Self-locating Priors and Cosmological Measures

Cian Dorr*and Frank Arntzenius†

Penultimate version: 21st January 2016‡

1 Introduction

It seems like bad news for a theory if it entails that almost all of those who perform a certain experiment get a certain result, and we actually perform that experiment and get a different result. But it is not immediately obvious why this should be bad news, or what kind of bad news it is, when that the theory in question is logically consistent with the fact that we got the result we actually got. This is something we need to understand better. It is not enough just to say that other things being equal, a theory's having this feature is a reason not to believe it, since other things never are equal. In all the interesting cases, the theories in question will have a great many other features which are, ceteris paribus, reasons to believe them—they may be attractively strong and simple; they may accurately predict the values of certain measured parameters; and so forth. We need a framework of thinking about the bearing of evidence on theories that can give us some guidance about how these factors trade off against one another.

Indeed we can see that it is not always bad news for a theory when we make observations that are atypical according to it. For consider the following theory (if you want to call it that): we will perform experiment E and get result A, although almost

*New York University
†University College, Oxford
‡To appear in *The Philosophy of Cosmology*, ed. Khalil Chamcham, John Barrow, Simon Saunders, and Joe Silk. Cambridge University Press: 2016.

everyone else who performs experiment E will get result B. Since our getting any result other than A would refute this theory, this result looks like good news rather than bad news. So, this is another place where we need some guidance.

The need for such a framework is especially pressing when we turns our attention from the elaborate experiments physicists are paid to perform to the "experiments" we perform all the time whether we like it or not—for example, the experiment of standing in front of a mirror and seeing what you look like. There might be gazillions of non-human observers in the universe, most of whom see completely different things when they look into a mirror. It seems foolish to reject a serious theory that posits a multitude of disparately-shaped aliens just on the grounds that you see a human-like form (two arms, two legs, ...) when you look in a mirror. Surely the place to look if you want to investigate a theory like that is a telescope, not a mirror! But it is unclear how this piece of common sense is to be reconciled with the idea that it is bad when a theory says that our observations are atypical. Small wonder that so many practicing physicists are suspicious of considerations having to do with the typicality of our observations—"anthropic" considerations—and would prefer to be able to make comparisons between theories without ever having to think about such matters.

Unfortunately for them, it is hard to see how one could possibly avoid the need to take such considerations into account. Since the work of Boltzmann—if not that of Democritus—physics has thrown up a succession of theories which seem to have the problematic feature that they make our actual observations excessively atypical, but are simple and attractive in other respects, and are logically consistent with our evidence, so that it is not clear what can be said against them without appealing to considerations of atypicality. In Boltzmann's picture, a fixed finite stock of particles continue to exist and to interact eternally, with the result that every possible dynamical state of those particles (with a fixed total energy) eventually comes arbitrary close to being realised, including all those possible dynamical states that subserve the existence of observers making observations of any humanly possible kind. Nothing that we have observed is inconsistent with this theory. Its only obvious defect is that according to it,

the vast majority of observations are utterly unlike our actual observations. They are, rather, the kinds of observations one would expect to be made by "Boltzmann Brains" (Albrecht and Sorbo 2004)—short-lived, isolated observers who came into existence as part of a recent, localised fluctuation from equilibrium. If one denies that this is any kind of problem for Boltzmann's theory, one seems forced into the position that observations could never bear in any way on a theory that entail that every possible observation is made at least once at some point in the history of the universe. But this conclusion would be a disaster, since cosmologists have considered a multitude of serious hypotheses with exactly this feature. If empirical investigation can never favour some of these hypotheses over others, we seem to be doomed to a paralysing level of scepticism.

These examples also remind us—if it wasn't clear enough already—that we need to think seriously about what it even means to say that our observations are 'atypical'. For given Boltzmann's theory, every possible observation is made not just once, but infinitely many times. The cardinalities are the same: there is a countable infinity of observations just like ours, and a countable infinity of observations unlike ours. So in what sense is it true to say that 'most' or 'almost all' observations are unlike ours? One could try to make sense of the 'most' claim by taking some kind of limit using a sequence of longer and longer finite temporal intervals. But what could make this the right way to compare the infinite sets? And how are we supposed to generalise it to more recent infinite-population theories which are set in relativistic spacetimes whose extent may be infinite in both temporal and spatial directions?

Quite apart from problems associated with infinity, the typicality or otherwise of our observations clearly depends on the class of objects you are considering: a feature that is typical among primates need not be typical among all animals. But what is the relevant class of things when we are trying to figure out whether our observations are, or are not, problematically atypical according to a certain theory?

In what follows we will develop a Bayesian framework for answering all these questions.

2 Bayesian background

We will assume that an ideally rational person at any time t has degrees of belief ('credences') which can be represented by a probability function $\mathbf{C}(\cdot|E_t)$: the result of conditionalising a certain other probability function $\mathbf{C}$—her 'priors'—on the total evidence E_t that she has at t.

In this paper, "probability functions" will always be functions P from *propositions*—things that can be true or false, and believed to various degrees—to real numbers in the unit interval [0,1], such that

(i) $P(A) = 1$ whenever A is logically true

(ii) $P(A) \leq P(B)$ whenever B is a logical consequence of A; and

(iii) $P(A \vee B) = P(A) + P(B)$ whenever A and B are logically inconsistent.[1]

We take the set of propositions on which rational peoples' credence functions are defined to include not only eternal, qualitative propositions (like *There exists or will exist or has existed at least one physicist*) as well as self-locating propositions (like *There is a physicist in front of me*) which attribute a qualitative property to the particular agent at the particular time in question.[2]

Note that we are merely assuming that rational people can be so represented, not that they need to have priors in their heads in any psychologically realistic sense, let alone that they need to have had them in their heads temporally prior to any given episode of rational belief-formation. We are also not assuming any particular account of evidence, e.g. that only propositions about one's conscious experience at t can be

[1] We do not require countable additivity, the analogue of (iii) for countably infinite disjunctions.

[2] Our purposes in this paper will not require taking any particular view as to the nature of propositions in general, or of self-locating propositions in particular. Perhaps the qualitative and the self-locating propositions are just two special subclasses of the class of all propositions, the former being those that are not "directly about" any particular objects at all and the latter being those that are "directly about" the person and time in question but nothing else. Or perhaps self-locating propositions should be treated as *sui generis*, as in the influential approach of Lewis 1979.

part of one's evidence at t. Everything we say will be compatible with externalistic views on which a rich body of truths about one's surroundings and history count as part of one's evidence (e.g. Williamson 2000). While the difference between these conceptions of evidence can make a big difference in some cases, we don't think it will matter to any of the theoretical comparisons we will be concerned with in this paper.

One advantage of characterising a rational person's credences as the result of conditionalising her priors on her total evidence at the relevant time, rather than as the result of conditionalising her previous credence function on the evidence that she just received, is that it determines, in a prima facie plausible way, how a rational person's credences evolve when she forgets things, and how her credences evolve in response to changing self-locating evidence.[3] Most importantly, it gives us a setting in which we can pose questions not just about how *new* evidence should modify our credences, but also about how the evidence that we already have bears on a given theory. It is vital to be able to make sense of this, since in many cases it is quite an achievement to extract any observational predictions at all from a theory, and often the predictions we manage to extract will concern observations that we have already made. Such "old evidence", whose relevance we want to be able to discuss, includes not just facts about experiments that have been done by physicists, but familiar facts of everyday life that are well-known to everyone. In the present framework, even when a person has already updated her credences on certain evidence E, we can still say that E confirms H "by that person's lights" if her prior $\mathbf{C}$ is such that $\mathbf{C}(H|E) > \mathbf{C}(H)$, and that E confirms H *simpliciter* if $\mathbf{C}(H|E) > \mathbf{C}(H)$ for any reasonable $\mathbf{C}$.

We stated above that the credences of a rational person are defined on self-locating as well as qualitative propositions. In recent years the epistemology of self-locating belief has been the focus of a substantial body of literature (for a survey, see Titelbaum 2013). That literature tends to focus on thought-experiments involving perfect duplication of experience between different people, or between the same person at different times. This might suggest that the inclusion of self-locating propositions in the framework

[3]See Arntzenius 2003 for further discussion of these issues.

is a technical innovation driven primarily by such thought experiments. But the idea that there is a crucial difference between learning that *you now* have a certain property and merely learning that *someone*, *sometime* has that property is really a completely intuitive one, which can be illustrated by any number of everyday cases. For instance, suppose that you and 20 friends have booked all the rooms in a 21-room hotel. You remember either reading in the hotel brochure that all but one room in the hotel is red, or reading that only one room is red. You just don't remember which, and you are initially about 50-50 as to the type of hotel that you have booked. Upon arrival you and your friends randomly pick rooms to go to. You then find that your room is red. If you took your evidence to be the qualitative proposition *Someone is in a red room* you would have no reason to update your credences regarding the type of hotel that you are in, since that proposition had to be true either way. But if your evidence is the self-locating proposition *I am in a red room*, it strongly supports the hypothesis that all but one room in your hotel is red. And it seems obvious that this is the correct way to reason. (This is presumably what you would conclude if you believed that you were the only person in the hotel, and it surely makes a relevant difference whether you believe that you have 20 friends with you or not.)

Of course, we have claimed that one should form credences by conditionalising one's priors on one's *total* evidence at any given time, and it is not realistic to think that *I am in a red room* is your total evidence. However it is not unrealistic to think that this is the only part of your evidence that we need to take into account in order to assess how your evidence bears on the comparison between the two live theories about what kind of hotel you are in. Formally, when E^- is a consequence of your total evidence E, E^- will exhaust the bearing of E on two hypotheses H_1 and H_2 whenever $\mathbf{C}(E|H_1E^-) = \mathbf{C}(E|H_2E^-)$. For in that case,

$$\frac{\mathbf{C}(H_1|E)}{\mathbf{C}(H_2|E)} = \frac{\mathbf{C}(E \wedge H_1)}{\mathbf{C}(E \wedge H_2)} = \frac{\mathbf{C}(E|E^- \wedge H_1)\mathbf{C}(E^- \wedge H_1)}{\mathbf{C}(E|E^- \wedge H_2)\mathbf{C}(E^- \wedge H_2)} = \frac{\mathbf{C}(E^- \wedge H_1)}{\mathbf{C}(E^- \wedge H_2)} = \frac{\mathbf{C}(H_1|E^-)}{\mathbf{C}(H_2|E^-)}$$

so that E^- can serve as a proxy for your total evidence E when assessing its bearing on these two hypotheses. In the present case, it is plausible that by the lights of your

priors, your total evidence in all its detail is just as likely on the assumption that you are in the only red room in the hotel as it is on the assumption that you are in one of twenty red rooms in the hotel, so that *My room is red* can serve as a proxy for your total evidence.

In this example, the hypotheses *All but one room in my hotel is red* is not a qualitative proposition, and it would be crazy in any case to think that your total qualitative evidence can serve as a proxy for your total evidence when the relevant hypotheses are self-locating propositions. But we could instead have focused on the qualitative hypothesis *There exists a 21-room hotel with 20 red rooms*. Given what you remember about the booking, you should clearly become more confident in this when you find yourself in a red room.

It is clear that your total evidence is not a qualitative proposition. We need not assume that it is a self-locating proposition either (i.e. one that attributes a qualitative property to the agent): perhaps we should instead take it to be a "de re" proposition like *I am in this particular red room*, which attributes to the agent a non-qualitative property involving a particular object. We will however be assuming that for the purposes of reasoning about qualitative and self-locating hypotheses, one's total self-locating evidence can serve as a proxy (in the sense explained above) for one's total evidence. Having absorbed the lesson the lesson that conditionalising on an existential generalisation often has very different effects from conditionalising on an instance of that generalisation, this assumption might seem implausible given the *de re* view of evidence—why would *I am in a red room* be any better as a proxy for *I am in this particular red room* than *Someone is in a red room*? But when we bear in mind that one's self-locating evidence might include propositions like *I am in a room that looks this highly distinctive way*, or *I am in a room that I remember having been in twenty years ago*, it becomes hard to imagine a plausible view on which the mere identity of the particular objects in one's environment would have any further capacity to discriminate among qualitative or self-locating hypotheses.

