Against Self-Location

Emily Adlam *

April 23, 2024

Abstract

I distinguish between pure self-locating credences and superficially self-locating credences, and argue that there is never any rationally compelling way to assign pure self-locating credences. I first argue that from a practical point of view, pure self-locating credences simply encode our pragmatic goals, and thus pragmatic rationality does not dictate how they must be set. I then use considerations motivated by Bertrand's paradox to argue that the indifference principle and other popular constraints on self-locating credences fail to be a priori principles of epistemic rationality, and I critique some approaches to deriving self-locating credences based on analogies to non-self-locating cases. Finally, I consider the implications of this conclusion for various applications of self-locating probabilities in scientific contexts, arguing that it may undermine certain kinds of reasoning about multiverses, the simulation hypothesis, Boltzmann brains and vast-world scenarios.

1 Introduction

Self-locating credences are used in a variety of contexts in physics and philosophy to draw substantive conclusions. For example, they play a central part in reasoning pertaining to the cosmological multiverse[1], the Everett interpretation[2, 3], the simulation hypothesis[4, 5], the arrow of time[6], Boltzmann brain scenarios[7], and so on. These applications presuppose that there exist certain privileged assignations of self-locating credences which we can use in scientific reasoning in much the same way as we would use ordinary non-self-locating credences or probabilities.

*Philosophy Department and Institute for Quantum Studies, Chapman University, Orange, CA92866, USA eadlam90@gmail.com

However, in this article, I will argue that pure self-locating credences are not sufficiently objective to bear the weight that is placed upon them in these kinds of scenarios. Of course, it is well-recognised that self-locating credences are not as objective as 'objective chances' and other kinds of probabilities employed in science: Bostrom tells us they are '*not physical chances but subjective credences*'[8]. However, in the literature on self-locating credences it is clear that they are not regarded as 'subjective' in the most radical subjective Bayesian sense, which would entail they are constrained only by the requirement of probabilistic consistency. Rather, it is assumed that there are rationally compelling ways to assign pure self-locating credences, and indeed much effort has been expended on determining the correct assignment in various problem cases[4, 9, 10, 11, 12]. That is the position I wish to criticize in this article. I will argue that pure self-locating credences are 'subjective' in the sense that they are not rationally constrained by *anything at all*, except possibly the requirement of probabilistic consistency. And I will argue that credences which are 'subjective' in this strong sense are largely not able to support substantive scientific conclusions.

I begin in section 2 by distinguishing between 'superficial' and 'pure' self-locating credences. In section 3 I will argue that assignations of pure self-locating credences cannot be rationally compelling with respect to pragmatic rationality, because in practical scenarios such as betting, an assignation of pure self-locating credences simply encodes one's practical goals. In section 4 I will argue that assignations of pure self-locating credences are not rationally compelling with respect to epistemic rationality either, since the 'Principle of Indifference' and other such principles are not a priori principles of epistemic rationality. In section 5 I will argue that the analogical strategy sometimes employed to argue for certain assignations of self-locating credences is undermined by some key disanalogies. Finally in section 6 I will discuss various scientific applications of pure self-locating credences, assessing the extent to which these applications are appropriate if there are no rationally compelling ways to assign pure self-locating credences.

Self-locating credences are used in a variety of contexts in physics and philosophy to draw substantive conclusions. For example, they play a central part in reasoning pertaining to the cosmological multiverse[1], the Everett interpretation[2, 3], the simulation hypothesis[4, 5], the arrow of time[6], Boltzmann brain scenarios[7], and so on. These applications presuppose that there exist certain privileged assignations of self-locating credences which we can use

in scientific reasoning in much the same way as we would use ordinary non-self-locating credences or probabilities.

However, in this article, I will argue that pure self-locating credences are not sufficiently objective to bear the weight that is placed upon them in these kinds of scenarios. Of course, it is well-recognised that self-locating credences are not as objective as 'objective chances' and other kinds of probabilities employed in science: Bostrom tells us they are '*not physical chances but subjective credences*'[8]. However, in the literature on self-locating credences it is clear that they are not regarded as 'subjective' in the most radical subjective Bayesian sense, which would entail they are constrained only by the requirement of probabilistic consistency. Rather, it is assumed that there are rationally compelling ways to assign pure self-locating credences, and indeed much effort has been expended on determining the correct assignment in various problem cases[4, 9, 10, 11, 12]. That is the position I wish to criticize in this article. I will argue that pure self-locating credences are 'subjective' in the sense that they are not rationally constrained by *anything at all*, except possibly the requirement of probabilistic consistency. And I will argue that credences which are 'subjective' in this strong sense are largely not able to support substantive scientific conclusions.

I begin in section 2 by distinguishing between 'superficial' and 'pure' self-locating credences. In section 3 I will argue that assignations of pure self-locating credences cannot be rationally compelling with respect to pragmatic rationality, because in practical scenarios such as betting, an assignation of pure self-locating credences simply encodes one's practical goals. In section 4 I will argue that assignations of pure self-locating credences are not rationally compelling with respect to epistemic rationality either, since the 'Principle of Indifference' and other such principles are not a priori principles of epistemic rationality. In section 5 I will argue that the analogical strategy sometimes employed to argue for certain assignations of self-locating credences is undermined by some key disanalogies. Finally in section 6 I will discuss various scientific applications of pure self-locating credences, assessing the extent to which these applications are appropriate if there are no rationally compelling ways to assign pure self-locating credences.

2 Pure versus Superficially Self-Locating Credences

It will be important in this article to distinguish *pure* self-locating uncertainty, and the associated pure self-locating credences, from a more superficial kind of self-locating uncertainty. The distinction is most easily expressed in the framework of Lewisian possible worlds, using the terminology of a 'centered world,'[13] to refer to a pair consisting of a possible world together with a 'center' within that world, which might be a place or time or a certain physically embodied observer[1]. Using this terminology, I will henceforth take it that *pure* self-locating (PSL) uncertainty refers to cases in which an observer is uncertain about which centered world they are in, out of a reference class of centered worlds which all belong to the same possible world. Whereas superficially self-locating (SSL) uncertainty refers to cases in which an observer is uncertain about which centered world they are in, out of a reference class of centered worlds which all belong to *different* possible worlds. There also exist scenarios involving mixtures of SSL and PSL uncertainty, in which some of the centered worlds belong to different possible worlds and some of them belong to the same possible world - this is the case in the famous Sleeping Beauty problem[9] - but I will not deal with these cases in this article[2].

For an example of superficially self-locating uncertainty, suppose that on days when I do not set an alarm, I do not know what time it is when I wake up. So after I have woken but before I have consulted a clock, I am in a state of uncertainty, and I may assign credences to various times that it might be. In a sense this is self-locating uncertainty, since it is about 'when' I am located. But it is only *superficially* self-locating uncertainty, since in every possible world there is exactly one time at which I actually wake up on any given morning, so all the different times that I assign credences to are in fact associated with centered worlds belonging to different possible worlds. Usually in cases of SSL uncertainty the actual 'location' is determined by a specific physical process, so for example in this case my actual location in time is determined by the set of biological processes which result in me waking up at a certain time. Hence

[1] For clarity, note that in this article an entire 'multiverse' is understood to be a single 'possible world,' because different universes in a multiverse are usually understood to be causally connectible or to have joint common causes. Thus credences to find oneself in one universe or another within the multiverse are pure self-locating credences.

[2] However, it seems clear that one consequence of the main thesis of this article is that the Double Halfer position[14] is the correct response to the Sleeping Beauty problem.

in cases of SSL uncertainty we should as far as possible assign credences which appropriately reflect relevant features of the underlying process.

For an example of pure self-locating uncertainty, consider Elga's 'Dr Evil' scenario[10], in which a person who currently believes himself to be Dr Evil receives a credible message telling him that a subjectively identical duplicate of Dr Evil has been created. This person is now in a state of self-locating uncertainty, because he does not know whether he is Dr Evil or the duplicate. And since Dr Evil and the duplicate exist within the same possible world, the two possibilities correspond to centered worlds within the same possible world, so this is *pure* self-locating uncertainty. In cases of PSL uncertainty there is not any physical process which determines the actual 'location,' so for example in this case there is no physical process by which a certain indexically individuated person is 'dropped' into Dr Evil or the duplicate: there are physical facts about the existence of Dr Evil, and physical facts about the existence of his duplicate, but then there are no further physical facts. Thus wherever PSL credences come from, they cannot simply reflect features of the physical process which determines the 'location,' because there is no such process.

Nearly all cases of self-location uncertainty that we encounter in our everyday lives are simply SSL uncertainty. For example, the most common kind of self-location uncertainty is uncertainty about what time it is, and as in the case above this can typically be understood as SSL uncertainty about the time at which some event occurs - e.g. the event of me waking up, or whatever other events may be happening around me simultaneously with my wondering about the time. Our intuitions around self-location are therefore largely driven by our experience of SSL uncertainty, so we should be wary of relying too much on these intuitions in the conceptually different PSL cases. This is important because although SSL uncertainty is much more common in ordinary life, a number of proposed applications of self-location in science do appear to be genuine cases of PSL uncertainty - for example, this is arguably the case for scenarios in which we assign credences over subjectively identical observers in different parts of the multiverse, or different branches of the Everettian wavefunction. Thus in this article I will seek to understand the epistemology of PSL cases specifically, without assuming that they work in the same way as the SSL ones.

Now, it should be noted that of course no real agent knows exactly which possible world she is in, so the definition of pure self-locating uncertainty as pertaining to set of centered worlds all belonging to a single possible world is an unrealizable idealization. Rather in realistic cases of PSL uncertainty there

is a set $\{P_1, P_2...P_N\}$ of possible worlds to which the agent assigns non-zero credence, where each P_i includes a set $\{C^1_{P_i}, C^2_{P_i}...C^M_{P_i}\}$ of centered worlds that she could be located in, and for each $X \in \{1...M\}$ there is some piece of pure self-locating information she could obtain which would allow her to conclude that she is in a centered world belonging to the set $\{C^X_{P_1}, C^X_{P_2}...C^X_{P_N}\}$ but which does not provide any *independent* information about which one of the worlds $\{P_1, P_2...P_N\}$ she is in (for now I will remain neutral about whether she might infer something about which possible world she is in from learning that she is in a centered world in the set $\{C^X_{P_1}, C^X_{P_2}...C^X_{P_N}\}$). For example, in Elga's Dr Evil case the observer is presumably uncertain about many things other than whether or not he is Dr Evil, so there will be a range of possible worlds that he could be in, all containing an individual who can be identified as that world's version of Dr Evil; so if the agent is now given reliable information that he is in fact Dr Evil, he learns that whatever possible world he should happen to be in, he is located in the centered world centered on that possible world's copy of Dr Evil, but this piece of information doesn't give him any independent information about which possible world he is in.

Throughout most of this article I will focus on the idealized case in which one is simply deliberating over centered worlds all existing in the same possible world; I take it that if it transpires that there is no rationally compelling way to assign credences in this simpler case, this likely means there is no rationally compelling way to assign credences across the centered worlds within each possible world in the more complex case described above. I will return to the more complex case in section 6.2.

2.1 Indexical Self-Reference

A key characteristic of genuine PSL uncertainty is that it involves scenarios where there are only two possible ways of singling out an individual observer from the reference class over which we are uncertain. From a first-person point of view, an observer belonging to the reference class can use indexical self-reference to identify herself; but from a third-person point of view, we can identify a specific observer only by specifying which centered world she is in, or by giving information which is equivalent to this specification. This is a necessary feature of PSL uncertainty, because if we could first identify an observer by describing some non-indexical feature F of hers and *then* ask whether she is in the centered world X or the centered world Y, we would be dealing with centered worlds

belonging to different possible worlds: one world in which the observer with property F is in centered world X, and another world in which the observer with property F is in centered world Y.

