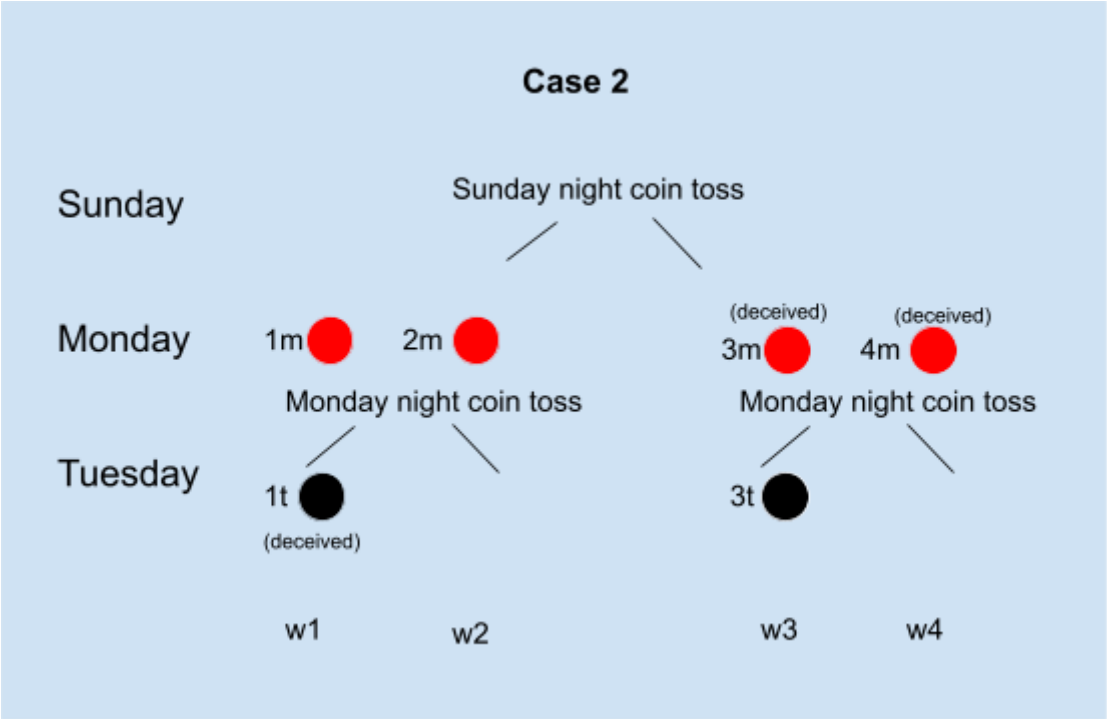
> CBB: Where 'Pr' is the ur-prior, if c_1 and c_2 are centers belonging to the same world, then $\Pr(c_1) > \Pr(c_2)$ if c_1 is undeceived and c_2 is deceived.

To show the promised inconsistency, we'll assume that if ECB true, then a slight generalization is true: if O1 and O2 are possible outcomes of a sequence of

coin flips, and O1 and O2 are not different with respect to who exists and for how long they exist, then the ur-prior probability of *I exist* is the same given either one; Pr(*I exist* | O1) = Pr(*I exist* | O2).

The following case shows the inconsistency between (generalized) ECB, *Deference to Future Chance*, and CBB.

Case 2: [*Look at the picture below!*] The case involves a single agent. On Sunday night, fair coin #1 is tossed. If coin #1 lands heads, the agent spends Monday veridically seeing red, and if it lands tails, the agent spends Monday non-veridically experiencing blue. On Monday night, fair coin #2 is tossed. If coin #2 lands tails, the agent sleeps through all of Tuesday. If coin #2 lands heads, then: if the agent saw red on Monday then on Tuesday they will non-veridically experience black, and if they agent experienced blue on Monday then on Tuesday they will veridically see black.



By ***Deference to Future Chance***, we have[44]

(a) 1m = 2m, and

[44] Again see footnote 18.

(b) 3m = 4m.

By ***ECB***, we have:

(c) 2m = 4m, and

(d) 1m + 1t = 3m + 3t.

But by ***CBB***, we have:

(e) 3m < 3t, and

(f) 1m > 1t.

And these claims are jointly inconsistent.

(*Proof sketch:* The first three claims, (a) - (c), imply that the Monday centers all equal each other. Adding the fourth claim, (d), then implies that the Tuesday centers are likewise equal to each other. This means we cannot make true both of the asymmetries that claims (e) and (f) describe.)

Of course, you could respond to all this by keeping *Deference to Future Chance* and *CBB* and giving up *ECB*. One lesson here is that it looks like there is no real advantage to having *CBB* without *CBI*, since either way, if we like *Future Chance Deference*, we will have to give up the *ECB* intuition.

ECB may have looked very tempting at first glance, but now the cost-benefit balance sheet will look different. If you think *ECB* has to stay, then you will have to also agree that the above constitutes a compelling argument against *CBB* (with or without *CBI*) from *Deference to Future Chance*.