Some theorists have taken seriously the "relevance-limiting thesis" according to

which only qualitative evidence needs to be taken into account when we are reasoning about qualitative hypotheses.[4] Their idea for dealing with apparent counterexamples like our hotel case is to say that our total qualitative evidence is quite rich—not just *Someone is in a red room*, but *Someone is in a red room experiencing such-and-such detailed pattern of light and shade, hearing such-and-such sounds, having such-and-such memories,. . . .*. When we are reasoning about hypotheses according to which the population of the universe is small enough that it is vanishingly unlikely that there would be more than one person satisfying such a rich description, this lets us recover reasonable-looking patterns of reasoning described above, at the cost of having to count all sorts of intuitively irrelevant aspects of one's evidence as crucially relevant. For example, it is plausible that a reasonable prior credence function will count it as approximately twenty times more likely that someone will satisfy the above rich description conditional on there being twenty people in red rooms than conditional on there being only one person in a red room. But as soon as we start thinking about scenarios in which we can no longer reasonably neglect the possibility that there is more than one witness to the existential quantification that is our total qualitative evidence, the approach that looks only at qualitative evidence will start to generate distinctive and implausible results. For example, consider the following case:

> *Measuring a Parameter:* Two qualitative theories T1 and T2 both entail that the population of the universe is vast but finite. They differ with regard to the value of a certain cosmological parameter α. T1 says that the true value of α is 34.31, and T2 says that it is 34.59. Because of this, T1 and T2 also differ as regards the distribution of results among the many repetitions in the history

[4]The label is due to Titelbaum (2013); defenders include Halpern and Tuttle 1993, Halpern 2004, Meacham 2008, and Neal 2006 (approvingly cited in Carroll 2010, p. 401). In a somewhat similar vein, Hartle and Srednicki (2007, p. 1) claim that "Cosmological models that predict that at least one instance of our data exists (with probability one) somewhere in spacetime are indistinguishable no matter how many other exact copies of these data exist", although their later work (Srednicki and Hartle 2010, 2013) suggests that their view may be a "permissivist" one on which ideal reasonableness permits, but does not require, the disposition to reason in such a way that one's credences in such models never evolve under the impact of new evidence.

of the universe of a certain experiment E which fairly reliability measures the value of α. Conditional on T1, the expected proportion of who get the result 34.31 among those who do E is approximately 1/20, while conditional on T2, the expected proportion is approximately 1/1000.

Intuitively, doing E and getting 34.31 very strongly favours T1 over T2. But if the populations are sufficiently large, our *qualitative* evidence will deserve high prior credence conditional on both T1 and T2, even if we are careful to include all manner of apparently irrelevant background details. Thus the view that only qualitative evidence matters in reasoning about qualitative hypotheses leads to a disastrous scepticism about our ability to bring empirical evidence to bear in distinguishing different large-population hypotheses.[5]

As if this weren't bad enough, the relevance-limiting thesis faces the further problem that it requires a counterintuitive boost in the posterior probabilities of theories according to which the population is large relative to their prior probabilities, simply because the more people there are, the less unlikely it will be (by the lights of a reasonable prior) that any given very detailed qualitative property has at least one instance. When combined with the approach's inability to allow for evidence-based discrimination between these large-universe hypotheses, this threatens to lead to a truly paralysing sceptical collapse.

One might still try to make a last gasp effort to make do only with purely qualitative evidence by adopting a ultra-fine-grained conception of evidence, on which your total qualitative evidence is the existential generalisation of a property so specific that you can legitimately neglect the possibility that more than one person has it, no matter how many people there are. For example, the property might specify the exact values

[5]Neal (2006) attempts to address this problem by (i) adopting a very fine-grained conception of evidence, on which the population of the universe would have to be quite large (he suggests something in the order of $10^{10^{10}}$) for there to be a substantial chance that the qualitative property attributed by our total evidence has multiple instances; and (ii) proposing that we should simply ignore the possibility that the population is this large. This "ignoring" strikes us as patently unreasonable, in spite of the dubious verificationist and ethical considerations which Neal offers in its favour. (On the ethical ramifications of infinite populations, see Arntzenius 2014.)

of certain continuous parameters (perhaps having to do with one's mental state), in which case it is plausible that reasonable priors will assign its existential generalisation probability zero, even conditional on number of people being (countably) infinite. Implementing this strategy would require an elaboration of the framework in which conditional priors are taken as primitive rather than defined by $\mathbf{C}(A|B) = \mathbf{C}(A \wedge B)/\mathbf{C}(B)$, so that one can meaningfully conditionalise even on propositions whose prior credence is zero (see Hájek 2003). The question how one's unconditional priors should be extended to allow for conditionalisation on probability-zero propositions raises tricky issues (see Dorr 2010, Myrvold 2015). Indeed, the project of formulating rules for extending unconditional priors to conditional priors involves puzzles that are in many ways analogous to those that arise for the project of formulating rules for extending qualitative priors to self-locating priors. Given that the ultra-fine-grained conception of evidence has severe foundational problems—intuitively, it vastly overstates the extent to which beings like us could ever hope to get their beliefs to correlate with the exact value of any continuous parameter—we will set it aside, while noting that the ideas about self-locating priors which we will discuss in this paper will have analogues within the fine-grained framework.

One final note: the Bayesian framework has often been combined with the idea that there are no rational constraints on priors. This would make the present enquiry trivial. We will be taking it for granted that there are better and worse priors to have, and that factors such as simplicity can legitimately be appealed to in saying what makes the better ones better. This means that we accept that in certain senses of 'simple', you should be pretty confident that the world is 'simple' in the absence of relevant evidence. Some will find this suspiciously rationalistic. They will be tempted to think that reasonable priors should be 'unbiased' or 'uniform'. But making sense of a notion of lack of bias or uniformity is extremely difficult (especially in the infinite case). And in the limited range of cases where one can make sense of it (e.g. by restricting to a finite range of possibilities or a uniquely natural measure), such 'uniform' or 'unbiased' priors turn out to have the feature that you typically don't learn anything about the

world given any finite amount of evidence. In any case, legislating some notion of uniformity or lack of bias seems equally a prioristic to us.

3 A principle about finite worlds

In our present state of understanding, it would be foolish to try to codify the features that make some priors more reasonable than others—simplicity and so forth—in the form of some precise collection of axioms. In general, we just have to muddle along as best we can by trusting our judgments about particular cases. However, there are a few special domains where we have the resources to formulate principles about what reasonable prior credence functions are like which go beyond the probability axioms, are not obviously false, and are precise enough to be worth arguing about. One of these domains is the epistemology of self-locating belief conditional on there being only finitely many observers. In this domain, we can formulate a general principle which, if true, will allow one to completely specify an reasonable assignment of prior credences to self-locating propositions (conditional on the population being finite), given as input a reasonable assignment of prior credences to qualitative propositions. This principle can be seen as a very limited principle of indifference: the intuition is that your self-locating priors, conditional on a certain qualitative state of affairs, should be indifferent among all the observers who exist in that state of affairs.[6] Or better—since self-locating propositions may address the question *when it is* as well as *who you are*—your conditional self-locating priors should be indifferent among all the portions of the lives of observers in the relevant state of affairs whose duration is the same.

To state this principle more rigorously, we will need to introduce some notation. Where P is a probability function and R is a real-valued random variable—which we

[6]This basic thought is what Bostrom (2002) calls 'The Self-Sampling Assumption': 'One should reason as if one were a random sample from the set of all observers in one's reference class'. PROPORTION below is intended as a precisification of this vague formulation. Our principle does not talk about 'reference classes': where Bostrom would ask 'What is the right way of specifying the reference class?', we will simply ask 'What is the right way of defining "observer"?'

can identify with a function that maps each real number x to a proposition "$R = x$" in such a way that the resulting propositions are pairwise inconsistent and jointly exhaustive—we will use '$\hat{P}(R)$' to denote the expectation value of R in P. Similarly we write '$\hat{P}(R|H)$' for the expectation value of R in $P(\cdot|H)$. When F and G are any properties, we write $\langle F : G \rangle$ for the random variable whose value is the ratio between the total, for all the things that are ever F, of the duration for which they are F and the total, for all the things that are ever G, of the duration for which they are G. For example, $\langle \text{physicist} : \text{philosopher} \rangle = 10$ is the proposition that the total duration of all philosophers' careers is positive and ten times smaller than the total duration across history of all physicists' careers.[7] Then

PROPORTION Where $\mathbf{C}$ is any reasonable prior credence function, H is any qualitative hypothesis which entails that the total duration of the lives of all observers is positive and finite, and F is any qualitative property:

$$\mathbf{C}(I \textit{ am } F|H) = \hat{\mathbf{C}}(\langle F \text{ observer} : \text{observer} \rangle|H)$$

In words: your prior probability that you are F given H should equal your prior expectation, given H, for the proportion of all observer-time taken up by F. Note that the expectation value of this random variable depends only on how $\mathbf{C}$ treats qualitative propositions. Thus PROPORTION fully determines one's self-locating priors (conditional on the total duration of the lives of all observers being positive and finite) as a function of one's qualitative priors.

To get a sense for the appeal of PROPORTION, let's return to the case of *Measuring a Parameter*. We can take it that T1 and T2 do not differ as regards the proportion of observers who ever do experiment E, or as regards how far along they are in their lives (on average) when they do it. In that case, the expected ratio between the total amounts of observer-time occupied by the properties *having done E and got 34.31* and

[7] We leave it open what the relevant notion of 'duration' is: it might be taken to be the physical notion of proper time; subjective (psychological) time; some measure of complexity of evolution in the relevant system's physical state; or something else again.

having done E will be equal to the expected proportion of trials of E that yield the result 34.31 according to that theory. So according to PROPORTION, **C**(*I did E and got 34.31*|T1) will be (approximately) twenty times greater than **C**(*I did E and got 34.31*|T2) for any reasonable **C**. Thus, assuming that none of the other details of your total evidence beyond the fact that you did *E* and got 34.31 are relevant to the comparison between T1 and T2, we can draw a conclusion about how the experiment should affect your credences: the ratio of your credence in T1 to your credence in T2 should be 20 times greater than it would have been with no relevant evidence.

Ideas in the vicinity of PROPORTION are often summed up with slogans like 'We should expect ourselves to be typical'. But such slogans need to be treated with care. If being a typical observer means having only those properties that most observers have, we should obviously expect *not* to be typical; indeed we should be confident that everyone has some properties that most observers lack. For the same reason, if being a typical observer means something like being an *average* observer, we should also expect not to be typical. To extract from PROPORTION the claim that we should be confident that we are typical observers, we need to devise an interpretation for 'typical observer' on which it is trivially true that if there are finitely many observers, most of them are typical. If we were concerned only with typicality in one particular quantitative *respect*, we could simply define 'typical' to mean 'having a value for the relevant quantity that is within so many standard deviations of the mean value among all observers'. But making sense of an "all things considered" notion of typicality is a much harder task, and not one that we have any need to take on.

Let us turn next to a more controversial application of PROPORTION.