For example, suppose we perform an experiment in which two identical copies of an observer are made and the original observer is destroyed, and then one copy has her hair dyed blue and the other has her hair dyed green. After the copying but before seeing the colour of their hair, the copies are in a state of PSL uncertainty, and they may assign self-locating credences over two centered worlds belonging to the same world, respectively centered on 'the copy with blue hair' and 'the copy with green hair.' Now, the two observers will have different physical locations after the copying, so one might think that we could give a third-person description of the experiment in which we first identify a copy by their physical location, and then assign non-trivial credences over whether that copy has blue or green hair. But if that is allowed, we end up with SSL uncertainty rather than PSL uncertainty, because we are thereby assigning credences over centered worlds belonging to *different* possible worlds - one possible world in which 'the observer at the far left has blue hair' and another possible world in which 'the observer at the far left has green hair.'

Therefore in order to construct a scenario involving genuine PSL uncertainty we must insist on a description in which all of the physical facts about the relevant centered worlds are completely fixed, so the only remaining question is an indexical one: 'Which of these centered worlds am I located in?' Additionally, for genuine PSL uncertainty all the observers in the reference class must be subjectively identical, for if some observer were not subjectively identical to the others in the reference class, then we could identify her without saying which centered world she is in by simply specifying the content of her subjective experience.

The literature on self-locating credences often does not distinguish between PSL and SSL cases. This is unfortunate, because the two classes are conceptually very different. In the PSL case we can't assign any non-trivial third-person credences over the set of centered worlds, because from the third-person standpoint we can only identify observers by saying what centered world they are in, and thus the only relevant propositions that we can formulate are of the form 'the observer in centered world X is in centered world Y' - and of course this proposition will necessarily be assigned credence 1 if $X = Y$, and 0 otherwise. So PSL credences really make sense only from a first-person point of view, since we need to define them using indexical self-identification. By contrast, SSL

credences can be formulated from a third-person perspective - for example, in the case where I am uncertain about what time I woke up, exactly the same credences may be assigned from a third-person perspective, in which case they will be interpreted as credences concerning the duration of time that a human with some particular causal history and biological features will sleep. Thus SSL credences can simply inherit their values from ordinary third-person physical probabilities, but PSL credences cannot be directly derived from any ordinary physical probabilities, so they are 'subjective' in a much stronger sense than SSL credences.

Because of these conceptual differences, conflating PSL and SSL uncertainty can lead to problematic equivocations. For example, when Bostrom argues that observer-relative self-locating credences don't require some kind of special non-physical facts, he imagines a situation which at first appears to be a PSL case with a number of copies of a human brain being made, but then he argues that we can understand these credences in physical terms, as follows: '*Let Alpha be the brain that was recently in states* A_1*,* A_2 *...* A_n*. The conditional probability of A being labeled 'the bookie' given that A is one of two existing brains is greater than the conditional probability of A being the brain labeled 'the bookie' given that A is one out of eleven brains*'[8]. But if it is possible to identify observers by appeal to their past brain-states in this way, then there is in fact a non-indexical means of identifying an observer without saying which centered world they are in, and thus we are switching from PSL uncertainty to SSL uncertainty: if we assume that only one of the brains has had this particular series of brain states (which Bostrom's argument seems to require) then there is one possible world in which the brain that has recently been in the states A_1, A_2 ... A_n is the bookie, and another in which it is not the bookie, so the relevant centered worlds now belong to different possible worlds. Thus there is some equivocation in this argument: Bostrom has successfully argued that SSL credences can be understood in purely physical terms, but this does nothing to assuage the concern that *PSL* credences cannot be understood in this way.

3 Self-Locating Credences and Betting

It is commonly assumed that in at least some scenarios involving self-locating uncertainty there is some particular way of assigning self-locating credences which is rationally compelling - and most people seem to have the intuition

that this extends to PSL cases as well as SSL cases. But on what grounds could such an assignation be rational?

There is an ongoing debate about whether reasons for belief should be pragmatic, epistemic, or both[15]; here I will not take a position on this debate, but will address both possibilities in turn. Let us begin with pragmatic rationality, i.e. the kind of rationality that concerns how best to achieve our practical ends. That is, once one has decided on a set of goals, pragmatic rationality prescribes how one ought to act to achieve those goals, in light of the practical realities to which one is subject.

Assigning credences may not initially appear to be a form of action, but we can make a connection to action via decision-theoretic representation theorems [16], which show that actions taken by a rational agent making choices under uncertainty can be modelled as if that agent is maximizing utility with respect to some particular credence assignation and utility function. So we can see how pragmatic rationality may be thought of as constraining credences: it may be that in order to achieve a certain goal, it is pragmatically rational to behave as if one is maximizing utility with respect to some particular utility function and assignation of credences.

For example, suppose you are placing bets on the outcome of some probabilistic process with a set of possible outcomes $\{i\}$, and let $W_i(S)$ be the winnings you will obtain when the outcome labelled i occurs, if you bet in accordance with strategy S. Suppose also that your goal is to achieve the greatest possible winnings over a large number of trials, i.e. you aim to obtain the highest possible value for the goal quantity $G = \sum_j \sum_i W_i(S)\delta(i, O(j))$ where $O(j)$ represents the outcome of the process on the j^{th} trial, and we sum over all outcomes i, and over a large number of trials j. Then we may appeal either to empirical tests or theoretical analysis to show that in order to obtain the highest value for G, you should choose a strategy S which maximizes the quantity $W(\{p_i\}) = \sum_i p_i W_i(S)$ with respect to a certain set of values $\{p_i\}$ - that is, you should act as if you are maximizing utility with respect to the credences $\{p_i\}$, with your utility function given by your winnings. Thus we may argue that you are rationally compelled to assign credences proportional to $\{p_i\}$, or at least, to *act* as if those are the credences you assign, since acting in accordance with any other assignation will achieve worse outcomes. As argued in ref [16], showing that one ought to behave *as if* one assigns certain credences is not necessarily the same as showing that one actually ought to assign those particular credences, but I will assume here that in order to make the case that that a

certain credence assignment is in some sense rationally compelling, it is enough to show that it is pragmatically rational to act as if one assigns these credences - for after all, giving up this assumption can only make it more difficult to argue that certain PSL credence assignments are rationally compelling.

In fact, pragmatic arguments like this have been made in support of certain assignations of PSL credences. For example, in Bostrom's Dungeon thought-experiment, he argues for a certain assignation of PSL credences on the grounds that if the prisoners bet in accordance with these credences '*then 90% of all prisoners will win their bets; only 10% will lose*' and later '*a probability of 90% is the only one which would make it impossible to bet against them in such a way that they were collectively guaranteed to lose money*'[8]. Likewise Leslie argues for a certain assignation of PSL credences on the grounds that '*if every emerald-getter in the experiment betted (in accordance with these credences), there would be five thousand losers and only three winners*'[17]. These arguments aim to show that a certain credence assignment $\{p_i\}$ is rationally compelling on the grounds that choosing a strategy which maximizes the quantity $W(\{p_i\}) = \sum_i p_i W_i(S)$ will yield the highest winnings summed over all centered worlds in some reference class, i.e. this will yield the highest value of the goal quantity $G = \sum_i W_i(S)$, where $W_i(S)$ is the winnings obtained by the observer in the centered world labelled i if they bet in accordance with strategy S, and the sum is taken over the complete set of centered worlds i described in the setup of the thought experiment.

However, there is something a little odd about this approach. For unless I am unusually altruistic, when I make bets what I care about is maximizing my *own* winnings: I don't care how much is won by other observers in some reference class! So why exactly should I be be required to adopt a strategy which aims to achieve the highest possible value of $G = \sum_i W_i(S)$?

Well, the problem I face in trying to design a strategy which benefits me specifically is that, if we take it that we are dealing with genuine PSL uncertainty, all of the observers in the reference class are subjectively identical and thus there's simply no way I can design a strategy which benefits *me* more than the other observers in the reference class. Indeed, since all of the observers in the reference class are subjectively identical, they must all use the same strategy, or the same approach to choosing a strategy. For example, if the observers choose their strategy probabilistically they may end up implementing different individual strategies, but since they are subjectively identical they will necessarily all be using the same probability distribution, so they will still ultimately

end up with the same convex combination of strategies. Thus in cases of genuine PSL uncertainty it's impossible to have different agents in the reference class adopting different approaches in order to benefit themselves specifically over their peers.

So it may be tempting to argue that under these circumstances, maximizing winnings over the whole reference class is my only viable option, even if I don't care at all about the other observers in the reference class. However, this is not true. For example, let us repurpose an idea suggested by Albert in the context of Everettian probabilities[18]. Suppose that I go to sleep and five copies of me are made, and the copies are put into induced comas for one year, while my original body is destroyed. The numbers from the set $\{1, 1, 2, 4, 10\}$ are assigned at random to the copies, and the amount of nutrition provided to each successor per day is proportional to the number assigned to her, so the five successors have widely varying masses when they wake at the end of the experiment. Imagine that upon waking, and before having an opportunity to gain any information about her current mass, each successor is asked to place a bet on the value of the number that she was assigned - again, we are assuming that the copies are subjectively identical and therefore they must all adopt the same approach. So what bet should the copies make? Well, if the goal is to obtain the greatest possible winnings summed over all five copies then they should all bet '1,' i.e. they should behave as if they are maximizing utility with respect to a credence distribution assigning equal credence to all five copies. But Albert envisions the possibility of an agent who cares more about her successors with greater mass, perhaps on the grounds that 'more is better.' And likewise we can imagine the copies in this experiment deciding that they assign greater utility to winnings accrued by a more massive successor, so the goal will be to maximize winnings over mass instead, i.e. to choose a strategy S which achieves the highest value for the goal quantity $G = \sum_i c_i W_i(S)$, where $c_i = m_i \ / \sum_i m_i$, where m_i is the mass of the copy labelled i. Then if we believe the successor assigned the number 10 will end up with more than twice as much mass as a successor assigned the number 1, the best way to maximize W is to have all the copies bet '10.' That is, the copies will now behave as if they are maximizing utility with respect to a set of credences proportional to mass, $\{c_i\}$, rather than assigning equal credence to all five copies.

For a more realistic example, recall that if standard statistical mechanics is right, there may be good reasons to think the world contains many more Boltzmann brains than actual people, and thus there will likely be a large num-

ber of Boltzmann brains having experiences subjectively identical to the ones I am having now[7, 19, 20]. So it may be argued that you ought to believe that you are probably a Boltzmann brain rather than a persisting human individual. However, suppose you are required to place a bet on whether or not you are a Boltzmann brain. If your goal is to obtain the greatest possible winnings summed over all the whole reference class of individuals subjectively identical to you, i.e. to maximize $G = \sum_i W_i(S)$, then you should assign equal credences over all subjectively identical centered worlds, including both persisting people and Boltzmann brains - but is that the only plausible goal here? After all, Boltzmann brains exist only for a moment, so even if they do win the bet, they will not last for long enough to enjoy their winnings. Thus it would surely be reasonable for you to decide that you don't care how much is won by the Boltzmann brains, in which case you should adopt a strategy which aims to obtain the greatest possible winnings for persisting individuals only, excluding Boltzmann brains. Then you and your subjectively identical fellows will be aiming to achieve the highest possible value for the goal quantity $G = \sum_i c_i W_i(S)$, where c_i has the same value for all persisting individuals and zero for Boltzmann brains, and this will lead to a strategy in which you always bet that you are not a Boltzmann brain. That is, if you all adopt this strategy you will behave as if you are maximizing utility with respect to a set of credences assigning zero credence to being a Boltzmann brain, and equal credence over all persisting individuals. So in this more realistic case, it is not true that you are rationally compelled to adopt a strategy aiming for the greatest possible winnings summed over all subjectively identical centered worlds - there are clear practical reasons why you might prefer to adopt a different strategy.