> *Brains or No Brains?:* Two theories *Brains* and *No Brains* are generally similar except that *Brains* predicts that, after the heat death of the universe, an enormous (but still finite) number of observers come into existence because of random vacuum fluctuations, while *No Brains* includes some mechanism that prevents this from happening. Most of the randomly-produced observers that exist according to *Brains* are "Boltzmann Brains"—i.e. things that just barely

> qualify as "observers" in the relevant sense, however we end up cashing it out. Although almost all the Boltzmann Brains are short-lived, *Brains* predicts that there are so many of them that the total duration of their lives is much larger than the total duration of all the ordinary observers' lives. And while a few of the Boltzmann Brains will, by chance, have misleading experiences as of living in a world like ours, fake memories, etc., almost all of them will spend their entire lives enduring the rather unpleasant sorts of experiences one would expect given the inhospitable conditions in which they have come into existence.

Consider our current evidence—evidence, perhaps, as of sitting in a comfortable room, drinking a cup of tea while typing on a computer keyboard. *Brains* and *No Brains* differ radically with regard to the proportion of observer-time occupied by this qualitative property. So according to PROPORTION, a reasonable prior credence function will assign this evidence much higher probability conditional on *No Brains* than conditional on *Brains*, so that the evidence counts heavily in favour of *No Brains*. Of course, this does not yet mean that we should be much more confident in *No Brains* than in *Brains*. This confidence depends on the priors for the two theories as well as on the evidence, and *No Brains* might have features in virtue of which it deserves a much lower prior credence—e.g. if the mechanism that prevents Boltzmann Brains from forming is an *ad hoc* postulate without any further motivation, it might detract greatly from *No Brains*'s simplicity. However, the larger the ratio of Boltzmann Brains to normal observers is according to *Brains*, the harder it will be to justify an asymmetry in the priors large enough to compensate for the force of the evidence.[8]

Brains or No Brains? shows that PROPORTION has some distinctive and controversial implications when combined with other plausible claims about reasonable prior credences. In that example, the required additional claim was to the effect that if theories

[8]Of course, we don't have to rest content with the evidence we can get by sitting in our armchairs: we can also go out and do some experiments. However, unless these experiments have an incredibly strong ability to discriminate the theories, they won't change the epistemic situation very much.

are roughly similar in respect of simplicity, etc., they should not be assigned very different prior credences. In other examples, the additional claim that combines with PROPORTION to generate controversial implications is one that is often taken for granted in applications of Bayesian methods: namely, that conditional on the hypothesis that a certain function is the one that maps each proposition to its *objective chance* of being true—its physical probability—our prior credences should agree with that function.

(PP) $$\mathbf{C}(A|P \textit{ is the objective chance function}) = P(A)$$

(This is one version of the 'Principal Principle' from Lewis 1980.) The combination of (PP) with PROPORTION makes for distinctive consequences when we are dealing with theories according to which are significant objective chances for substantially different total numbers of observers. For example, consider the following case from Bostrom 2002:

> *The Incubator:* Stage (a): The world consists of a dungeon with one hundred cells. The outside of each cell has a unique number painted on it (which cannot be seen from the inside); the numbers being the integers between 1 and 100. The world also contains a mechanism which we can term the incubator. The incubator first creates one observer in cell #1. It then. . . flips a fair coin. If the coin falls tails, the incubator does nothing more. If the coin falls heads, the incubator creates one observer in each of the cells ##2–100. Apart from this, the world is empty. It is now a time well after the coin has been tossed and any resulting observers have been created. Everyone knows all the above.
>
> Stage (b): A little later, you have just stepped out of your cell and discovered that it is #1. (Bostrom 2002, p. 64)

Let S be the qualitative description of this setup, supplemented with the stipulation that all observers created by the incubator live equally long lives; let H and T respectively be the conjunctions of S with the propositions that the coin lands Heads and that the coin lands Tails. Since S specifies that the chances of H and T are equal, (PP) says that $\mathbf{C}(H|S) = \mathbf{C}(T|S) = 1/2$ for any reasonable $\mathbf{C}$. Since H entails that

there are exactly 100 equally-long lived observers of whom one is in a cell numbered #1, while T entails that all observer-time is spent in a cell numbered #1, we have $\hat{\mathbf{C}}(\langle\text{observer in cell \#1 : observer}\rangle|H) = 1/100$ and $\hat{\mathbf{C}}(\langle\text{observer in cell \#1 : observer}\rangle|T) = 1$. So by PROPORTION, $\mathbf{C}(\textit{I am in cell \#1}|\mathrm{H}) = 1/100$ while $\mathbf{C}(\textit{I am in cell \#1}|\mathrm{T}) = 1$. We can thus apply Bayes's theorem to the probability function $\mathbf{C}(\cdot|S)$ to get

$$\begin{aligned}\mathbf{C}(H|\textit{I am in cell \#1} \wedge S) &= \frac{\mathbf{C}(\textit{I am in cell \#1}|H)\mathbf{C}(H|S)}{\mathbf{C}(\textit{I am in cell \#1}|S)} \\ &= \frac{\mathbf{C}(\textit{I am in cell \#1}|H)\mathbf{C}(H|S)}{\mathbf{C}(\textit{I am in cell \#1}|H)\mathbf{C}(H|S) + \mathbf{C}(\textit{I am in cell \#1}|T)\mathbf{C}(T|S)} \\ &= \frac{0.01 \times 0.5}{0.01 \times 0.5 + 1 \times 0.5} = \frac{1}{101}\end{aligned}$$

So if we assume that at stage (a) you have no relevant evidence beyond S, and that you gain no relevant evidence at stage (b) beyond the proposition that you are in cell 1, your credence in Heads will decrease from 1/2 at stage (a) to very low at stage (b).

In thinking about cases like *The Incubator*, some have been attracted to an alternative view according to which your credence in Heads should be 1/2 at stage (b). On the less plausible version of this view, your credence should *also* be 1/2 at stage (a). But this is hard to take seriously, since the discovery that you are in cell #1 looks like strong evidence in favour of Tails (which entails it).[9] On the more plausible version of the view, your credence in Heads should be high at stage (a), so that it can be 1/2 even after the impact of this strong evidence.

Given our Bayesian framework, there are two ways to generate this high credence in Heads at stage (a): we could either revise PROPORTION in such a way that your evidence at stage (a) will count as heavily favouring Heads, or we could revise (PP) in such a way as to require reasonable priors to favour Heads over Tails. One natural thought that would motivate the relevant sort of revision to PROPORTION is the idea that the more observers there are, the less surprising it is that *you* are one of them, so that the self-

[9]If we add the stipulation that all the observers have exactly the same evidence at stage (a), the claim that your credence should not change between stage (a) and stage (b) follows from the relevance-limitation thesis discussed in section 2. There are also some, such as Bostrom (2002), who are sympathetic to this claim but not to the relevance-limitation thesis.

locating proposition that you *exist* (or that you are an observer) should count as strong evidence for Heads (Bartha and Hitchcock 1999). This is inconsistent with PROPORTION, which entails that your prior credence that you are an observer conditional on the total duration of observer-time being positive and finite should be 1. If we wanted existence or observerhood to have evidential force, we could easily modify PROPORTION so as to concern not your prior unconditional credences, but your prior credences conditional on your existence or observerhood. The problem with this approach is that it is not really general enough. Deriving the judgment that your credence in Heads at stage (b) should be 1/2 in all variants of *The Incubator* that differ just with regard to the two population numbers associated with Heads and Tails would require a prior credence function in which the probability of *I am an observer* conditional on *There are n observers* increases *linearly* in proportion to n. And of course this is impossible, since conditional probabilities cannot exceed 1. The better option, then, is to keep PROPORTION while revising (PP), by building into the priors a proportional bias towards hypotheses according to which there are many observers, even when chances are equal.[10]

Those who favour a high credence in Heads at stage (a) in *The Incubator* will presumably take an analogous view about other comparisons between hypotheses that disagree about the total number of observers. For example, in *Brains or No Brains?*, they will hold that your prior credence in *Brains* conditional on *Either* Brains *or* No Brains *is true and I am an observer* should be very high, so that the posterior credences in *Brains* and *No Brains* given normal evidence (e.g. as of sitting drinking tea and typing) end up close. But considered as a general model for good reasoning about the population of the universe, this seems quite crazy. Consider our current state of ignorance as regards how hard it is for intelligent life to evolve in an arbitrary solar system. When combined with a cosmological theory according to which spacetime as a whole is finite (but very large), different answers to this question will generate radically different expected numbers of total observers. The "bias towards high populations" idea will thus lead us to the absurd result that we should right now be confident, conditional

[10]For further discussion of this strategy, including the details of the required modification of (PP), see Arntzenius and Dorr MS.

on the universe being finite, that it is *very easy* for intelligent life to evolve—probably easy enough that even when we conditionalise on the (by the lights of this approach surprising) fact that we have not yet encountered any alien life, we should still be confident that the average galaxy contains many inhabited solar systems. While this "abundant life" hypothesis is not itself unreasonable, it seems clear that there are also perfectly reasonable hypotheses on which life is far rarer than this, and that in our current state of ignorance, it would be quite unreasonable to assign a very low credence to these hypotheses.[11]

Some have argued that PROPORTION itself should be rejected on the grounds that it makes it easier than it should be to do astrobiology from the armchair. For example, Sean Carroll argues as follows against the claim that "we should make predictions by asking what most observers would see":

> Imagine we have two theories of the universe that are identical in every way, except that one predicts that an Earth-like planet orbiting the star Tau Ceti is home to a race of 10 trillion intelligent lizard beings, while the other theory predicts there are no intelligent beings of any kind in the Tau Ceti system. Most of us would say that we don't currently have enough information to decide between these two theories. But if we are truly typical observers in the universe, the first theory strongly predicts that we are more likely to be lizards on the planet orbiting Tau Ceti, not humans here on Earth, just because there are so many more lizards than humans. But that prediction is not right, so we have apparently ruled out the existence of that many observers without collecting any data at all about what's actually going on in the Tau Ceti system. (Carroll 2010, p. 225)

Carroll does not tell us what the two theories in his example say about life other than on Tau Ceti and Earth. This matters when we apply PROPORTION: if both theories entail that there are quadrillions of observers elsewhere, and say the same things about

[11]Our argument here echoes Bostrom's "Presumptuous Philosopher" argument against what he calls the 'Self-Indication Assumption' (Bostrom 2002).

how many of those observers are likely to have the qualitative property ascribed by our total self-locating evidence, the lizards will make no appreciable difference. But perhaps Carroll was taking it for granted that both theories entail that every observer is either on Earth or on Tau Ceti. In that case, PROPORTION does entail that our armchair evidence counts strongly against the lizard theory. If our priors are not biased towards high populations, and the theories really are on a par in the other relevant respects, we should be confident that the lizard theory is false, just as in *The Incubator* we should be confident that we are alone in the universe when we know that we are in cell #1.[12]

If, like Carroll, you think this result is wrong, it is worth trying to get clear on what it is about the lizards that is driving your intuition. Consider a range of theories.

T0 In any given solar system there is a certain tiny chance ε for life to evolve at all, but if observers do come into existence they are likely to be human-like creatures whose DNA uses two base pairs, having five toes on each foot.

T1 In any given solar system, the chance of human-like life evolving is ε, but the chance of lizard-like life evolving is $10,000\varepsilon$.

T2 In any given solar system, the chance of human-like creatures with five toes on each foot evolving is ε, while the chance of human-like creatures with six toes on each foot evolving is $10,000\varepsilon$.

T3 In any given solar system, the chance of human-like creatures whose DNA uses only two base pairs is ε, while the chance of human-like creatures whose DNA uses three or more base pairs is $10,000\varepsilon$.

Suppose that all the theories are on a par as regards simplicity, and agree that the number of solar systems is finite but far greater than $1/\varepsilon$.

[12]Hartle and Srednicki 2007 make a similar argument about aliens living in the atmosphere of Jupiter. They argue that it would be unreasonable to reject a theory according to which there are many such aliens 'solely because humans would not then be typical of intelligent beings in our solar system'. Of course we agree with this, but we note that the theories in their example are not described in enough detail—in particular, with regard to what they say about life outside the solar system—to determine what PROPORTION says about them.

According to PROPORTION, our actual evidence (as of being human-like, with five toes and two base pairs) strongly supports T0 over all of T1–T3. In the case of T3, this result is quite intuitive. Suppose we had known about T0 and T3 and their predictions before investigating our DNA: then, surely, the discovery that we have two base pairs would have strongly favoured T0. The situation seems similar in every relevant respect to *Measuring a Parameter*. Perhaps there is some temptation to think that the situation with T2 is different, given that we have always known how many toes we have. But how could that difference matter? The order in which we get evidence does not normally have much significance as regards what we should believe once we have the evidence; the mere fact that the toe-counting experiment was performed long ago is not in itself any kind of reason to discount its significance. This takes us back to T1 and the lizard beings. Carroll doesn't tell us very much about what is driving his judgment about them. Is it their scaly skin? Their cold blood? Their bizarre social structures? Their alien sensory experiences? We have the feeling that as the scenario is fleshed out in more detail, the sense that there could be any important difference between T1 and T2 will start to fade away.

The ways of fleshing out T1 that make it most plausible that we should think about it differently from T2 and T3 involve the lizard beings having a mental life that is in some deep respect very different from ours. But in these versions of the case it is no longer obvious what PROPORTION says, because it is no longer clear whether the lizards count as "observers". So far we have been treating PROPORTION as a single univocal principle; but we should now admit that as we are conceiving it, it it is really schema that can be filled in in many different ways depending on how one interprets 'observer'. Some of the instances of the schema have crazily counterintuitive consequences. For example, if we plug in 'five-toed biped' for 'observer', we will get the absurd result that no amount of evidence should shake your confidence that you are a five-toed biped, conditional on there being a positive finite number of five-toed bipeds. If we understand 'observer' as 'living being or rock', we will end up with the absurd result that our actual evidence heavily favours hypotheses according to which the ratio of

rocks to living beings is low over hypotheses that according to which it is high. We can refine our understanding of how the schematic notion of an 'observer' should be understood by considering our intuitive judgments about such cases. But there will still, inevitably, be hard cases where the intuitions are unclear.

Some friends of PROPORTION might hope to draw some principled, sharp line between observers and non-observers.[13] We don't think this can be done, but we also don't think that this is a problem for the basic thought underlying PROPORTION. Given that all the relevant factors are continuous, we should probably allow reasonable prior credence functions that blur the line in one way or another—e.g. by assigning real-valued "degrees of observerhood" which one integrates over time, instead of simply looking at the duration for which a single property is instantiated, and/or by taking a weighted average of many different probability functions each of which obeys PROPORTION on a different conception of observerhood. Probably, too, we should be somewhat permissive, allowing different reasonable prior credence functions corresponding to different ways of cashing out observerhood. Perhaps, if Carroll's lizards are sufficiently mentally alien, they may be among the things that can reasonably be excluded altogether, or assigned an observerhood score much less than that of humans, or excluded by some of the probability functions that enter into the final weighted average.

The general dialectical situation here is quite similar to the situation we are in in connection with the idea that reasonable prior credence functions favour simpler theories over more complex ones: there are many different ways of making the notion of simplicity precise, and we doubt that there is any uniquely natural, rationally compulsory way of measuring simplicity and taking it into account. Things are messy, but this shouldn't stop us from trusting our intuitive judgments about particular cases where it is relatively obvious how the simplicity comparisons pan out. The general methodology for working out how simplicity relates to reasonableness seems to be one

[13] A natural thought for those inclined towards some kind of mind-body dualism is that the sharp line in question is the one between consciousness and lack of consciousness (see, e.g. Page this volume).

of "reflective equilibrium"; insofar as we are drawn to something like PROPORTION, our attitude about what to count as an observer should be worked out in the same way.

Fortunately, the details about what counts as an observer for the purposes of PROPORTION do not seem to be relevant to any of the theoretical comparisons that have arisen in physics. For example, in realistic ways of filling in the details of *Brains or No Brains?*, it won't actually matter whether we count disembodied brains as observers when calculating the relevant proportions. We will get almost the same results if we only count fully embodied creatures, or creatures that live lives long enough to realise certain cognitive capacities, or creatures that achieve certain kinds of interaction with their environments. In each case, *Brains* still will entail (a) that the total duration of the lives of the post-heat-death "observers" is far greater than the total duration of the pre-heat-death "observers", and (b) that the qualitative property ascribed by our total evidence takes up a much lower proportion of post-heat-death observer-time than of pre-heat-death observer-time.

One further point: if we were only concerned with the question how we should respond to some change in our evidence, taking for granted that our credences *before* the change are reasonable, there would be no reason to concern ourselves with the question what to count as an observer. For given certain very weak assumptions, we can use PROPORTION to derive a formula which specifies our new credences conditional on some H in terms of our old credences conditional on H, in which the notion of observerhood does not appear at all except in delimiting the possible values of H:

PROPORTIONAL UPDATE When a reasonable person has evidence *I am E* and credence function $\mathbf{C}_t$ at t, and evidence *I am* E^+ and credence function $\mathbf{C}_{t^+}$ at t^+, and H is a qualitative proposition that entails that the total duration of observer-time is finite and that the total duration of E is positive if the total duration of E^+ is,

$$\mathbf{C}_{t^+}(\textit{I am } F|H) = \frac{\hat{\mathbf{C}}_t(\langle FE^+ : E\rangle|H)}{\hat{\mathbf{C}}_t(\langle E^+ : E\rangle|H)}$$

whenever the right-hand side is defined.

To derive PROPORTIONAL UPDATE from PROPORTION, the only assumption we need to

make about the property of observerhood is that it is entailed by both E and E^+.

Proof. Begin with a definition and two preliminary observations. *Definition:* when R and S are two random variables, let $R \times S$ be the random variable such that for any nonzero x, $R \times S = x$ is the disjunction of all conjunctions $R = y \wedge S = z$ where $yz = x$, while $R \times S = 0$ is just $R = 0 \vee S = 0$ (and thus can be true even when one of R and S is undefined). *First observation:* given that E and E^+ entail observerhood and that H entails that E has positive duration whenever E^+ does, H entails that $\langle FE^+ : \text{observer}\rangle = \langle FE^+ : E\rangle \times \langle E : \text{observer}\rangle$. *Second observation:* PROPORTION yields the following expression for the posterior expectation of any qualitative random variable R conditional on H:

$$\hat{\mathbf{C}}(R|\textit{I am } E \wedge H) = \frac{\hat{\mathbf{C}}(R \times \langle E : \text{observer}\rangle|H)}{\hat{\mathbf{C}}(\langle E : \text{observer}\rangle|H)}$$

This gives us what we need to establish PROPORTIONAL UPDATE:

$$\mathbf{C}_{t^+}(\textit{I am } F|H) = \frac{\mathbf{C}(\textit{I am } F \textit{ and } E^+|H)}{\mathbf{C}(\textit{I am } E^+|H)} = \frac{\hat{\mathbf{C}}(\langle FE^+ : \text{observer}\rangle|H)}{\hat{\mathbf{C}}(\langle E^+ : \text{observer}\rangle|H)}$$
$$= \frac{\hat{\mathbf{C}}(\langle FE^+ : E\rangle \times \langle E : \text{observer}\rangle|H)}{\hat{\mathbf{C}}(\langle E^+ : E\rangle \times \langle E : \text{observer}\rangle|H)} = \frac{\hat{\mathbf{C}}(\langle FE^+ : E\rangle|\textit{I am } E \wedge H)}{\hat{\mathbf{C}}(\langle E^+ : E\rangle|\textit{I am } E \wedge H)} = \frac{\hat{\mathbf{C}}_t(\langle FE^+ : E\rangle|H)}{\hat{\mathbf{C}}_t(\langle E^+ : E\rangle|H)}$$

where the equalities are justified respectively by the fact that E^+ is your evidence at t^+, by PROPORTION, by the first observation, by the second observation, and by the fact that E is your evidence at t. □

PROPORTIONAL UPDATE in turn yields a rule that we can use in the same way that people standardly use Bayes's rule, to express how much a change in evidence favours one qualitative hypothesis over another (when the conditions of PROPORTIONAL UPDATE are met).[14]

$$\frac{\mathbf{C}_{t^+}(H_1)}{\mathbf{C}_{t^+}(H_2)} = \frac{\hat{\mathbf{C}}_t(\langle E^+ : E\rangle|H_1)}{\hat{\mathbf{C}}_t(\langle E^+ : E\rangle|H_2)} \frac{\mathbf{C}_t(H_1)}{\mathbf{C}_t(H_2)}$$

The question what counts as an observer for the purposes of PROPORTION can thus be

[14] *Proof:* $\mathbf{C}_{t^+}(H_1)/\mathbf{C}_{t^+}(H_2) = \mathbf{C}_{t^+}(\textit{I am such that } H_1|H_1 \vee H_2)/\mathbf{C}_{t^+}(\textit{I am such that } H_2|H_1 \vee H_2) = \hat{\mathbf{C}}_t(\langle E^+\text{-such-that-}H_1 : E\rangle|H_1 \vee H_2)/\hat{\mathbf{C}}_t(\langle E^+\text{-such-that-}H_2 : E\rangle|H_1 \vee H_2)$ (by PROPORTIONAL UPDATE) $= (\hat{\mathbf{C}}_t(\langle E^+ : E\rangle|H_1)\mathbf{C}_t(H_1|H_1 \vee H_2))/(\hat{\mathbf{C}}_t(\langle E^+ : E\rangle|H_2)\mathbf{C}(H_2|H_1 \vee H_2)) = (\hat{\mathbf{C}}_t(\langle E^+ : E\rangle|H_1)\mathbf{C}_t(H_1))/(\hat{\mathbf{C}}_t(\langle E^+ : E\rangle|H_2)\mathbf{C}(H_2))$

bracketed when we are only concerned with assessing the impact of new evidence. However, unlike many Bayesians, we are interested in questions about synchronic rationality (what credences are reasonable given certain evidence) as well as diachronic rationality (how one's credences should evolve given certain changes in one's evidence, assuming they were reasonable to begin with). So we do not take this as a dissolution of the question what counts as an observer.[15]

In conclusion, it seems to us that once we modify PROPORTION so as to remove the suggestion that there is a unique, binary notion of observerhood that all reasonable prior credence functions have to respect, the result is an attractive principle that yields defensible results across a wide range of cases. If we only ever had to think about finite populations, the simplicity and strength of this principle combined with the plausibility of its consequences would constitute good grounds for accepting it. However, we cannot reasonably assign a credence of zero to the hypothesis that there are infinitely many observers. And given this, we should surely want whatever we say about the epistemology of self-locating belief in finite worlds to emerge as a special case of some more general epistemological theory that also has something to say about infinite worlds. Thus, we will have to investigate the infinite case before forming a final view about PROPORTION. In the remaining sections we will make a start on this project.

4 Infinite populations

Let us begin with some especially straightforward infinite-world hypothesis, where there is an obvious, uniquely natural way of generalising PROPORTION-style reasoning by taking limits.

> *Chessboard:* Black and white houses are arranged on a two-dimensional plain, in a chessboard pattern. At a certain time, one person is born in each house,

[15]Garriga and Vilenkin (2008) also note that for the purposes of assessing the impact of new evidence, there is no need to talk about any "reference class" other than the one given by the old evidence. They seem to be interested only in what we called "diachronic rationality", and thus take their method to be a full solution to the problem of defining observerhood. PROPORTIONAL UPDATE is an improvement on the updating method they describe, which does not take account of the time-relativity of evidence.

and lives the next twenty years inside that house. At that point all the doors of the houses are unlocked, and the people get to leave their houses, see what colour they are, and explore their immediate neighbourhood. Sixty years later, everyone dies. No other living creatures ever exist.