3.1 Caring Measure

A notable feature of the PSL cases above is that if we take it that the 'rational' credences to assign are the credences $\{p_i\}$ such that choosing a strategy which maximizes the quantity $W(\{p_i\}) = \sum_i p_i W_i(S)$ yields the highest total value of the goal quantity G, then once we have determined the goal quantity G, the 'rational' credences $\{p_i\}$ are immediately fixed. If the goal quantity G weights all the observers in the reference class equally it will be rational to assign equal credence to all of them; if G weights the observers proportional to mass it will be rational to assign credences proportional to mass; if G excludes Boltzmann brains it will be rational to assign zero credence to all Boltzmann brains; and all

of this is true completely independently of any empirical observations we might make if we actually perform the experiment relevant to the scenario.

This point can be generalized - in principle observers in this scenario could adopt any goal quantity $G = \sum_i c_i W_i(S)$, where c_i is an arbitrary set of weights for the winnings. In fact, it can be shown that assigning negative weights c_i leads to a scenario in which a 'Dutch book' can be made[21], which in the self-locating case means that agents seeking to maximize this quantity will accept bets which are guaranteed to result in *all* of them losing money. So perhaps we can make the case that rational observers must not choose negative weights - but it would seem that any non-negative real number weights are permissible. And then it follows immediately that if we say the rational credences to assign are the values $\{p_i\}$ such that a strategy S chosen so as to maximize $W(\{p_i\}) = \sum_i p_i W_i(S)$ will yield the highest total value of G, the rational credences will always be given by $p_i = c_i \ / \sum_i c_i$.

That is, once we have chosen a goal G in a self-locating betting scenario, there is no point in doing empirical tests or any kind of theoretical analysis to decide which credences lead to the best results with respect to that goal - the credences follow immediately from the choice of goal. So if the 'rationally compelling' credences in a PSL scenario are the ones which are optimal to achieve our practical goals, then those credences are nothing more or less than a direct encoding of those practical goals. Thus in a sense, the practical function of PSL credences in decision-making scenarios is to act as something like a 'caring measure,' as proposed by Greaves in the context of the Everett interpretation[22]: from a practical point of view, PSL credences simply describe the extent to which we value winnings accrued by various observers in the reference class.

Armstrong[23] makes a similar observation, noting that the pragmatically rational way to proceed in scenarios of self-locating uncertainty necessarily depends on some specification of which agents in one's reference class one cares about, and thus he proposes an 'anthropic decision theory' characterizing the decisions that an agent under self-locating uncertainty ought to make, given a specification of how much they care about the agents in their reference class. And variations on this observation have also been made with respect to the Sleeping Beauty case[24, 25] - if we try to determine the credences for this case by appeal to empirical accuracies, the result depends on whether we count the *total* accuracy over all wakings or the *average* accuracy over all wakings, so the 'rational' credences are determined by a choice we make regarding how much

we care about various wakings[3].

But if PSL credences should really be understood as something like a caring measure, this undermines the idea that pragmatic rationality dictates how we ought to set them. For nobody is rationally obliged to care in a particular way, or at all, about the members of a certain class of subjectively identical observers; and thus if PSL credences are in practical terms just an encoding of our goals, it can't be the case that pragmatic rationality prescribes some specific way we ought to set them. Any caring measure is 'rational,' provided that it is probabilistically consistent: as Price puts it, '*Rationality may dictate choice in the light of preference, but it doesn't dictate preference itself*'[26]. So it appears that there is simply no space for pragmatic rationality to play any role in constraining our PSL credences, because as soon as we have chosen our goals, this immediately fixes how we ought to act.

Note that this criticism applies regardless of whether the observers in the reference class are distinct observers, or different time-slices of the same observer - for in the latter case we are still free to make various different choices about how to prioritize gains accrued by different temporal parts of ourselves. This can be regarded as an instance of Hedden's notion of 'Time Slice Rationality,' based on the idea that '*determining what an agent ought to believe does not require first figuring out the correct theory of personal identity over time*'[27]. That is, the way we assign credences over a reference class should not depend on whether or not the observers in the reference class are regarded as being identical, so if the credences can be understood as a 'caring measure' in a case which is naturally described as involving completely distinct observers, they can also be interpreted as a caring measure in a case which is naturally described as involving temporal parts of the same observer. In particular, it is not the case that an observer subject to self-locating uncertainty over different time-slices of herself is rationally compelled to maximize winnings over all the time-slices just because the situation happens to have been described as one of identity between time-slices, since after all she would not be so compelled if the situation were not described as one of identity.

Now, there is one possible response the proponent of PSL credences might make at this stage. This would involve arguing that the aim of science is not to

[3]As noted earlier, the Sleeping Beauty case involves a mixture of PSL and NSL credences, so the credences we arrive at in this case do have an empirical component which reflects the NSL part, but also depend on a specification of a 'caring measure' which reflects the PSL part.

produce a description of reality which is true in an absolute sense, but rather to produce a description of reality which is convenient for some set of observers. Then it could be argued that the role of PSL credences in scientific practice is not to encode physical content, but rather to encode a stipulation about which set of observers the scientific reasoning in question is intended to work for - hence, it plays the role of a 'caring measure' by dictating the set of observers that the reasoning is designed for. Now, presumably not everyone would agree that this is really the aim of science, but in any case, if this is what proponents of PSL credences intend, then conclusions drawn using PSL credences really ought to be indexed to the practical context for which they are intended - so for example, rather than offering as a conclusion that 'you should believe you are in a simulation,' proponents of the simulation hypothesis ought to say 'if you care equally about all subjectively identical copies of yourself, then you should assign a high self-locating credence to being in a simulation.' But such conclusions are seldom presented in this way: they are usually presented as if they are rationally compelling in an absolute sense. So if PSL credences are really to be thought of as simply encoding practical priorities, a number of conclusions drawn using such credences should be moderated in very significant ways.

3.2 NSL and SSL vs PSL

It is informative to contrast the pragmatics of PSL credences with NSL and SSL credences. For in the NSL and SSL cases, just fixing the goal of our betting procedure does not already determine which credences will best achieve that goal: in the probabilistic case described in section 3, if we say that the rational credences to assign are the values $\{p_i\}$ such that a strategy S which maximizes $W(\{p_i\}) = \sum_i p_i W_i(S)$ will yield the highest value for the goal quantity $G = \sum_j \sum_i W_i(S)\delta(i, O(j))$ summed over a large number of experiments, we can't immediately infer what those values $\{p_i\}$ are, since we still have to perform empirical tests or theoretical investigations to determine which credences will in fact lead to the greatest winnings.

Note that the difference is not merely that there is only one possible goal quantity G in the NSL case. For we could imagine different possible goals in an NSL scenario too: for example, rather than just maximizing the total cumulative winnings, we could discount winnings further into the future using a risk premium, so the goal quantity might be something like $G = \sum_j \sum_i W_i(S)\delta(i, O(j))e^{-|c|j}$. But nonetheless, once we have chosen such a goal,

we still typically need to appeal to experiment or theory to determine which credences do in fact yield the highest value for the quantity G.

In fact, the difference between the NSL and PSL cases has its origin in the fact that in a NSL scenario, we can identify observers from a third-person point of view without saying which outcome they observe. For example, we can choose a description according to which there is just one observer who persists through the whole experiment and try to maximize her total winnings, or we can single out a particular temporal part of that observer who performs the sixth experiment in the sequence and try to maximize *her* winnings, and so on. The point is that after we have identified the observer(s) whose winnings we are trying to maximize, there is a further empirical question about what outcome(s) they do in fact see, meaning that we can assign non-trivial credences to various possible outcomes and then determine either theoretically or empirically how close those credences are to the actual values or relative frequencies. This explains why pragmatic rationality can place meaningful constraints on these credences even after our pragmatic goals have been fixed.

By contrast, in a genuine PSL case we can only identify observers by saying what centered world they are in, or by idexical self-reference. Beginning with the first horn of this dilemma, in a PSL case the relevant 'outcome' to which we are assigning credences is simply the centered world in which one is located, so if we identify observers by saying what centered world they are in, then we have already determined their 'outcome' and thus there is no remaining empirical fact over which any nontrivial credences could be defined. So if we identify observers this way, non-trivial PSL credences cannot play the pragmatic role of reflecting knowledge or expectations about actual values or relative frequencies, as they would in the non-self-locating case: the only thing that non-trivial credences could possibly do is encode something like a caring measure over various third-person identified observers.

Meanwhile, taking the second horn of the dilemma, if we identify an observer indexically from a first person point of view, then there is indeed a further (self-locating) fact about what that observer will observe in a single experiment, but then there is no possible pragmatic justification for any non-trivial credences. One might initially imagine that we could repeat the experiment to see which credences produce higher winnings for this indexically individuated observer over time - for example, Bostrom suggests '*if we imagine the experiment repeated many times, the only way a given participant could avoid having a negative expected outcome when betting repeatedly against a shrewd outsider would be by*

setting her odds in accordance with SSA'[8]. But while this approach would work for NSL or SSL credences, it is impossible for PSL credences, because the definition of PSL credences requires that observers can only be identified either indexically or by saying what centered world they are in; whereas if it is possible to track 'the same' observer across several experiments, then it *would* be possible to identify an observer without saying what centered world they are in during the current experiment, since we could simply point to the observer who obtained certain results in previous experiments and ask what centered world that observer is in now. This is exactly why many thought experiments about PSL credences involve extreme measures such as creating a set of clones and destroying the original, in order to get rid of causal histories which might identify 'the same' observer over time. So in a genuine case of PSL uncertainty, indexically individuated observers cannot be identified across experiments, and thus each indexically individuated person only ever sees one outcome, so the empirical facts cannot possibly favour any credence other than 1 (for the actual outcome) or 0 (for all other outcomes). Thus again, if we identify observers this way, non-trivial PSL credences cannot play the pragmatic role of reflecting knowledge or expectations about actual values or relative frequencies, as they would in the non-self-locating case, so the only thing that non-trivial credences could possibly do is encode something like a caring measure over various observers.

4 The Principle of Indifference

If the claim that certain assignations of PSL credences are 'rationally compelling' cannot be understood in terms of pragmatic rationality, perhaps it is instead referring to *epistemic* rationality. But what could render certain assignations of such credences rationally compelling from the epistemic point of view?

Perhaps the most common way of defining epistemic rationality is in terms of aiming towards truth: '*An epistemically rational agent must strive to hold a system of full beliefs that strikes the best attainable overall balance between the epistemic good of fully believing truths and the epistemic evil of fully believing falsehoods*'[28]. But here we encounter an immediate problem for the idea that epistemic rationality prescribes certain ways of assigning one's self-locating beliefs. For we already saw in section 3.1 that it is impossible to demonstrate empirically that certain PSL credence assignations are more pragmatically suc-

cessful than others, and this also applies to demonstrating that certain PSL credence assignments are more likely to produce true beliefs than others: before we can say that assigning PSL credences in a certain way is a good way of coming to believe true things, we must specify for *whom* it is a good way of believing true things, and making that specification completely fixes which credences will best result in our chosen observers believing true things. Thus much as in the pragmatic case, there appears to be little room for epistemic rationality to play any meaningful role in constraining credences once we have decided on an epistemic goal.