What prior credence should you have that you were born in a black house, conditional on *Chessboard*? Or equivalently: if you are still locked inside your house, how confident should you, conditional on *Chessboard*, that you will find it to be black when you get let out? The natural answer is 1/2.

It is sometimes suggested that there is a deep conceptual problem about endorsing this natural answer. Perhaps the thought is that in claiming that your credence that you are in a black house conditional on *Chessboard* should be 1/2, we are somehow forgetting about the fact that it is not true to say that the *proportion* of observers are in black houses if *Chessboard* is true is 1/2 (or any other number), since there is no such thing as the ratio of infinity to infinity. But this assumes that claims about proportions provide the only possible basis for favouring some credences over others in this case. We see no grounds for any such assumption.

Finding a fully general principle that entails the natural answer concerning *Chessboard* is a very tall order. But we can take a step in that direction by formulating a principle that tells us how to assign prior credences to self-locating propositions conditional on hypotheses like *Chessboard*, which describe approximately static arrangements of observers in a fixed background space.

LIMITING PROPORTION Suppose that F is some qualitative property, and H is a qualitative proposition that entails that every finite region only ever contains finitely many observers each of whom has a finite life, and that

(i) There is a certain real number x such that, for any all-encompassing nested sequence of concentric spheres $\sigma_1, \sigma_2, \sigma_3 \ldots$, x is the limit of the sequence $x_1, x_2, x_3 \ldots$, where x_i is the proportion of the total duration of the lives of observers whose lives are confined to σ_i during which they are F.

(ii) Observers don't move around too much: there is a finite upper bound to the lengths of the journeys they take over the course of their lives.[16]

Then $\mathbf{C}(I\ am\ F|\mathrm{H}) = x$ for any reasonable prior credence function $\mathbf{C}$.

(We call a sequence of regions 'all-encompassing' just in case its union is the entire space.)

One might worry that there is something objectionably arbitrary about using the family of orderings of observers generated by the nested spheres to set a constraint on priors. After all, so long as the cardinality of F observers is the same as the cardinality of non-F observers, one can find orderings of the observers in which the limiting proportion of F observers takes any value one pleases. However, in general the definition of one of these competitor orderings—or of a family of such orderings that agree on the limiting proportion of F observers—will be far more complicated than the definition of the family of orderings generated by nested sequences of spheres. For example, in the case of *Chessboard* the sequences of observers in which the limiting frequency of observers in black houses is anything other than 1/2 are, intuitively, quite crazy, jumping around in ever-larger leaps with no discernible logic beyond the imperative to make the limiting frequency come out at a specified value. Thus, insofar as one is comfortable with the idea that considerations of simplicity can play a legitimate role in making a difference between reasonable and unreasonable priors, it is hard to see how there could be any deep problem with the thesis that prior credences based on taking limits in nested spheres are more reasonable than prior credences based on taking limits using some ordering that gives a limiting proportion other than 1/2.

Someone might object that our judgment that the nested-sphere-based orderings are simpler than the jumpy orderings that give different limiting frequencies is a merely

[16]Given this condition, we will get the same limiting proportion whether we look, for each sphere, at the observers whose lives are confined to that sphere; or at the observers whose lives overlap that sphere; or at the portions of the lives of observers spent in that sphere. These limits can come apart in far-fetched possibilities where the observers move about at unbounded speeds. We consider the puzzles raised by such possibilities in Arntzenius and Dorr MS.

"relative" one. The notion of a sphere is defined in terms of a certain metric; but given any ordering of the observers, one could always define a new metric according to which that ordering counts as being derived from a sequence of nested "spheres". Relative to the new metric, the sequences that looked well-behaved relative to the old metric will make arbitrarily large jumps. But we don't see any problem here. There are very important differences between the real metric—the one that matters in physics—and the cooked-up quantities relative to which the crazy orderings look simple, and these seem like exactly the sorts of differences we should expect reasonable people to be sensitive to. Moreover, since just about any theory can be made to look simple by expressing it in a language with appropriately cooked-up vocabulary, it is hard to see how considerations of simplicity could play any substantive role in an epistemology that gave no role to the contrast between natural quantities and cooked-up ones.[17]

A different way of motivating the claim that there is a special conceptual problem about infinite populations comes from the idea that reasonable prior credence functions should be *permutation-invariant*, in the following sense:

PERMUTATION-INVARIANCE When H entails that every observer bears R to exactly one observer and that exactly one observer bears R to every observer, $\mathbf{C}(I\ am\ F|H) = \mathbf{C}(I\ bear\ R\ to\ someone\ who\ is\ F|H)$ for any reasonable prior credence function $\mathbf{C}$.

PERMUTATION-INVARIANCE is inconsistent with LIMITING PROPORTION. Suppose we define an ordering of all the observers with no first or last member, in which every third observer is in a black house and the rest are in white houses. If we let R be the relation every observer bears to the next observer in this ordering, PERMUTATION-INVARIANCE entails that $\mathbf{C}(I\ am\ in\ a\ black\ house|Chessboard) = \mathbf{C}(I\ bear\ R\ to\ someone\ in\ a\ black\ house|Chessboard) = \mathbf{C}(I\ bear\ R\ to\ someone\ who\ bears\ R\ to\ someone\ in\ a\ black\ house|Chessboard)$. Since *Chessboard* entails that every observer falls into exactly one of these categories, $\mathbf{C}(I\ am\ in\ a\ black\ house|I\ am\ an\ observer\ and\ Chessboard\ is\ true)$ must equal 1/3 (if it is defined at all). But for the same reason, since we can define periodic orderings of

[17]For more on naturalness and simplicity, see Lewis 1983 and Dorr and Hawthorne 2013.

the observers corresponding to any rational number between 0 and 1, PERMUTATION-INVARIANCE also entails that $\mathbf{C}(\textit{I am in a black house}|\textit{I am an observer and Chessboard is true})$ is either ill-defined or equal to x for every other rational $x \in (0,1)$. This requires that either $\mathbf{C}(\textit{Chessboard})$ is zero, or $\mathbf{C}(\textit{I am an observer}|\textit{Chessboard})$ is zero. Since the same reasoning will apply to other infinite-world hypotheses, the upshot is that we should be sure that the number of observers is not infinite. We take this to be a decisive reason to give up PERMUTATION-INVARIANCE.

5 Infinite worlds with multiple simple measures

LIMITING PROPORTION is not applicable to most of the infinite-population hypotheses that arise in the context of cosmology. The reason for this is that in relativistic spacetimes there is no useful notion of a 'four-dimensional sphere'—the closest analogues of spheres are regions bounded by hyperboloids, but these regions will in general contain infinite numbers of observers and hence be useless for the purpose of taking limits. One possible response to this limitation would be to embark on a quest for a generalisation of LIMITING PROPORTION: a single natural rule which prescribes reasonable prior self-locating credences conditional on any infinite-world hypothesis that physicists are likely to take seriously. But merely formulating a principle at this level of generality, let alone arguing for its truth, would be a very difficult task.

We don't think this is the right way to go. The moral we want to draw from the previous section's qualified defence of LIMITING PROPORTION is not even that LIMITING PROPORTION is true without exception, but that those who are comfortable with the idea that considerations of simplicity play a role in making the difference between reasonable and unreasonable priors face no special conceptual problem when it comes to infinite worlds. In typical cases where the method of taking limits in nested sequences of spheres yields well-defined self-locating priors, it is also far simpler than any method yielding different results. But this is not always the case; and in cases where there are simple self-locating probability functions that disagree with the method of nested spheres, the claim that reasonable priors must accord with that method is much

more tendentious. Consider:

> *Uneven Road:* Inhabited houses are arranged along an infinite road running east-west. At one point there is a wall across the road. To the west of the wall, the houses are 100 metres apart; to the east, they are 10km apart.

How confident should you be, conditional on *Uneven Road*, that you are in the western (thickly settled) part of the road? LIMITING PROPORTION entails that your credence should be 100/101, since this is the limiting proportion of observers to the west of the wall in any all-encompassing sequence of nested concentric spheres. But this is not the only principled answer that that could be given in this case: there is also some temptation to think that your credence should be 1/2. This will seem natural insofar as one is gripped by the thought that the spacing of the houses is rationally irrelevant. And this answer can also be generated by a reasonably simple (albeit rather less general) method of assigning probabilities to self-locating propositions, namely a method which looks, not at nested sequences of concentric spheres, but at nested sequences of segments of the road in which each member of the sequence expands on its predecessor by adding the same number of houses in both directions.

The prior credences prescribed by LIMITING PROPORTION in this case strike us as somewhat dogmatic: finding that you do not live in a densely packed region does not seem like *very* strong evidence against *Uneven Road*. But the claim that you should be equally confident that you are in the western and eastern regions also seems unpromising—it is hard to see what plausible general principle could underlie such a prescription. We suggest that in cases like this, where there are multiple simple recipes for assigning credences to self-locating propositions conditional on some qualitative hypothesis, the most reasonable approach is to split the difference. Conditional on such a hypothesis, reasonable prior credences will be generated by taking a *weighted average* of the credences that result from the different simple methods, in which the simpler ones get weighted more heavily.

To make this talk of 'recipes' more precise: let a *cosmological measure* be a function μ from qualitative properties to qualitative random variables, such that conditional on

the proposition that $\mu(G)$ is defined for at least one property G:

(i) $\mu(F) = 1$ follows from *Always, everything is F*

(ii) $\mu(F) \leq \mu(F')$ follows from *Always, everything F is F'*

(iii) $\mu(F) + \mu(F') = \mu(F'')$ follows from *Always, everything F'' is either F or F'' but not both*

Given a measure μ and any probability function P on qualitative propositions which assigns probability 1 to μ being defined, we can extend P to a self-locating probability function $P^{[\mu]}$ simply by taking $P^{[\mu]}(I\ am\ F)$ to be $\hat{P}(\mu(F))$. A self-locating probability function thus corresponds to the combination of a qualitative probability function and a cosmological measure. In these terms, our proposed "compromising" approach can be stated as follows:

COMPROMISE For any reasonable prior $\mathbf{C}$ and sufficiently specific qualitative H,

$$\mathbf{C}(I\ am\ F|H) = \sum_i w_i \hat{\mathbf{C}}(\mu_i(F)|H)$$

where μ_i are simple measures which are well-defined according to H, and w_i are weights summing to 1, generally higher for simpler μ_i.

The 'sufficiently specific' H should be, at the minimum, specific enough to settle, of each simple measure, whether it is well-defined or not. More generally, the point is to focus on hypotheses that pin things down in enough detail that there is no controversy about what the *qualitative* priors should be conditional on them, so that all of the debate pertains to the self-locating priors.

(We can be more precise about the restriction to 'sufficiently specific' H if we derive COMPROMISE from the following attractive account of the role of simplicity in reasonable priors:

SIMPLICITY A reasonable prior credence function $\mathbf{C}$ is a weighted average $\sum_i w_i P_i$ of self-locating probability functions P_i, where w_i is generally higher for simpler P_i.

Plausibly all or almost all of the weights should go to probability functions P_i that are of the form $Q_i^{[\mu_i]}$ for some qualitative probability function Q_i and measure μ_i. In this setting, the 'sufficiently specific' H mentioned by COMPROMISE can be characterised as those for which $Q_i(\cdot|H)$ and $Q_j(\cdot|H)$ are everywhere approximately identical whenever both are defined and Q_i and Q_j are simple enough to receive significant weight. If this condition is met, $\hat{\mathbf{C}}(R|H)$ will be approximately the same as $\hat{Q}_i(R|H)$ for any qualitative random variable R; and so we have $\mathbf{C}(I\ am\ F|H) = \mathbf{C}(I\ am\ F \wedge H)/\mathbf{C}(H) = \sum_i w_i P_i(I\ am\ F \wedge H)/\mathbf{C}(H) = \sum_i w_i \hat{Q}_i(\mu_i(F\text{-such-that-}H))/\mathbf{C}(H) \approx \sum_i w_i \hat{\mathbf{C}}(\mu_i(F\text{-such-that-}H))/\mathbf{C}(H) = \sum_i w_i \hat{\mathbf{C}}(\mu_i(F)|H)$.)

COMPROMISE is especially plausible when we turn to hypotheses in which LIMITING PROPORTION does not apply, but where there are multiple other simple methods of assigning probabilities to self-locating propositions. In many such cases, the quest for a general principle which would privilege a particular simple method as the one corresponding to a reasonable self-locating prior credence function seems misguided. Consider:

> *The Cliff:* There is an infinite half-plane dotted with black, grey and white houses, terminated to the north by a straight, infinite cliff-edge. The houses and their inhabitants get exponentially smaller and more tightly packed as we approach the cliff. The distance between the centre of a house and the cliff is always a power of two (in metres). The black houses on the line 2^n metres from the cliff are distributed randomly, in such a way that the expected number of houses in each segment of length l is equal to $l/2^n$: thus the average spacing between black houses on a given line is equal to the distance between that line and the cliff. The distribution of grey and white houses is determined by the distribution of black houses, as follows: for each black house, there is a grey house exactly halfway between it and the cliff, and for each grey house, there is a white house exactly halfway between it and the cliff. These are the only grey and white houses. (See Figure 1.)

There are two simple ways of assigning probabilities to self-locating propositions

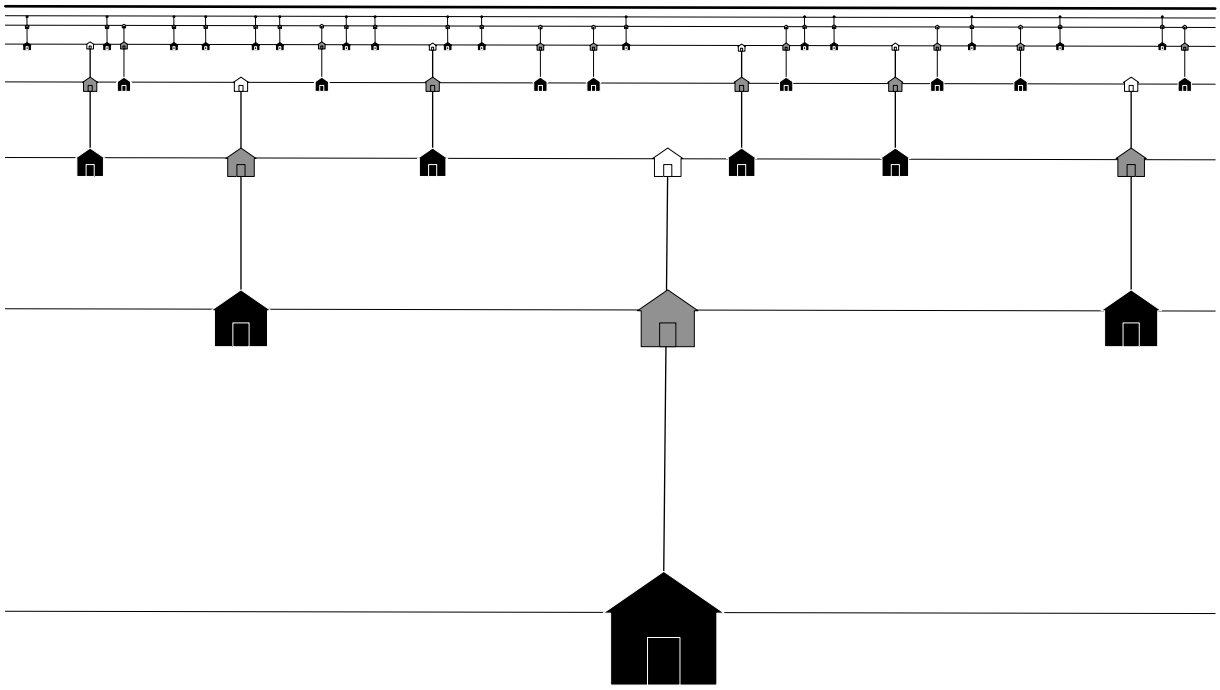

Figure 1. *The Cliff*

conditional on *The Cliff*, which assign different probabilities to *I am in a black house*. One approach assigns a probability of 1/3, based on the fact that each north-south line of houses contains three houses of which one is black. The other approach assigns a probability of 4/7, based on the fact that on any given *east-west* line of houses, black houses are twice as common as grey houses, and grey houses are twice as common as white houses, so that (with chance 1) the limiting proportion of black houses along any east-west line is 4/7. For the same reason, the limiting proportion of black houses in any all-encompassing nested sequence of *rectangular* regions (all with finite populations) will also be 4/7.[18]

Instead of the compromising approach, one might suggest a permissivist approach according to which *both* of the simple self-locating probability functions correspond to some optimally reasonable prior credence function conditional on *The Cliff*. More generally, the thought would be each of any sufficiently simple measure can be used

[18]In this case there are no all-encompassing, nested sequences of *concentric* circles each of which has a finite population: since circles that extend past the cliff-edge have infinite populations, any all-encompassing sequence of nested circles in which each circle has a finite population must have centres that get further and further away from the cliff edge. The limiting proportion of black houses in any such sequence is also 4/7.

to generate a maximally reasonable prior credence function conditional on a given detailed qualitative hypothesis. (There is some pressure for those who take this permissivist view to allow that weighted averages of maximally reasonable prior credence functions are also maximally reasonable.) But this view is implausibly liberal. By modifying the details of *The Cliff*, one can make the two candidate credences as close as one pleases to 1 and 0—just let each black house be the southernmost member of a very long series of non-black houses, and allow the density of black houses to increase by an arbitrarily large factor with each step towards the cliff.[19] The extreme credence functions that assign *I am in a black house* probabilities close to 0 or 1 in these cases seem clearly less reasonable than the weighted-average credence functions that assign intermediate credences—it seems unnecessarily dogmatic to treat either the discovery that one is in a black house, or the discovery that one is not, as incredibly weighty evidence against *The Cliff*.

In endorsing COMPROMISE, we do not mean to commit ourselves to the strong claim that simplicity is the *only* factor relevant to setting the weights assigned in a reasonable prior. There may be some further conditions that measures need to meet in order to deserve any weight at all. Note that all the measures we have considered make reference to a notion of observerhood, something which—as we discussed in connection with PROPORTION—could be understood in many different ways. A flat-footed extension of the compromising approach to the question what should count as an observer, according to which which we simply take a weighted average of all probability functions which can be defined by appealing to simple criteria of observerhood—even crazy ones that count rocks as observers!—would have quite implausible consequences. Thus, we can already see that a fuller articulation of the compromising approach will need to appeal to some considerations other than simplicity to keep the final weighted average from being dominated by probability

[19] Indeed, if we modify the case so that each black house is the southernmost member of an *infinite* series of non-black houses, the north-south way of assigning probabilities will require assigning probability zero to being in a black house, while we can still make the east-west proportion of black houses arbitrarily close to 1.

functions defined using crazy (but simple) criteria of observerhood. The question what these considerations should look like is closely bound up with the question whether PROPORTION is true. It could easily turn out that the best way of excluding the crazy probability functions has as a consequence that the only probability functions that should receive any positive weight are functions that agree with PROPORTION in finite worlds (as do all the limiting procedures that we have considered). If this proved to be so, it would be good news for PROPORTION. If not, PROPORTION might start to look like an ugly and *ad hoc* addition, out of keeping with the spirit of the compromising approach. We leave further investigation of this question as a topic for future research.

6 The compromising approach and the measure problem in cosmology

Let us consider one way in which modern cosmology prompts us to take seriously the possibility that there are infinitely many observers. According to the theory of inflation, if you follow the geodesic paths that are currently occupied by observable galaxies back far enough, you eventually—after 14 billion years or so—reach an "inflationary" era during which the paths (as we follow them backward) approach one another at an exponential rate. This theory is fantastically successful by normal scientific standards. But models of the universe as a whole which provide a mechanism for such inflation typically feature *eternal* inflation—a kind of universe in which pockets of ordinary, non-inflating space keep forming, but in such a way that that the inflating portion of space is never completely filled, but keeps expanding and giving rise to new non-inflating pockets. In the most plausible such models, there are many different kinds of pockets, only a few of which are hospitable to life. Nevertheless there is plenty of life: in fact there will be infinitely many life-friendly pockets as well as infinitely many life-unfriendly ones, and the life-friendly pockets will typically contain infinitely many observers each. We attempt to illustrate the general picture in Figure 2.

Since these hypotheses are set in relativistic spacetime, the proposal to assign probabilities by taking limiting relative frequencies in sequences of nested spheres doesn't even make sense. Nor does any other way of assigning self-locating probabilities look

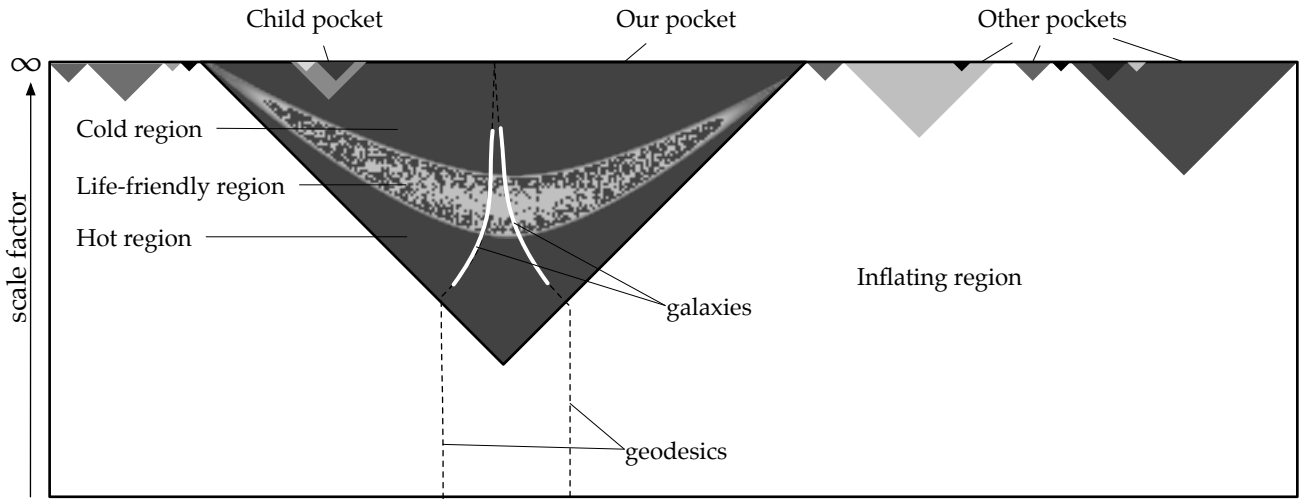

Figure 2. Eternal inflation

to be overwhelmingly simpler than all the others. Instead, what we generally find is a multiplicity of non-equivalent, reasonably simple cosmological measures. Here are a just a few examples. The "proper time measure" (see Garcia-Bellido and Linde 1995, Linde 1986) is defined by starting with an (almost) arbitrary bounded, smooth, spacelike hypersurface; using it to construct a nested sequence of four-dimensional regions, each of which is the union of all timelike geodesic segments of a particular finite length, perpendicular to and extending futurewards from the chosen hypersurface; and assigning self-locating probabilities by taking limits within this sequence of regions, as in LIMITING PROPORTION. The "scale factor measure" (De Simone, Guth, Linde, Noorbala, Salem and Vilenkin 2010) is similar, except that instead of following the geodesics along by constant amounts of proper time, we use a time co-ordinate given by the scale factor—essentially, we follow each geodesic as far as we need to to reach a hypersurface on which the distances between nearby geodesics are a constant multiple of the distances between them on the initial hypersurface. For the "causal diamond measure" (Bousso 2006), we choose a timelike geodesic and take limits in a nested sequence of four-dimensional regions sandwiched between the forward light cones of points increasingly early on that geodesic and the backward light cones of points increasingly late on that geodesic. (Unlike the sequences of nested spheres considered in LIMITING PROPORTION, the sequences of regions considered by these three measures

need not be all-encompassing; nevertheless, if the definitions of the measures are filled in properly, all the sequences of regions meeting the relevant criteria should yield the same limiting proportions.) There are also measures that are not based on taking limits in sequences of bounded regions at all: for example, the "pocket-based measure" of Garriga, Schwartz-Perlov, Vilenkin and Winitzki (2006) works in two stages, first generating a separate measure for each vacuum state, and then aggregating these using a separate recipe for assigning weights to the different vacuum states.