So instead of arguing that certain PSL credences are rational in virtue of leading to true beliefs, proponents of PSL credences typically argue that the correct PSL credences are determined according to certain principles. For example, such arguments often employ something like Elga's Principle of Indifference[10] (PSL-POI) which says that '*similar centered worlds deserve equal credence*' (here the term 'similar' refers to centered worlds which belong to the same possible world and which are subjectively identical). Since we cannot hope to demonstrate empirically that the PSL-POI is a good way of coming to believe true things, the claim that we are rationally required to set our credences according to such a principle must presumably be interpreted as asserting that the PSL-POI is something like an a priori principle of epistemic rationality. But *is* it? To answer this question, it will be informative to take a brief detour to consider the status of a similar principle often employed in NSL cases.

4.1 The Non-Self-Locating Principle of Indifference

In scenarios of non-self-locating uncertainty, it is common to employ a principle which is sometimes referred to as 'the principle of indifference,' (NSL-POI) or else 'the principle of (in)sufficient reason,' which mandates that in the absence of any relevant evidence distinguishing between various mutually exclusive possible outcomes, we should distribute our credences equally between these outcomes[29]. One might be tempted to think that the NSL-POI is an a priori principle of epistemic rationality, in which case it would make sense to think its PSL analogue is also an priori principle of epistemic rationality. And indeed, there are various theoretical justifications one might offer for such an a priori principle - for example, it can be shown that the NSL-POI is a special case of Jaynes' entropy principle[30], which may be interpreted as showing that it mimimizes bias, or that it is the minimally committed option[21].

However, the idea that the NSL-POI is an a priori principle of epistemic rationality is undermined by Bertrand's paradox[31, 21], which refers to the fact that there are generally different ways of dividing an outcome space up into individual 'outcomes,' and applying the principle of indifference to different divisions will result in different probability assignations. A classic example is Buffon's needle experiment, in which a needle is to be dropped onto the floor, and the task is to calculate the probability that the needle crosses the cracks between floorboards. One way to apply the principle of indifference here would involve dividing the outcome space up with respect to the angle that the needle makes with the vertical axis; another possibility would involve dividing the outcome space up with respect to the vertical distance between the top and the bottom of the needle. And these two choices will result in different predictions for the probability of crossing the cracks, so we can't solve the problem using the NSL-POI without first making a choice about how to partition the outcome space[32].

Now, Bertrand's paradox is sometimes thought to apply only when the set of possible outcomes is continuous, so one might still try to argue that the NSL-POI is an a priori principle of epistemic rationality when the number of outcomes is finite. However, there is a sense in which the set of possible outcomes is continuous in *all* realistic scenarios, since there will always be an (effectively) continuous range of possible final states for any real physical system. In some cases, such as rolling a die or flipping a coin, there is a particularly obvious way of dividing these final states up into discrete outcomes, but the mere fact that such a description exists does not guarantee that the right way to apply the NSL-POI is to assign equal credences to the outcomes thus described: for example, when rolling a die, I can choose to characterize the outcomes as '1' and 'not 1,' but this does not entail that we should assign probability 50% to the outcome '1.' So in real physical situations we cannot simply take for granted that the outcomes as they are initially described are the right partition to use in applying the NSL-POI: we must pay attention to the details of the actual physical situation in order to decide if the way of partitioning the outcome space provided in the problem description is physically plausible. Thus Bertrand's paradox is relevant even in cases which are initially described as if they have a finite set of discrete outcomes, because these may not always be the right set of outcomes to which to apply the principle of indifference.

So what is it exactly that determines the appropriate choice of partition in cases of NSL uncertainty? Well, typically the best partition reflects relevant

features *of the process which determines the outcome* - in particular, symmetries of that process[33, 34]. For example, in the case of Buffon's needle, most experimenters will drop the needle in a way which is blind to rotation angle, since experimenters are typically not aiming at the cracks; and thus the right way to assign credences is often to use a distribution which is invariant under rotations, which amounts to applying the principle of indifference to a partition with all outcomes spanned by equal rotation angle[34]. We can verify empirically that this distribution matches the observed results for a needle dropped blindly. On the other hand, if we design the experiment differently by having experimenters deliberately aim at the cracks, then we should instead use a distribution which is not invariant under rotations, which will amount to applying the principle of indifference to a different partition.

The key point here is that the correct way of applying the NSL-POI is not knowable a priori just from an abstract description of the outcome space - we need to examine the features of the process by which the outcome is selected in order to know what probability distribution to use. As van Fraassen puts it, '*This method always rests on assumptions which may or may not fit the physical situation. Hence it cannot lead to a priori predictions. Success, when achieved, must be attributed to the good fortune that nature fits and continues to fit the general model with which the solution begins*'[34]. Of course, in the actual world it often turns out that the intuitively natural partition of the outcome space is related to symmetries in the relevant probabilistic process producing the outcomes, so it is a reasonable first guess to apply the principle of indifference to the 'natural' partition. However, this does not amount to an priori principle of epistemic rationality: the NSL-POI is a rule of thumb which helps us guess the underlying symmetries of the process producing the outcome, but it should subsequently be subjected to actual experimental investigations in which we empirically establish the actual probability distribution and/or properly determine the nature of the process which produces the outcome.

4.2 The PSL Principle of Indifference

These observations give us reason to question the status of the PSL-POI - for if the NSL-POI is not an a priori principle of epistemic rationality, why would that be different in PSL cases? In particular, one might worry that concerns along the lines of Bertrand's paradox would apply in the PSL case as well - for Elga's formultion of the PSL-POI assumes we should use a partition of the out-

come space in which each conscious observer corresponds to one outcome, but although such a partition follows naturally from the way in which we usually describe PSL scenarios, the mere existence of such a description does not in and of itself entail that the corresponding partition is always the one over which we should assign equal credences. There are certainly other possible partitions - for example, in the variant mass case discussed in section 3, one could partition the outcome space so as to have equal mass per outcome, leading to a probability distribution which assigns higher probability to observers with larger mass, rather than equal probability to all observers.

Now, Builes defends his principle of Center Indifference (a variant on Elga's principle of indifference) on the grounds that, unlike the NSL-POI case, it comes with a partition already specified: (NSL-POI) ' *doesn't specify a unique way one should partition the space of possibilities that one is indifferent over, but Center Indifference specifies that one should be indifferent between maximally specific similar centered worlds*'[7]. However, the fact that Center Indifference has been formulated in this way does not in itself guarantee that the specified partition is right. After all, there are many ways in which we *could* strengthen the NSL-POI to give a unique way of partitioning the space of possibilities in certain kinds of cases - for example, we might adopt a 'Die Principle' stipulating that in any case of uncertainty involving dice, one should always choose a partition in which each side of the die corresponds to a single outcome. But we don't typically deal with Bertrand's paradox by adopting such strengthened principles, because we recognise that in the actual world, the correct choice of partition is not something which can be known a priori - it must be determined empirically with reference to the real physical process producing the outcome. It would be a mistake to adopt the Die Principle as an a priori principle of epistemic rationality, because sometimes dice are weighted. So it is no virtue of PSL-POI or Center Indifference that they specify a way of selecting a partition, unless we can give some reason to think this particular partition is always the correct one.

Now, given the similarities between the PSL and NSL principles of indifference, one might naturally think the right way to determine the partition in a PSL case should be similar to the NSL case, so it would require us to consider the symmetries and other features of the process which produces the outcome. But here we arrive at an important disanalogy. For in a PSL case the 'outcome' - i.e. which centered world an indexically individuated observer turns out to be located in - is not produced by any physical process, since the observer is not literally dropped into one location rather than another. Therefore we cannot

determine the right partition by appealing to features of the process producing the outcome, since there *is* no such process. And therefore one main justification for using the principle of indifference in an NSL case is absent in the PSL case - it doesn't make sense to appeal to a rule of thumb designed to help us guess the underlying symmetries of the process which produces the outcome if there is no such process in the first place!

Of course, there will typically be some symmetries present in the outcome space for a PSL scenario, or in the process which produces the relevant set of observers in toto, and indeed, there have been attempts to use such symmetries to justify either the PSL-POI, or some more specific way of assigning self-locating credences. For example, some Everettians have argued that the assumed preference for applying the principle of indifference to a partition with one consciousness per outcome can be overridden by knowledge of symmetries. In particular, Sebens and Carroll argue for the Epistemic Separability Principle: '*ESP: The credence one should assign to being any one of several observers having identical experiences is independent of the state of the environment*'[2], which amounts to requiring that our credences should be invariant under transformations affecting only the environment, which are taken to be symmetry transformations. Similarly, Vaidman and McQueen adopt a principle requiring that when an experiment respects a symmetry, it will lead to a symmetry between descendants corresponding to the measurement outcomes[3].

However, note that in the NSL case the mere existence of symmetries in the general vicinity of the relevant scenario is not enough to tell us what probability distribution we ought to adopt. For example, in the outcome space for the Buffon's needle case we can identify various possible symmetries of the outcome space, including a possible rotational symmetry, which is encoded in the 'equal angle' partition, and a possible translational symmetry in the direction orthogonal to the floorboard cracks, which is encoded in both the 'equal angle' and the 'equal distance' partition. But we cannot determine a priori that the appropriate probability distribution should be invariant under one or both of these symmetries. To establish that, we have to consult the details of the actual process by which the outcome is generated in order to determine which symmetries are relevant to the way in which the outcome is actually determined - and if we change that process by dropping the needle in a different way, the appropriate probability distribution will change, even though the outcome space and the rest of the experimental setup remains the same.

And there is surely no reason why PSL cases should be any different: to

know which symmetries are relevant we cannot just look at the outcome space or general features of the experimental setup, we need to consider the symmetries of the process which *produces the outcome*. Yet we cannot do this in a PSL scenario, since there is no such process. For example, in the Everettian scenarios studied by the authors above, there is of course a branching process which creates the set of post-measurement branches and observer in toto, and this process will have certain symmetries, as identified by refs[2, 3]. But a process which produces the set of centered worlds in totality is importantly different from a process in which a specific observer is placed into one particular centered world rather than another, and clearly there is no process of the latter kind in the standard Everettian picture. And in the absence of such a process, there is no possible way of demonstrating any direct link between general symmetries of the experimental setup and the credences one should assign over finding oneself in various branches, since none of these symmetries play any role in determining which branch one finds oneself in. These symmetry-based arguments may initially look compelling, but this is at least to some extent because we are familiar with NSL cases in which it is a reasonable first guess to hypothesize that the process producing the outcomes in a given scenario may be invariant under 'natural' symmetries of the experimental setup or outcome space, such as transformations of the environment - and again, in the NSL case this is simply a rule of thumb which stands in for actual knowledge of the process producing the outcome, so its use in NSL cases offers no justification for using the same rule in PSL cases where there cannot be any such process.

4.3 Factors Specific to the Centered Case

It appears that the kinds of factors which determine the appropriate choice of partition for typical applications of the NSL-POI are not present in putative applications of the PSL-POI. So if there is nonetheless a rationally compelling way to apply the POI in PSL cases, it is most likely determined by factors which are specific to PSL cases. What could those factors be?

Perhaps the most obvious point of difference between the NSL and PSL cases is that in the PSL case outcomes are attached to centered worlds and hence to distinct consciousnesses, rather than just subdivisions of a set of physical states. So one may be tempted to argue that there is something about the nature of consciousness which means that we are rationally compelled to apply the principle of indifference to a partition with one outcome per consciousness.

But this would amount to assigning a privileged role to the consciousnesses which define the centered worlds, which physicalists at least should look on with suspicion - this approach seems uncomfortably close to letting dualism in through the back door, by treating consciousness itself as a fundamental constraint on the way in which we should partition an outcome space. After all, if we do not think that consciousness itself is fundamental or specially privileged, then 'centered worlds' are simply ways of identifying certain events or locations in the actual physical world, and there are other possible way of partitioning these actual events and locations into outcomes.

Additionally, we saw in section 4.2 that in the Everettian case, it has sometimes been argued that a naive application of the principle of indifference using 'one consciousness per outcome' is *not* always correct. So if we are willing to entertain these kinds of arguments, we are by implication accepting that there is no a priori principle of rationality which mandates that we must *always* assign equal credence to every consciousness. This suggests that we must determine the right way of assigning credences by appeal to the features of the actual physical situation - and yet, as we saw in section 4.2, the kind of features which determine the credences in the NSL case are absent in the PSL case, and it's unclear that there is any suitable replacement for them.

If we don't want to give a privileged role to consciousness in justifying the PSL principle of indifference, we might instead try to take the approach suggested by Builes, who argues for Center Indifference on the grounds that '*the usual reasons for why one might favor one possibility over another don't seem to be present in Center Indifference*'[7]. Now, an immediate problem with this is that Builes appears to be presupposing a choice of partition rather than offering any argument for it - for if we were to choose a partition which makes finer subdivisions of the centered worlds, the elements of that partition would presumably still have the property that there is no reason to favor any of them over any other, so Builes' approach makes sense only if we have already decided that a partition with one outcome per consciousness is the only option.

But in addition, is the absence of any possible reasons really favourable to Center Indifference? Builes focuses here on reasons pertaining to theoretical virtues, noting that the PSL hypothesis that I am in one centered world cannot be simpler or more explanatory than the hypothesis that I am in another centered world. But this point can be taken further - we saw in section 4.2 that in PSL cases there also cannot be any reasons arising from the nature of the process determining the outcome which would favour one possibility over

another. So in the PSL scenario, it does not just happen to be the case that there are no reasons favoring either of the outcomes - there is simply no kind of reason which could *possibly* favour one outcome over another, and thus our credences in this scenario are completely unconstrained by any 'reasons.'

Note that this is markedly different from NSL cases. In NSL scenarios, our applications of the principle of indifference do not typically have the feature that there are *no possible kinds of reasons* which could ever lead to one outcome being favored over the other - rather it just happens to be the case that in some particular instance the 'reasons' present favor all of the outcomes equally. This is important, because it means that in general, if we consider some alternative partition of the outcome space then it will typically no longer be the case that the reasons present favor all of the outcomes equally, and thus there is an objectively correct way to decide which partition is the one to which we should apply the NSL-POI - i.e. the one for which the reasons *do* favor all of the outcomes equally. Whereas in the PSL case, no matter how we partition the outcomes there will never be any reasons favoring one outcome over another, and thus there is no fact of the matter about which partition is the one to which we should apply the PSL-POI, since they are all equally good in this regard. So it is quite unclear that we ought to respond to a scenario in which there could not possibly be any reasons favouring one choice over another in exactly the same way as we respond to a scenario in which the reasons present happen to favour all outcomes equally. One might think that in the former case the right response is to simply accept that there is no rationally compelling assignation of credences, precisely because there is no possible way in which any 'reasons' could ever constrain such credences.

Builes also offers a second argument: '*Another way to support Center Indifference is by noting that violations of Center Indifference require a strange kind of forced epistemic disagreement. Suppose you deviated from Center Indifference in some way, say by being more confident in c1. Then, so long as you are self-aware, it will be implied by your evidence that you are more confident in c1. This implies that your evidential twin will also think that they are more likely to be located in c1*'[7]. However, this argument is compelling only if we take for granted that there is a uniquely rational way of assigning PSL credences. In that case, if two agents assign contradictory credences, at least one of them must be irrational. But if there is no uniquely rational way of assigning PSL credences, such disagreements do not indicate irrationality - agents may freely choose how to assign these credences, and thus the fact that their choices do not

agree does not imply that either of them is wrong. Thus this kind of argument does nothing to show that there exists a rationally compelling way of assigning credences in the first place, so it also does not prove that any particular assignation is rationally compelling.

5 Analogical Arguments

Because the PSL-POI is markedly similar to the NSL-POI, arguments for the PSL-POI can be regarded as instances of a more general strategy in which it is argued that certain distributions of PSL credences are rationally compelling in virtue of an analogy with structurally similar NSL or SSL cases. Moreover, the problems we have encountered in discussing the PSL-POI generalize to other such analogical arguments. For in using analogies between scenarios to establish what is 'rational' in one of those scenarios, it is important to first consider whether any possible disanalogies between the scenarios might undermine the comparison, and we have just seen that there is indeed a potentially fatal disanalogy between NSL/SSL and PSL scenarios: in the NSL/SSL case there is an actual process which produces the outcome over which we are assigning credences, so there are physical facts about the process which ground certain rationally compelling assignations of credences over the outcomes, but in the PSL case there is no such process, and thus we are missing the kinds of facts which often ground the rationally compelling credences in the NSL case. That is, the reasons we have for assigning certain credences in NSL cases will not in general be the same as the reasons we might have for assigning credences in PSL cases, so we should not assume that credences from NSL cases will automatically transfer across to PSL cases, even if they are structurally similar.

We can see an example of the analogical strategy in Elga's argument for the PSL principle of indifference[10], which involves a chain of reasoning such that at each step of the chain we are asked to agree that two scenarios are relevantly similar, so the rational credences for one case can be inferred from the rational credences for the other case. In particular, Elga considers scenario TOSS&DUPLICATE in which an agent Al and a subjectively identical duplicate are put to sleep, then a coin is flipped, and then both the agents are woken. Elga contends that if an agent is awakened in TOSS&DUPLICATE and then given the information I: 'either the coin landed on Heads and you are Al, or the coin landed on Tails and you are Dup,' he should assign probabil-

ity 10% to the coin landing on Heads. Elga then contends that this means that in TOSS&DUPLICATE we should assign probability 10% to $p(HEADS|HeadsAl \vee TailsDup)$ - that is, his argument depends on the claim that this conditional probability is correct even in the case where the information I is not actually provided.

But we should be careful here. The conditional probability $p(X|Y)$ is indeed usually interpreted as the credence you should have in X if you come to know Y. However, this interpretation is problematic in the case of self-locating information, because providing the information Y is a real physical process, so the scenario where the agent comes to know Y is physically different from the one in which they do not know Y. In particular, the provision of information is liable to shift us from a PSL case to a SSL case - and that is exactly what happens here. In scenario TOSS&DUPLICATE without the communication of information I there are two possible worlds in which the agent could be located: W_H and W_T in which the coin lands Heads and Tails respectively. And there are four centered worlds in which the agent could be located: (W_H, C_A), (W_H, C_D), (W_T, C_A), (W_T, C_D), with C_A corresponding to Al and C_D corresponding to Dup. Thus the agent has a mix of NSL credences over the two possible worlds W_H, W_D, and then PSL credences distributed over the centers C_A and C_D corresponding to Al and Dup respectively within those two worlds. But as soon as the agent is given the information I, his credences become SSL credences, since I rules out the centered worlds (W_H, C_D), (W_T, C_A), and thus now there are only two possible centered worlds, (W_H, C_A), (W_T, C_D), which belong to *different* possible worlds. So in this case, it makes sense to say that the rational credence to assign to $p(HEADS|HeadsAl \vee TailsDup)$ is 0.1, since at this point all we are really doing is assigning credences over possible worlds, which here is simply equivalent to assigning credences over the result of the coin flip. But this does not mean that the rational credence to assign to $p(HEADS|HeadsAl \vee TailsDup)$ is also 0.1 in the physically distinct scenario where the information I is not provided, since that is a PSL case and not a SSL case.

In fact, I would argue that the expression $p(HEADS|HeadsAl \vee TailsDup)$ is not really meaningful in the case where the information I is not provided, because this expression conditions on information which the agent cannot possibly have in a pure self-locating scenario. The only sensible interpretation of this expression is to think of it as referring not to the PSL case but to the hypothetical SSL case in which the information I is provided, but if that is so then the probability assigned to this expression does not automatically imply anything about

the credences we ought to assign to centered worlds in the physically distinct PSL case. The structural similarity between the cases is not relevant unless we have roughly the same reasons for assigning certain credences in the two scenarios, and in this example we cannot have the same reasons, since the rational credences to assign in the case where the information I is given are grounded on features of the actual physical process by which the coin is flipped and then some information is physically provided to one agent or another, whereas the rational credences to assign in the case where no information I is provided cannot be grounded on any actual physical process. Therefore we are not obliged to transfer the credences over from one case to another as Elga's argument demands.

With that said, let me acknowledge that there is a deflationary way of reading analogy-based arguments like Elga's on which the move from a SSL scenario to a PSL scenario is reasonable. That is, we could think of Elga's principle of indifference as simply aiming to characterize the way in which it is intuitively natural for agents like us to assign credences. Then it may be argued that since the self-locating aspects of the scenario TOSS&DUPLICATE are outside of our usual experience of probabilistic reasoning, the most natural thing for us to do is to reason as we would in the most closely analogous SSL case, which is probably something like the case where the information I is physically provided. So Elga's chain of reasoning may well be successful if the goal is just to arrive at a statement about the way in which it is intuitively natural for beings like us to assign credences.

But the problem is that the principle of indifference is *not* usually understood as merely characterizing natural intuitions; it is frequently invoked as a scientifically weighty principle from which significant conclusions can be drawn. Yet if it can only be understood as characterising reasoning which feels natural to us, then the PSL credences it recommends are surely not sufficiently objective to be used in these scientific applications. So while the arguments of this section do not necessarily entail that we should refrain from employing the principle of indifference when we find ourselves in scenarios of PSL uncertainty, they do indicate that we should be cautious about deriving any serious scientific conclusions from it.

5.1 Certainty

A particular subspecies of analogical arguments involves making comparisons to cases involving certainty. For example, suppose that in Case A I know that none of the subjectively identical observers that I could possibly be are simulations, while in Case B I know that one of the subjectively identical observers that I could possibly be is a simulation, while the remaining 999 are not simulations. It seems natural to say that in Case A I am entitled to be certain that I am not a simulation; but Case A and Case B are extremely similar, so surely if I am entitled to have certainty in Case A, I am entitled to have credence very close to 1 in Case B? We could then imagine moving through something like a sorites series to arrive at a more general argument for something like the principle of indifference[4].

In response to this argument, note first that that case A is not a counterexample to the claim that PSL credence assignations are constrained by nothing but probabilistic consistency. For in case A, I am still free to distribute my credences in any way that I like over the entire reference class of subjectively identical observers. It just so happens that in this case the entire class shares some property P (not being a simulation), so my conclusion that I have property P is independent of the choices I make about how to distribute my credences. But such independence of the choice of credence distribution occurs only in the specific limiting case when all the observers have the same property P: as soon as we add even one observer without that property, the credence I assign to having property P will depend on how I distribute credences over observers. Therefore cases A and B are conceptually quite different despite their numerical similarity.

Moreover, note that there are two importantly different ways of understanding the claim in case A that 'I am certain that I am not a simulation.' It could be understood as a self-locating claim, of the form P_1: 'I myself am one of the observers in my reference class who is not a simulation.' But it could also be understood as a *non*-self-locating claim, of the form P_2: 'My experiences are not compatible with being a simulation.' Assuming we are able to provide a non-indexical characterisation of the nature of the relevant experiences, credences assigned to P_2 can be understood entirely from a third-person point of view - for example, we might arrive at them in an entirely impersonal way on the basis of hypotheses about what kinds of experiences are possible for simulations.