Some of these measures have turned out to be "pathological" in that they assign vanishingly low probability to our actual evidence—according to them, "almost all" observers in the relevant models are, in some way or other, incredibly unlike us. For example, the proper time measure suffers from what is sometimes called the 'youngness paradox' (see Bousso, Freivogel and Yang 2008, Tegmark 2005). Since new pocket universes are being created in the inflating region at such a high rate, at each region in one of the relevant nested sequences, the pocket universes that have only just been added constitute a high proportion of all the pocket universes in that region. Because there are so many "new" pockets, the observers in the new pockets who have come into existence quite soon after their local Big Bang (the initial boundary of their pocket) outnumber all the observers in the older pockets. As a result, when we take limits using these sequences, observations like ours—e.g. of measuring the temperature of the microwave background to be 2.7K—will be assigned far lower probabilities than the kinds of observations that would be expected closer to the Big Bang, e.g. measurements of higher background temperatures. Indeed, the measure is so drastically skewed towards early observers that by its lights, "almost all" observers are "Boltzmann Babies"—freak observers who fluctuate into existence at times when the universe was so hot as to be extremely inhospitable to life, and spend the entirety of their short lives being roasted to death by their infernally hot surroundings.[20] Some

[20]It is also worth noting that, even conditional on our actual evidence about the present and past, a probability function generated by the proper time measure will assign high probability to the proposition that we are about to *start* getting roasted to death.

other measures suffer from a more familiar kind of pathology: they are dominated not by observers who live much *closer* than us to their local Big Bang, but by observers who live much *further*—Boltzmann Brains who have fluctuated into existence in the endless freezing vacuum. At least some versions of the pocket-based measure suffer from this pathology (Page 2008). The problem is that there is a sense in which each and every inhabited pocket universe with a positive cosmological constant is dominated by Boltzmann Brains: if you begin with a finite-volume portion of the initial boundary of such a pocket universe, and evolve this region further and further forwards into the future (excluding any 'child' pockets that might form inside the initial one), you will continue to add Boltzmann Brains without bound, but you will only ever have come across finitely many ordinary observers.[21] This makes it quite challenging to devise a simple measure on observers in a given vacuum state that is not dominated by Boltzmann Brains (and, as usual, also dominated by Boltzmann Brains of the usual sort, living brief and bizarre lives).[22]

There is a rough analogy here with *The Cliff*. The spacelike surfaces represented by horizontal lines in Figure 2 are dominated by Boltzmann Babies, just as the east-west lines in *The Cliff* are dominated by black houses, whereas timelike paths (ignoring child pockets) are dominated by Boltzmann Brains, just as the north-south lines in *The Cliff* are dominated by non-black houses. Fortunately, whereas in the case of *The Cliff* there don't seem to be any other comparably simple measures, in the case of eternally inflating spacetime there is a wider array of alternatives that have not been shown to

[21] In Figure 2, imagine that the two dotted geodesics are the boundaries of a region that is open to the future but bounded in spacelike directions. Then the intersections of this region with the "hot" and "life-friendly" parts of our pocket universe have finite spacetime volume and contain finitely many observers, whereas the intersection of this region with the "cold" part has infinite spacetime volume, and contains infinitely many Boltzmann Brains.

[22] These claims about the existence of Boltzmann Brains in the very late parts of pockets with positive cosmological constants are standard, but are sensitive to issues about the interpretation of quantum theory. Boddy, Carroll and Pollack this volume and Goldstein, Struyve and Tumulka n.d. point out ways in which the question of Boltzmann Brains may require rethinking on Everettian and pilot-wave interpretations, respectively. By contrast, the earlier remark about the domination of the proper time measure by Boltzmann Babies seems much less interpretation-sensitive.

be in any way pathological.

In introducing the multiplicity of measures, cosmologists often characterise it as a deep foundational problem—the 'measure problem'. Tegmark (2014, p. 314) goes so far as to call it "the greatest crisis facing physics today". Some regard this problem as a weighty reason to reject inflation in favour of some rival theory: for instance, Steinhardt (2011, pp. 42–3) says that 'The notion of a measure, an ad hoc addition, is an open admission that inflationary theory on its own does not explain or predict anything', and rhetorically asks 'If inflationary theory makes no firm predictions, what is its point?'. Many others regard the problem as analogous to the divergences in quantum field theory: not a reason to reject the relevant hypotheses altogether, but a reason to believe that they are mere approximations to, or characterisations of some emergent behaviour in, some yet-to-be-discovered underlying theory that does not suffer from the problem. For instance, Tegmark (2014, p. 316) thinks that the measure problem is 'telling us' that we will have to give up on the idea that 'space can have an infinite volume, that time can continue forever, and that there can be infinitely many physical objects'. Similarly, Bousso and Freivogel (2007) treat the pathological features of certain measures as a reason to treat the infinite population of observers as 'figments of our imagination', instead favouring the minimalistic (and, *prima facie*, objectionably anthropocentric) view that a causal diamond that includes everything causally accessible from our worldline is 'all there is, as far as the semiclassical description of the universe goes'.[23]

While we have no objections to physicists following their hunches, one moral we want to draw from our discussion in previous sections is that the "measure problem" does not constitute a *reason to disbelieve* the infinite-population hypotheses that give rise to it. There is no good a priori or empirical reason to be confident that our universe is one of the "well-behaved" infinite ones with a unique simple measure, let alone that it is finite. The fact that there are *some* simple self-locating probability functions that give high probability to a certain hypothesis about the structure of the world as a whole while giving a vanishingly small probability to our evidence does

[23]Bousso and Freivogel seem to think that this eliminativist attitude is required by the use of their causal diamond measure, but we do not see why this should be so.

not constitute any kind of reason to reject that hypothesis. What would constitute a reason to reject the hypothesis would be the discovery that *every* simple self-locating probability function that assigns it substantial probability assigns vanishingly small probability to our evidence.[24] And we have not discovered anything like this in the case of eternal inflation.

Despite the hand-wringing about foundational problems, the actual practice that cosmologists have adopted in reasoning scientifically about infinite-population hypotheses looks very much like what would be recommended by our "compromising" account of the role of simplicity considerations in reasonable priors. A theorist will come up with a definition of a measure that works for a certain model of the cosmos, and try to figure out what kinds of observations are probable according to that measure (often a difficult task). When a particular measure is shown to have some pathological feature such as assigning high probability to being a Boltzmann Baby, the theorist's reaction is not to give up straight away on the relevant cosmological model, or to suddenly start treating sceptical scenarios in which we actually *are* Boltzmann Babies (or freak observers of some other sort) as live options. Rather, the theorist will start looking for some alternative simple measure which does not suffer from any such pathology. The goal is to find a pair of a simple cosmological hypothesis and a simple measure that together give reasonably high probability to certain characteristic properties attributed to us by our actual evidence, such as observing a background temperature not too far from what we actually observe. And the cosmologists seem to be making considerable progress towards this goal—indeed the whole enterprise looks like science at its best.

All of this is exactly as it should be on the compromising approach. When generating reasonable self-locating priors from reasonable qualitative priors by setting $\mathbf{C}(I\ am\ F|H) = \sum_i w_i \hat{\mathbf{C}}(\mu_i(F)|H)$, it will often happen that some of the μ_i simple enough to for the weight factor w_i to be substantial will be "pathological" in the sense that

[24] 'Vanishingly small' here really means: small by comparison with the probability assigned to our evidence by other simple self-locating probability functions that do not assign high probability to the given cosmological hypothesis.

they a very low measure even to the barest outlines of our actual evidence. For example, there may be some simple μ_i such that H entails that μ_i(Boltzmann Baby in the midst of being roasted to death) is close to 1. But this is not a problem: so long as there are *any* reasonably simple measures that are not pathological in this way, our posterior credences will be dominated by those terms. That is: our posteriors will be approximately as they would be if we had started, not with our actual priors, but with a weighted average that excluded the pathological measures. Indeed, it is reasonable to hope that by doing the right experiments, we can enrich our total evidence to the point where one simple measure μ_k will assign it a vastly higher probability than any other comparably simple measure. In that case, our posterior self-locating credence $\mathbf{C}_t(I\ am\ F) = \mathbf{C}(I\ am\ F \wedge E)/\mathbf{C}(\mathrm{E})$ will be approximately equal to $\hat{\mathbf{C}}(\mu_k(F \wedge E))/\hat{\mathbf{C}}(\mu_k(E))$. For the purposes of making predictions, we won't have to think about anything other than the particular simple measure μ_k, and we will not need to concern ourselves with detailed questions about how exactly simplicity should be measured and weighted.

7 'Which measure does nature subscribe to?'

Cosmologists sometimes say rather mysterious things in describing what we are learning about the world when we engage in this process of defining different measures and trying to find out which ones assign high probability to our evidence. For example, Max Tegmark describes the enterprise as follows:

> There is some correct measure that nature subscribes to, and we need to figure out which one it is, just as was successfully done in the past for the measures allowing us to compute probabilities in statistical mechanics and quantum physics. (Tegmark 2005, p. 2)

This remark suggests the following picture. In addition to facts of the familiar sort studied by physics (facts about the disposition of fields in spacetime, the wavefunction, and so on), there are facts about *which measure nature subscribes to*. We can investigate such facts empirically, since our evidence that we have a certain property F counts in favour of the hypothesis that nature subscribes to a measure μ for which $\mu(F)$ is

high. More generally: for any measure μ, when we conditionalise any reasonable prior credence function $\mathbf{C}$ on the proposition that nature subscribes to μ, the result (if well-defined) is just given by μ itself:

DEFER TO NATURE $$\mathbf{C}(I \text{ am } F | \text{Nature subscribes to } \mu) = \hat{\mathbf{C}}(\mu(F))$$

Given this, all that remains to be done to to pin down a particular reasonable prior credence function $\mathbf{C}$ is to specify $\mathbf{C}(\text{Nature subscribes to } \mu)$ for every μ. Presumably this should be higher for simpler μ.

Although Tegmark's remark is unusually explicit, the picture is not unique to him. It is also suggested by a certain way of using the word 'theory' that some cosmologists favour in this context, on which a theory is something that 'builds in' or 'comes with' a particular measure or self-locating probability function. As Linde (2007, p. 32) puts it: 'the probability measure becomes a part of the theory, and we test both the theory and the measure by comparing them with observations'. By itself this is a harmless terminological choice; but it becomes consequential when it is combined with the natural assumption that theories are propositions, things capable of being true or false.[25] For clearly it makes no sense to say that a measure—which is just a function from properties to real numbers obeying certain axioms—is true or false. The question has to be whether the measure enjoys some special status that plays something like the epistemological role characterised by DEFER TO NATURE. And the choice becomes a serious metaphysical commitment if it is combined with the further assumption that the truth or falsity of the relevant theories remains a qualitative question, so that the relevant distinguished status is not merely a distinguished relation that the relevant measure stands in to *us*.[26]

We find these putative facts about what measure nature subscribes to obscure and

[25]We do not attribute this assumption to Linde.