[4]Thanks to Kelvin McQueen for suggesting this argument

And if we focus on proposition P_2 rather than P_1, Cases A and B do not look so similar. For clearly in Case B I am obliged to assign credence 0 to proposition P_2; whereas in Case A there is some room for debate, but arguably in that case I must be assigning credence 1 to P_2. For any first-person evidence I might have which provides evidence for the proposition that 'No observer subjectively identical to me is a simulation' must also be part of the first-person evidence available to all observers subjectively identical to me. So if my experiences are not incompatible with being a simulation, then it is possible for an observer to have this very evidence while being a simulation, and thus this evidence cannot be reliable evidence for the claim that no observer subjectively identical to me is a simulations. So plausibly the only way I can come to be certain that no observer subjectively identical to me is a simulation is by coming to be certain that my experiences are incompatible with being a simulation, i.e. by assigning credence 1 to P_2.

Thus there appears to be a discontinuous change in the credences assigned to the proposition P_2 between cases A and B, despite their apparent similarity. Moreover, this is true even if we increase the ratio of non-simulations to simulations in case B - the credence we assign to P_2 will not approach 1 as this ratio approaches infinity. And therefore if 'certainty' in case A is interpreted not in terms of the self-locating claim P_1 but in terms of the non-self-locating claim P_2, it follows immediately that even though I am entitled to be 'certain' in case A, I am not obliged to be 'close to certain' in case B.

So although it may seem counterintuitive to make such a strong distinction between cases where all relevant observers have some property and cases where almost all relevant observers have some property, this distinction is perfectly reasonable once we recognise that 'certainty' in case A need not be understood as just a limiting case of PSL credence - it is arguably better analysed in terms of non-self-locating claims about the compatibilty of my experiences with being a simulation, and thus it does not imply anything about how we ought to distribute PSL credences in either case A or case B.

6 Scientific Applications

Suppose it is accepted that there is never any rationally compelling way of assigning PSL credences. If this is true, it will have consequences for various common applications of PSL credences - and in particular, applications in sci-

entific contexts.

The PSL principle of indifference and similar principles like Bostrom's Self-Sampling Assumption (SSA) are commonly invoked in various scientific debates. Bostrom argues that we should see such principles as '*methodological prescriptions. They state how reasonable epistemic agents ought to assign credence in certain situations and how we should make certain kinds of probabilistic inferences*'[35]. But if these methodological prescriptions are not rationally compelling, nor susceptible to empirical verification, what exactly are their credentials as methodological prescriptions? At one point Bostrom considers the possibility that SSA may not be a requirement of rationality, arguing that even so, '*It suffices if many intelligent people do in fact - upon reflection - have subjective prior probability functions that satisfy SSA. If that much is acknowledged, it follows that investigating the consequences for important matters that flow from SSA can potentially be richly rewarding*'[35]. But it seems possible that intelligent people may have these subjective probability functions only because of intuitions that have been inappropriately transferred from SSL or NSL cases - so it may indeed be interesting to investigate the consequences of these probability functions, but we must be very careful about what exactly has been achieved in such an analysis. If the SSA, the PSL-POI and so on are not rationally compelling, and this is all just a matter of what 'feels right,' we should be cautious about using this kind of reasoning to make strong claims about reality.

After all, most of us - with the possible exception of radical subjective Bayesians - presumably believe that probabilities and/or credences used in science are responsive in various ways to physical facts. For example, statistical mechanics involves probabilities or credences, and it is clear that there are right and wrong ways to assign these probabilities or credences: there will not typically be *exactly one* uniquely correct way to assign such credences, but it is not the case that 'anything goes.' Indeed, as Myrvold points out, the idea that there is a dichotomic distinction between 'objective chances' and what are sometimes called 'subjective' or alternatively 'epistemic' probabilities or credences[36, 37] is somewhat limiting - we should make space for an intermediate conception of probability which '*combines epistemic and physical considerations*'[37]. Myrvold argues that many of the probabilities appearing in statistical mechanics belong in this intermediate category, which explains why it is reasonable to base predictions and other substantive scientific conclusions on them - they have subjective aspects but they also encode real physical content, and it is that physical content which we are accessing when we use these probabilities in

scientific reasoning. But if it is true that PSL credences are not constrained by anything other than probabilistic consistency, then they do not belong in this intermediate category, because they do not have *any* physical content. So PSL credences seem much more 'subjective' than other so-called 'subjective probabilities/credences' commonly employed in science, which raises questions about their role in scientific practice.

With that said, it should be emphasized that many 'self-locating credences' appearing in practical or scientific applications are in fact merely SSL credences, and the arguments of this article do not threaten such applications. For example, Bostrom describes an application of self-locating credences to predicting how fast cars will move in different lanes, based on treating yourself as a random sampling from the set of all drivers on the motorway[35]. This case is an instance of SSL rather than PSL uncertainty, because there is a causal history which results in you being in one position rather than another in the traffic jam, and thus different possible positions that you could have in the traffic jam correspond to centered worlds in different possible worlds, rather than different centered worlds within the same possible world. So the arguments I have made in this article don't undermine this kind of reasoning.

Thus in what follows I will focus on types of application in which what appear to be genuine PSL credences are used to draw scientific conclusions. Evidently a possible strategy one might adopt in response to my concerns would be to try to show that these cases can be understood as SSL rather than PSL. And indeed I think this would be an interesting route to explore, but I do not have space do so here, so in what follows I will simply assume that the relevant credences in these cases are in fact PSL credences.

6.1 Personal Circumstances

The first kind of application involves using PSL credences to directly draw conclusions about your personal circumstances, as in claims such as 'you are very likely to be a simulation'[4], 'you are very likely to be a Boltzmann brain'[7], or, in the Doomsday argument[35], 'you are likely to have been born at around the midpoint of the birth order of all humans who will ever exist.'

For example, arguments for the simulation hypothesis[4, 5] typically start off by asserting that we have good reasons to believe that the world contains many more simulations than actual people. Then it appears reasonable to think that large numbers of these simulations might be having experiences subjectively

identical to yours, so the PSL-POI can be invoked to argue that you should believe you are most likely in a simulation. Now, this is perfectly reasonable as long as we understand the conclusion as saying simply that 'one intuitively natural way of assigning subjective credences in this case would be to assign high credence to being a simulation.' However, advocates of the simulation hypothesis typically seem to want to say something stronger than this: they appear to be saying that there is something rationally compelling about the conclusion that you are likely a simulation, so those who deny this conclusion are involved in some kind of error. And yet the self-locating credences here appear to be PSL credences, since there is no physical process by which you are dropped either into a real person or a subjectively identical simulation. So if there is never any rationally compelling way to assign PSL credences, then there cannot be a rationally compelling way to assign a credence to being a simulation, and thus although it would be permissible to assign high credence to being a simulation, it is equally permissible to assign high credence to *not* being a simulation, and thus the simulation argument by itself does not establish very much[5].

Much the same applies to the Boltzmann brain case. If you believe that the world likely contains many more Boltzmann brains than persisting human individuals, then a straightforward application of the PSL-POI suggests you are most likely a Boltzmann brain. But again, the credences here appear to be PSL credences, since there is no physical process by which you are dropped into a real person or a Boltzmann brain. So while it would be reasonable to assign high credence to being a Boltzmann brain, it would also be reasonable to assign low credence to being a Boltzmann brain - and indeed in section 3 we saw that this would make sense from a practical point of view. Thus again, we are not rationally compelled to believe that we are probably Boltzmann brains, so this argument by itself does not establish very much.

[5]One might seek to avoid this problem by considering a reference class of simulated and non-simulated observers who are not subjectively identical, in which case we are not in a PSL scenario. But then the ratio of simulations to non-simulations is significantly less relevant to our assessment of our situation, since we can alternatively base our credences on the compatibility of our own experiences with being a simulation, without reference to other observers. Thus the simulation argument still runs into problems in this case, since it is no longer clear that the high ratio of simulations to non-simulations in and of itself entails that we must assign high credence to being a simulation.

6.2 Empirical Confirmation

The second kind of application involves using pure self-locating credences to perform empirical confirmation by means of Bayesian updating. This occurs in some multiverse scenarios[38], but is perhaps most prominent in the Everett interpretation. In that context it is often argued that mod-squared amplitudes should be interpreted as (pure) self-locating credences for finding oneself in one branch of the wavefunction rather than another, and that inhabitants of an Everettian world can perform Bayesian updating based on observed measurement outcomes by using these PSL credences in exactly the same way as we would customarily use non-self-locating probabilities[2, 3]. For example, this means that if I am considering two versions of an Everett-style theory which assign different mod-squared amplitudes to a certain measurement outcome, and then I do in fact see that outcome, I ought to update my credences to assign higher probability to the version of the theory which has a larger mod-squared amplitude for that outcome, following the usual Bayesian updating formula.

Now, one reason to think there may be something wrong with this approach to empirical confirmation follows from a view that Titelbaum calls the Relevance-Limiting Thesis (RLT)[5, 39], which suggests that learning a piece of self-locating information should never cause us to update our non-self-locating credences. Evidently the RLT entails that self-locating information cannot be used in empirical confirmation, which is all about updating non-self-locating credences. One important argument for the RLT is that approaches to belief-updating which do not uphold the RLT typically lead to extremely counterintuitive results in cases where the number of subjectively identical observers in a given world can increase over time[14, 40, 4, 41]. But nonetheless, a number of philosophers (including Titelbaum himself[5, 39]) believe that the RLT is false.

In fact, I think these disagreements over the RLT arise from a failure to distinguish properly between PSL and SSL scenarios. For there are compelling reasons to believe that the RLT is true for PSL information, but not for SSL information - indeed, this seems to follow almost by definition. If you learn pure self-locating information, then learning that information only tells you which centered world you are in from a set of centered worlds all belonging to the same possible world, so it cannot tell you anything new about *which* possible world you are in, i.e. it cannot change your non-self-locating credences. Whereas if the information you learn is only superficially self-locating, then when you learn which centered world you are in you also learn which possible

world you are in, so clearly you do have reason to update your non-self-locating beliefs.

And indeed, if we examine putative counterexamples to the RLT, at least the most obvious kinds of cases turn out to concern SSL credences rather than PSL credences. For example, in the case considered in section 2 about knowing the time upon waking, the unqualified RLT would seem to suggest that I shouldn't update any non-self-locating beliefs when I check the time and see that it is seven o'clock, but that is surely wrong - on seeing that it is seven o'clock I learn how long a certain human being slept on a given occasion, and that can potentially lead me to update various non-self-locating beliefs about the state of health and sleep hygiene of that human being (who happens to be me, but no part of the belief-updating I am doing rests upon this fact). However, checking the time gives me SSL information, since I learn whether I am in a possible world where a certain human being slept for six hours or an alternative possible world where that human slept for seven hours, and thus it is no surprise that I end up updating some non-self-locating beliefs on the basis of this information. So this counterexample supports the view that the RLT is false for SSL information, but this is perfectly compatible with the hypothesis that the RLT as it pertains to *PSL* information is correct.

With that said, the argument give above is of course an oversimplification, for as noted in section 2, in real scenarios you will never know exactly which possible world you are in, so any time you are deliberating over a range of centered worlds there must in some sense exist duplicates of those centered worlds in various different possible worlds. So the real question is, if there is a set $\{P_1, P_2...P_N\}$ of possible worlds to which you assign non-zero credence, where each P_i includes a set $\{C^1_{P_i}, C^2_{P_i}...C^M_{P_i}\}$ of centered worlds that you could be located in, and you then learn a piece of information X which tells you that you are in the set of centered worlds $\{C^X_{P_1}, C^X_{P_2}...C^X_{P_N}\}$ but which does not give you any independent information about which possible world you are in, can this nonetheless cause you to update the credences you assign over the possible worlds in $\{P_1, P_2...P_N\}$?

The RLT suggests that it cannot, but here is one approach one might take to argue that the RLT is wrong. Imagine that the worlds P_1 and P_2 have laws or symmetries which entail different assignations of PSL credences to their corresponding centered worlds $C^1_{P_1}, C^1_{P_2}$, and suppose the PSL credence to find myself in $C^1_{P_1}$ mandated by the laws of world P_1 is higher than the PSL credence for $C^1_{P_2}$ mandated by the laws of P_2. Then suppose I learn a piece of pure

self-locating information which tells me that I am in a centered world in the set $\{C^1_{P_1}, C^1_{P_2}\}$. Surely in that case I ought to update my NSL credences to assign higher credence to P_1 and lower credence to P_2, thus changing my non-self-locating beliefs? If this is right, the RLT seems to be false even for PSL information.

However, if it is true that there is never any rationally compelling way of assigning pure self-locating credences, it follows that laws or symmetries *cannot* entail anything about PSL credences, and therefore this argument does not work. For example, we examined the symmetry case in section 4.2, and concluded that the conditions for symmetries to mandate certain assignments of credences are not met in the PSL case, since in PSL cases symmetries do not ever play any role in determining the 'outcome,' i.e. the centered world in which one finds oneself. If this is right, then the symmetries that hold in a given possible world cannot possibly entail any particular PSL credences that one ought to have in that world, and much the same goes for laws, and thus the situation described above simply cannot ever occur. This suggests that the RLT as it pertains to PSL information is indeed correct: although there might sometimes be certain choices of PSL credences which feel intuitively natural given a certain set of laws and symmetries, if they are only intuitively natural as opposed to rationally compelling, then it would be a mistake to use them in Bayesian updating as if they are the same as ordinary non-self-locating probabilities, and thus learning PSL information can't cause us to change our non-self-locating credences in the way described above.

Additionally, if we agree there is never a rationally compelling way to assign PSL credences, but we think there *is* sometimes a rationally compelling way to assign NSL credences, then in order to maintain the rationality of our NSL credences we should avoid allowing them to be 'infected' by the subjectivity of our PSL credences, and thus we have good reason to adopt an approach to belief-updating which keeps NSL and PSL credences clearly distinct. That is, we should probably adopt an approach to belief-updating in which we '*first assign (NSL) credences to possible worlds and then somehow distribute those credences over the centered worlds corresponding to the possible worlds*'[41], such as the system proposed by Halpern and Tuttle[42] or Meacham's compartmentalized conditionalization[14]. And as noted by ref [41], any such approach to belief updating will automatically uphold the RLT as it pertains to *pure* self-locating information.

Bradley[43] makes a somewhat similar point, arguing that the RLT is true if

it pertains to 'Mutation - belief change in virtue of a change in the truth-value of the content of the belief' but false as it pertains to 'Discovery - belief change in virtue of the discovery of the truth of the content of the belief, where the truth-value did not change over the period of interest.' An example of Mutation is watching the hands of a clock move and changing one's beliefs about the time (because the statement 'it is now twelve o'clock' changes in value) and an example of Discovery is being uncertain about the time and then looking at one's watch. Applying the schema I have used here, we can see that examples of Mutation typically involve gaining PSL information, whereas examples of Discovery often seem to involve learning SSL information - for example, when you look at your watch, you do not just learn which centered world you are in, you learn that certain events (the event of you checking your watch, or other events going on around you as you check your watch) occur at twelve o'clock, so you learn that you are in a possible world in which those events occur at twelve o'clock rather than at some other time. Thus Bradley's way of distinguishing the cases in which the RLT is true from thsoe in which it is false would likely agree with my approach in many instances. However, Bradley is particularly concerned with cases where the uncertainty is about one's *temporal* location, and his Discovery vs Mutation categorisation does not seem so straightforwardly applicable to other kinds of cases, such as being uncertain about which one of a set of subjectively identical clones one is at a certain fixed time. So the difference between PSL and SSL cases looks like a more generalizable way to distinguish good and bad applications of the RLT.

In summary, if it is accepted that there are no rationally compelling assignations of PSL credences, this suggests strongly that the RLT is correct as it pertains to PSL information. And if this is the case, it immedately follows that PSL credences should not be used to do empirical confirmation, either in the Everettian context or in any other context. Thus this provides further reason to be wary of the use of PSL credences in scientific applications.

6.3 Anthropic Reasoning

The third kind of application involves using PSL credences in anthropic explanations. For example, it has been proposed that we can explain the apparent fine-tuning of various fundamental parameters by first assuming we are in a certain kind of multiverse, and then arguing that in such a multiverse the appropriate self-locating credence to assign to finding oneself in a universe with

fundamental parameters in the relevant range is relatively high[44, 45].

Now, it is well-known that this approach runs into problems if the multiverse in question is infinite, since in that case we must choose a measure to determine the relevant self-locating credences, and there seems to be no rationally compelling choice of measure[44, 46, 47]. However, it is commonly thought that at least in the finite case explanations of this kind can be given successfully. But if there is no rationally compelling way of assigning PSL credences, then even in the finite case we are not obliged to assign credences over universes in any particular way. From this point of view, the only real difference between the infinite and the finite case is that in the finite case there happens to be a particular choice of 'measure' (i.e. a way of assigning credences over worlds) which has a strong intuitive appeal; but the fact that a measure is intuitively appealing does not make it rationally compelling. Friederich argues that in the infinite case, '*even if some specific measure were established as physically privileged in the context of external inflation, this would not by itself show that this measure should guide our assignment of probabilities*'[44] and I would contend that in fact the same goes for the finite case - credences obtained from simply taking ratios of numbers of universes may be the most obvious choice, but that does not mean we are rationally obliged to set our credences that way.

What does this mean for explanations which rely crucially on these pure self-locating credences? The answer may depend on the view of explanation that one adopts. Certainly if one is working with a deductive-nomological or inductive-statistical approach[48], explanations which depend on PSL credences look problematic if there is no rationally compelling way of assigning such credences, for that means we will not be able to derive the explanandum either deterministically or statistically from just a set of initial conditions plus some laws of nature: we must in addition make use of a special assignation of self-locating credences which simply encodes some kind of subjective attitude, such as how much we care about various individuals. It seems doubtful that such a thing is a legitimate ingredient in either a DN or an IS explanation. Similarly, it is hard to see how we could give a satisfactory causal explanation[49] relying crucially on purely subjective PSL credences. And even in an approach to explanation based on unification[50], it's not obvious that an explanation can be considered significantly unifying if it relies on an essentially arbitrary input which does not come from any relevant theory but which simply stipulates PSL credences as a subjective attitude.

On the other hand, it is true that in certain cases there is a particular

assignation of PSL credences which feels intuitively natural, so if our main desideratum for explanations is that they should provide an intuitive feeling of understanding, perhaps the use of intuitively natural PSL credences may be acceptable. However, I want to highlight two problems which could follow from taking these 'explanations' too seriously. The first is that if we are satisfied by such explanations, this may stop us from exploring paths to more physically-grounded explanations, and then we could potentially miss out on useful insights into physical reality that would follow from such explanations. The second is that if we are satisfied by such explanations, we may be tempted to employ them in the context of inference to the best explanation - for example, this often occurs in arguments for the multiverse, where the idea that the existence of the multiverse would explain the values of certain fundamental parameters is used to argue that we ought to believe in such a multiverse[44, 45]. But if the explanation in question is the 'best' explanation only in the sense that it gives us a intuitive feeling of understanding, it's unclear that we are really justified in making strong inferences about existence from it. So although explanations using PSL credences may in certain circumstances be acceptable, if our primary focus is on achieving an intuitive feeling of understanding, we should be careful about using such explanations to motivate any stronger scientific conclusions.

With that said, it should be emphasized that the concerns I have raised in this article about the status of PSL credences do not necessarily impugn all kinds of anthropic reasoning. For example, Carter's original formulation of the anthropic principle states that '*what we can expect to observe must be restricted by the conditions necessary for our presence as observers*'[51] and this principle does not appear to depend on the existence of rationally compelling PSL credence distributions. Rather it simply mandates that as a matter of certainty we must find ourselves in a universe belonging to the set of universes obeying various conditions - and as discussed in section 5.1, certainty in this sense need not be understood as merely a limiting case of PSL credence, since it can be analysed as a third-person claim about the compatibility of certain experiences with certain physical circumstances. So although the concerns I have raised about PSL credences suggest that there may be no further fact of the matter about how we ought to assign credences over universes *within* the relevant set, that does not undermine the objectivity of the original statement that we will definitely find ourselves somewhere in this set, and thus certain kinds of anthropic reasoning will remain available even if there are no rationally compelling assignments of PSL credences.

6.4 Vast World Scenarios

The fourth kind of application pertains to vast world scenarios. I have argued that PSL credences are not suitable to play a role in most scientific applications. However, it might be objected that we have no choice but to use PSL credences in science, due to the possibility we are in a vast world. For example, Bostrom notes that modern cosmology gives us good reason to think we live in a very large universe, and that such vast worlds '*imply, or give a very high probability to, the proposition that every possible observation is in fact made. This creates a challenge: if a theory is such that for any possible human observation that we specify, the theory says that that observation will be made, then how do we test the theory?*'[35]. Bostrom thus contends that in a vast-world scenario, PSL credences are indispensable for prediction: any time an inhabitant of a vast world assigns non-trivial probabilities to outcomes of a certain event, what she is really doing is assigning PSL credences over herself being located in a part of the universe where a certain observation is made. Similarly, Srednicki and Hartle argue that PSL credences are needed to make sense of empirical confirmation in such a context: '*in a large universe the likelihood that at least one instance of our data exists somewhere approaches unity for any theory that is consistent with our data. The third-person Bayes procedure is therefore not effective for discriminating between theories in a very large universe.*' They maintain that '*a further assumption is ... needed to connect the third-person probabilities of theory with the first-person probabilities for our observations. We call this assumption the xerographic distribution*'[20]. This 'xerographic distribution' can be thought of as a distribution of self-locating credences over centered universes within a single large universe.

Now, Bostrom and Srednicki and Hartle seem to assume here that *all* predictions made without appeal to self-location must take the form 'the probability that this datum occurs somewhere in the universe/multiverse,' in which case useful predictions and empirical confirmation using purely third-person NSL propositions will indeed be virtually impossible in a vast-world scenario, or indeed a sufficiently large multiverse. However, although there may be some special cases, such as predictions of the values of constants or global features of a universe, which do involve probabilities of this form[45], most scientific predictions don't just predict that an observation will be made somewhere. Rather, scientific predictions are typically conditional and relational, involving probabilities of the form 'the probability that a certain system is in state Y at time

$t + \delta$ given that it was in state X at time t.' And to assess a probability of this kind we need not take into account how many systems of the relevant kind exist: we can simply consult the dynamics of our theory, which will typically provide us with a well-defined, non-trivial conditional probability for the state transition. Thus empirical confirmation here proceeds just as it would in a small universe, where an observation of X transitioning to Y confirms theories whose dynamics assign a high probability to this transition and disconfirms theories whose dynamics assign a low probability to this transition[6].

Of course, there is always some possibility that a low probability transition will occur and thus we will be misled by our data, but this is true regardless of whether we are in a small universe or a large one: there are certainly conceptual difficulties surrounding the nature and justification of empirical confirmation for probabilistic theories, but these difficulties need not be any more severe in a vast-world context. So for most kinds of scientific reasoning in a vast-world context, there's no need to invoke PSL credences, since the appropriate way to define credences is to simply match them to relevant features of the dynamical process connecting the states at different times, much as we saw in section 4.1 that the appropriate credence distributions for probabilistic processes typically reflect symmetries or other features of the process that produces the outcome.

One might object, as do Bostrom and Srednicki and Hartle, that in a vast-world scenario I should take seriously the possibility that I am just a Boltzmann brain or another temporary fluctuation, and if I am indeed a Boltzmann brain then it is likely that the system S that I am trying to make predictions for is not real and thus has no actual dynamics of its own. So the suggestion appears to be that in a vast-world scenario, in order to make predictions about a system S that I am observing, I must first invoke some PSL credences to convince myself that I am probably not a Boltzmann brain, in order that I can justifiably conclude that S really exists and will obey standard dynamical laws: '*if we assume a xerographic distribution ... such that we are not likely to be (Boltzmann brains), then we get a predictive, testable framework*'[20].

[6]Note that the main example used by Srednicki and Hartle in ref [20] involves checking a 'global' non-relational property - a colour - which may have a different value in different (temporal) regions of the universe. Since this property is not relational or conditional, consulting the dynamics of any given theory will not provide any prediction for it, and thus it does seem correct to say that the only way we could possibly make a prediction about this property would be to employ self-locating credences over centered universes corresponding to the different temporal regions. But this is not the kind of property that we actually use to empirically confirm real theories - in general our data takes the form of relations between properties, not individual global properties.

But even if it's possible to get the numbers to work out right (as noted in section 6.1, a naïve application of the principle of indifference would seem to suggest I *am* quite likely to be a Boltzmann brain) it's simply not obvious that this is the right way to think about what is going on in this prediction. For the idea that I must first justify the claim that I am not a Boltzmann brain before making predictions using ordinary dynamical laws has a strongly foundationalist flavour, and yet it has often been argued that foundationalism is not a viable way of conceptualizing scientific knowledge[52, 53, 54]. So we can plausibly deal with the Boltzmann brain hypothesis without appealing to self-locating credences by simply rejecting foundationalism and instead adopting a coherentist[55] or perhaps progressive coherentist[56, 57] approach, which would involve seeking to formulate an overall coherent system of beliefs which has the consequence that I am not a Boltzmann brain.

Or alternatively, we might simply take a pragmatic stance and stipulate that the possibility of being a Boltzmann brain should be ignored. For after all, if I am really a Boltzmann brain there's little point in me trying to do science at all, since I have effectively no chance of making any successful predictions. That is, in the course of doing science, we might as well begin by assuming that our circumstances are at least somewhat conducive to our doing so successfully, so we should simply take it for granted that we are not Boltzmann brains. If we ultimately end up with a coherent belief system which affirms our original assumption that we are not Boltzmann brains, then we obtain at least some prima facie justification for thinking that the original assumption was a reasonable one, and that retrospective justification as understood within a progressive coherentist framework is arguably more convincing than foundationalist arguments based on a supposedly prima facie assignation of subjective self-locating credences.

7 Conclusion

A famous dilemma for self-locating credences involves a 'presumptuous philosopher' [8] who uses self-locating credences to conclude that a certain theory of physics must be right, and then advises the physicists that they need not even bother performing the experiment to distinguish between two competing theories. Bostrom's response to this scenario is to criticize the particular way in which this philosopher arrived at these self-locating credences[8]. But the argu-

ments given this article suggest a much more general response: it is 'presumptuous' under any circumstances to use PSL credences to arrive at substantive conclusions about physics or the content of reality, because there is no rationally compelling way to assign PSL credences, so such credences are not a suitable basis for scientific reasoning. Thus in fact, if we take it that the credences involved in the presumptuous philosopher case should be understood as PSL credences, then no matter how the philosopher arrives at them he is in the wrong for trying to answer this question by appeal to PSL credences alone!

More generally, if the thesis of this article is true, then we should not expect to resolve substantive scientific questions using PSL credences, and this has important consequences for reasoning around multiverses of various kinds, the simulation hypothesis, Boltzmann brains and so on. Of course, it is entirely possible that much of this reasoning can be rewritten in such a way as to explicitly invoke SSL credences rather than PSL credences - in this article I have not attempted to determine whether or not this can be done. But even if such rewriting is possible, simply showing explicitly how to achieve it would surely in and of itself represent a major step forward in our understanding of the epistemology of such scenarios. Thus distinguishing clearly between PSL and SSL credences in these applications may help demarcate scientific and unscientific applications of the notion of self-location, which has the potential to significantly clarify ongoing discussions on these topics.

8 Acknowledgements

Thanks to Kelvin McQueen for discussions which inspired the writing of this paper.

References

[1] Yann Benétreau-Dupin. *Probabilistic Reasoning in Cosmology*. PhD thesis, The University of Western Ontario, 2015.

[2] Charles T Sebens and Sean M Carroll. Self-locating Uncertainty and the Origin of Probability in Everettian Quantum Mechanics. *The British Journal for the Philosophy of Science*, 69(1):25–74, 07 2016.

[3] Kelvin J. McQueen and Lev Vaidman. In defence of the self-location uncertainty account of probability in the many-worlds interpretation, February 2018.

[4] Don Fallis and Peter J. Lewis. Simulation and self-location. *Synthese*, 202(6):1–13, 2023.

[5] Michael G. Titelbaum. The relevance of self-locating beliefs. *The Philosophical Review*, 117(4):555–605, 2008.

[6] Eddy Keming Chen. Time's arrow and self-locating probability. *Philosophy and Phenomenological Research*, 105(3):533–563, 2021.

[7] David Builes. Center indifference and skepticism. *Noûs*, forthcoming.

[8] Nick Bostrom. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, New York, 2002.

[9] Adam Elga. Self-locating belief and the sleeping beauty problem. *Analysis*, 60(2):143–147, 2000.

[10] Adam Elga. Defeating dr. evil with self-locating belief. *Philosophy and Phenomenological Research*, 69(2):383–396, 2004.

[11] Bradley Monton. Sleeping beauty and the forgetful bayesian. *Analysis*, 62(1):47–53, 2002.

[12] Ruth Weintraub. Sleeping beauty: A simple solution. *Analysis*, 64(1):8–10, 2004.

[13] David Lewis. Attitudes de dicto and de se. *Philosophical Review*, 88(4):513–543, 1979.

[14] Christopher J. G. Meacham. Sleeping beauty and the dynamics of de se beliefs. *Philosophical Studies*, 138(2):245–269, 2008.

[15] Darren Bradley. Reasons for belief in context. *Episteme*, pages 1–16, forthcoming.

[16] Christopher J. G. Meacham and Jonathan Weisberg. Representation theorems and the foundations of decision theory. *Australasian Journal of Philosophy*, 89(4):641–663, 2011.

[17] JOHN LESLIE. *The Anthropic Principle Today*, pages 163–187. Catholic University of America Press, 1997.

[18] David Albert. Probability in the everett picture. In Simon Saunders, Jonathan Barrett, Adrian Kent, and David Wallace, editors, *Many Worlds?: Everett, Quantum Theory & Reality*. Oxford University Press, 2010.

[19] Sean M. Carroll. Why boltzmann brains are bad, 2017.

[20] Mark Srednicki and James Hartle. Science in a very large universe. *Physical Review D*, 81(12), June 2010.

[21] Jonah Schupbach. *Bayesianism and Scientific Reasoning*. Elements in the Philosophy of Science. Cambridge University Press, 2022.

[22] H. Greaves. Understanding Deutsch's probability in a deterministic multiverse. *eprint arXiv:quant-ph/0312136*, December 2003.

[23] Stuart Armstrong. Anthropic decision theory for self-locating beliefs. 2017.

[24] David Builes. Time-slice rationality and self-locating belief. *Philosophical Studies*, 177(10):3033–3049, 2020.

[25] Rachael Briggs. Putting a value on beauty. *In Tamar Szabo Gendler and John Hawthorne (Eds.), Oxford Studies in Epistemology, Volume 3. Oxford University Press*, pages 3–34, 2010.

[26] Huw Price. Decisions, decisions, decisions: Can savage salvage everettian probability?, 2008.

[27] Brian Hedden. Time-slice rationality. *Mind*, 124(494):449–491, 2015.

[28] James M. Joyce. A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65(4):575–603, 1998.

[29] Benjamin Eva. Principles of indifference. *Journal of Philosophy*, 116(7):390–411, 2019.

[30] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[31] Nicholas Shackel. Bertrand?s paradox and the principle of indifference. *Philosophy of Science*, 74(2):150–175, 2007.

[32] Bas Van Fraassen. Explanation through representation, and its limits. *Epistemologia*, 1:30–46, 2012.

[33] Edwin T. Jaynes. The well-posed problem. *Foundations of Physics*, 3:477–492, 1973.

[34] Bas C Van Fraassen. Laws and symmetry. 1989.

[35] Nick Bostrom. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, 2002.

[36] Wayne C. Myrvold. Probabilities in statistical mechanics: Subjective, objective, or a bit of both? 2011.

[37] Wayne C. Myrvold. *Beyond Chance and Credence: A Theory of Hybrid Probabilities*. Oxford University Press, 02 2021.

[38] Feraz Azhar. Testing typicality in multiverse cosmology. 2015.

[39] Darren Bradley. *Bayesianism and Self-Locating Beliefs*. PhD thesis, Stanford University, 2007.

[40] Nick Bostrom. Sleeping beauty and self-location: A hybrid model. *Synthese*, 157(1):59–78, 2007.

[41] Emily Adlam. The Problem of Confirmation in the Everett Interpretation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 47:21 – 32, 2014.

[42] Joseph Y. Halpern and Mark R. Tuttle. Knowledge, probability, and adversaries. *J. ACM*, 40(4):917–960, September 1993.

[43] Darren Bradley. Self-location is no problem for conditionalization. *Synthese*, 182(3):393–411, 2011.

[44] S. Friederich. *Multiverse Theories: A Philosophical Perspective*. Cambridge University Press, 2021.

[45] Feraz Azhar. Prediction and typicality in multiverse cosmology. 2013.

[46] Andrei Linde. Sinks in the landscape, boltzmann brains and the cosmological constant problem. *Journal of Cosmology and Astroparticle Physics*, 2007(01):022–022, January 2007.

[47] Raphael Bousso, Roni Harnik, Graham D. Kribs, and Gilad Perez. Predicting the cosmological constant from the causal entropic principle. *Physical Review D*, 76(4), August 2007.

[48] James Woodward and Lauren Ross. Scientific Explanation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

[49] Phil Dowe. *Physical Causation*. Cambridge University Press, New York, 2000.

[50] Philip Kitcher. Explanatory unification and the causal structure of the world. In Philip Kitcher and Wesley Salmon, editors, *Scientific Explanation*, pages 410–505. Minneapolis: University of Minnesota Press, 1989.

[51] B. Carter. In M.S. Longair, editor, *Confrontation of Cosmological Theories with Observational Data*, International Astronomical Union Symposia. Springer Netherlands, 2013.

[52] Erik Olsson. Coherentist Theories of Epistemic Justification. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

[53] Wilfrid Sellars. *Empiricism and the Philosophy of Mind*. Harvard University Press, Cambridge, Mass., 1997.

[54] Laurence Bonjour. Can empirical knowledge have a foundation? *American Philosophical Quarterly*, 15(1):1–13, 1978.

[55] Donald Davidson. A coherence theory of truth and knowledge. In Ernest LePore, editor, *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson*, pages 307–319. Blackwell, 1986.

[56] Hasok Chang. Scientific progress: Beyond foundationalism and coherentism. *Royal Institute of Philosophy Supplements*, 61:1–20, 2007.

[57] Hasok Chang. *Inventing Temperature: Measurement and Scientific Progress*. OUP Usa, New York, US, 2004.