[26]For example, Page (this volume) proposes that each 'complete theory for a universe' should 'give normalized probabilities for the different possible observations O_j that it predicts, so that for each T_i, the sum of $P(O_j|T_i)$ over all O_j is unity', while also defining a 'complete theory for a universe' as one that 'completely describes or specifies all properties of that universe', in a context where it it is clear that the 'properties' in question are *qualitative* properties.

problematic. We see no good reason to posit them at all, rather than taking seriously a more economical picture of reality as fully characterised by facts of a more familiar physical kind (such as facts about fields in spacetime).

You may have noticed that DEFER TO NATURE is structurally parallel to (PP), the standard 'Principal Principle' that specifies how reasonable prior credence functions behave conditional on a hypotheses about what probability function enjoys a different special status, that of being the true *objective chance* function. This might suggest that those who do not regard facts about objective chance as objectionably spooky should have no metaphysical objection to facts about what measure nature subscribes to. But there is at least the following disanalogy between the two cases. One widely-held view about objective chance is Humean reductivism, which in its best-developed form (Elga 2004, Lewis 1994), identifies the proposition that a probability function P is the objective chance function with the proposition that P optimally balances the desiderata of simplicity and "fit" (assigning high probability to truths, especially simple truths). It is reasonable to hope that by finding the right weighting of these factors, we could define a property F, necessarily instantiated by at most one probability function, such that $P(P \textit{ is } F)$ will be close to 1 for any P simple enough to deserve substantial weight in a reasonable prior credence function. In that case, $\mathbf{C}(\cdot|P \textit{ is } F)$ will be well-approximated by P, so that (PP) will be approximately true (see Lewis 1994, §10). We could certainly adopt a similar, reductivist Humean account of "being the measure to which nature subscribes". But on this understanding, the talk of 'nature' is deeply misleading, since the proposition that nature subscribes to μ will be a self-locating proposition rather than a qualitative one. If there is a simple measure μ that assigns a much higher probability to the properties that truly characterise *me* than any other simple measure, then I will speak truly when I say 'Nature subscribes to μ'. But someone else (some Boltzmann Brain, say), whose qualitative properties are atypical by the lights of μ but typical by the lights of some other simple measure μ^*, will say something false in uttering the same sentence. If we want to conceive of facts about what measure nature subscribes to as qualitative, Humean reductivism is not an option.

Our metaphysical objection to positing irreducible facts about what measure nature subscribes to will, of course, not weigh so heavily with those who already endorse an anti-Humean realist account of objective chance. But even they should want some argument for positing such facts. In the case of facts about objective chance, one can point to the pervasive role the concept seems to play in our ordinary thought and in the sciences. By contrast, the concept "measure to which nature subscribes" seems like a recent innovation; so far, no-one seems to have given anything that looks like an argument for positing a domain of facts answering to this concept.

One argument that might be given for positing such facts is that there is no way to understand the rationality and cognitive significance of the cosmologists' enterprise of looking for simple cosmological models and simple measures on those models which assign reasonably high probability to our evidence—without making such a posit. But one of the morals of our discussion in the previous sections is that this argument is unsuccessful. Take a reasonable prior credence function as conceived by someone who posits facts about what nature subscribes to, and simply delete all the probabilities assigned to propositions about that peculiar subject-matter, leaving the probabilities of ordinary qualitative and self-locating propositions unchanged. So long as the original probability function gave simpler measures a higher probability of being subscribed to, the output of this procedure will fit our compromising picture of reasonable priors: it will be a weighted average of self-locating probability functions in which simple ones receive higher weight. Thus, insofar as we are interested in how our evidence bears on these ordinary questions, there is no need to take seriously the idea that we are uncovering hidden facts about what nature subscribes to. We can treat this as nothing more than a colourful manner of speaking: for example, we can say 'We have learnt that nature subscribes to μ' when what we really mean is something like 'We have received evidence such that the result of conditionalising our priors on it is approximately equal as the result of conditionalising a self-locating probability function corresponding to μ on it'.

8 Conclusion

To sum up: for completely ordinary reasons, we should realise that the propositions we believe and that constitute our evidence are not always qualitative, and we should implement the familiar Bayesian idea that a reasonable credence function is derived from reasonable priors by conditionalisation on evidence in a way that treats self-locating propositions and qualitative propositions as being on a par. In such a framework there is nothing mysterious or surprising about the idea that our total self-locating evidence *I am E* might support one qualitative proposition over another, even when both propositions entail that *someone* is *E*, or entail that there are infinitely many people. The question *when* this happens reduces to the question what priors are reasonable. We should not, in general, expect to be able to establish our claims about reasonable priors by deducing them from precisely stated, uncontroversial principles, or regard it as a deep problem for some hypothesis about the world if we cannot find principles of this sort which completely pin down the result of conditionalising any reasonable prior credence function on that hypothesis. (If this is what it means for a hypothesis to make "firm predictions", firm predictions are overrated.) Once we give up on the misguided hope for a knockdown demonstration that reasonable priors have to work in a certain way, we can get a long way with the familiar, vague idea that reasonable priors should be influenced by considerations of simplicity. In particular, the idea that reasonable priors are weighted averages of simple probability functions (in which simpler probability functions are weighted more heavily) yields prescriptions for reasoning about infinite-population hypotheses that are both intuitively plausible, and a good fit for the methodology that cosmologists have actually adopted in practice.

References

Albrecht, Andreas and Lorenzo Sorbo (2004). 'Can the Universe Afford Inflation?' *Physical Review D* 70.6, p. 063528. arXiv: hep-th/0405270.

Arntzenius, Frank (2003). 'Some Problems for Conditionalization and Reflection'. *Journal of Philosophy* 100, pp. 356–70.

— (2014). 'Utilitarianism, Decision Theory and Eternity'. *Philosophical Perspectives* 28.1, pp. 31–58.

Arntzenius, Frank and Cian Dorr (MS). 'What to Expect in an Infinite World'.

Bartha, Paul and Christopher Hitchcock (1999). 'No-one Knows the Date or the Hour: An Unorthodox Application of Rev. Bayes's Theorem'. *Proceedings of the Biennial Meeting of the Philosophy of Science Association* 66, S339–S353.

Boddy, Kimberly K., Sean M. Carroll and Jason Pollack (this volume). 'Why Boltzmann Brains Don't Fluctuate Into Existence From the De Sitter Vacuum'.

Bostrom, Nick (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge.

Bousso, Raphael (2006). 'Holographic Probabilities in Eternal Inflation'. *Phys. Rev. Lett.* DOI: 10.1103/PhysRevLett.97.191302. arXiv: hep-th/0605263.

Bousso, Raphael and Ben Freivogel (2007). 'A Paradox in the Global Description of the Multiverse'. *Journal of High Energy Physics 06* 2007, p. 18. arXiv: hep-th/0610132.

Bousso, Raphael, Ben Freivogel and I-Sheng Yang (2008). 'Boltzmann Babies in the Proper Time Measure'. *Physical Review D* 77.10, p. 103514. arXiv: 0712.3324 [hep-th].

Carroll, Sean M. (2010). *From Eternity to Here*. New York: Dutton.

De Simone, Andrea, Alan H. Guth, Andrei Linde, Mahdiyar Noorbala, Michael P. Salem and Alexander Vilenkin (2010). 'Boltzmann Brains and the Scale-Factor Cutoff Measure of the Multiverse'. *Physical Review D* 82.6, p. 063520. arXiv: 0808.3778 [hep-th].

Dorr, Cian (2010). 'The Eternal Coin: A Puzzle About Self-Locating Conditional Credence'. *Philosophical Perspectives* 24.1, pp. 189–205.

Dorr, Cian and John Hawthorne (2013). 'Naturalness'. In *Oxford Studies in Metaphysics*, vol. 8, ed. Karen Bennett and Dean Zimmerman. Oxford: Oxford University Press, pp. 3–77.

Elga, Adam (2004). 'Infinitesimal Chances and the Laws of Nature'. *Australasian Journal of Philosophy* 82.1, pp. 67–76.

Garcia-Bellido, Juan and Andrei Linde (1995). 'Stationarity of Inflation and Predictions of Quantum Cosmology'. *Physical Review D* 51.2, p. 429. arXiv: hep-th/9408023.

Garriga, Jaume, Delia Schwartz-Perlov, Alexander Vilenkin and Sergei Winitzki (2006). 'Probabilities in the Inflationary Multiverse'. *JCAP* 0601, p. 017. arXiv: hep-th/0509184v3.

Garriga, Jaume and Alexander Vilenkin (2008). 'Prediction and Explanation in the Multiverse'. *Physical Review D* 77, p. 043526. arXiv: 0711.2559v3 [hep-th].

Goldstein, Sheldon, Ward Struyve and Roderich Tumulka. 'The Bohmian Approach to the Problems of Cosmological Quantum Fluctuations'.

Hájek, Alan (2003). 'What Conditional Probability Could Not Be'. *Synthese* 137, pp. 273–323.

Halpern, Joseph (2004). 'Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems'. In *Oxford Studies in Epistemology*, vol. 1, ed. Tamar Szabo Gendler and John Hawthorne. Oxford University Press, pp. 111–142.

Halpern, Joseph and Mark Tuttle (1993). 'Knowledge, Probability, and Adversaries'. *Journal of the Association for Computing Machinery* 40, pp. 917–62.

Hartle, James B. and Mark Srednicki (2007). 'Are We Typical?' *Physical Review D* 75.12, p. 123523. arXiv: `0704.2630v3 [hep-th]`.

Lewis, David (1979). 'Attitudes *De Dicto* and *De Se*'. *Philosophical Review* 88, pp. 513–43.

— (1980). 'A Subjectivist's Guide to Objective Chance'. In *Studies in Inductive Logic and Probability*, vol. 2, ed. R. C. Jeffrey. Berkeley: University of California Press, pp. 263–93.

— (1983). 'New Work for a Theory of Universals'. *Australasian Journal of Philosophy* 61, pp. 343–77.

— (1994). 'Humean Supervenience Debugged'. *Mind* 103, pp. 473–90.

Linde, Andrei (1986). 'Eternally Existing Self-reproducing Chaotic Inflanationary Universe'. *Physics Letters B* 175.4, pp. 395–400.

— (2007). 'Sinks in the Landscape, Boltzmann Brains and the Cosmological Constant Problem'. *Journal of Cosmology and Astroparticle Physics* 2007, p. 022. arXiv: `hep-th/0611043`.

Meacham, Christopher (2008). 'Sleeping Beauty and the Dynamics of *De Se* Beliefs'. *Philosophical Studies* 138, pp. 245–69.

Myrvold, Wayne C. (2015). 'You Can't Always Get What You Want: Some Considerations Regarding Conditional Probabilities'. *Erkenntnis* 80, pp. 573–603.

Neal, Radford M. (2006). *Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning*. Technical Report 0607. Department of Statistics, University of Toronto. arXiv: `math/0608592v1`.

Page, Don (2008). 'Is Our Universe Likely to Decay Within 20 Billion Years?' *Phys. Rev. D* 78, p. 063535.

— (this volume). 'Cosmological Ontology and Epistemology'.

Srednicki, Mark and James B. Hartle (2010). 'Science In a Very Large Universe'. *Physical Review D* 81.12, p. 123524. arXiv: `0906.0042v3 [hep-th]`.

— (2013). 'The Xerographic Distribution: Scientific Reasoning in a Large Universe'. In *Journal of Physics: Conference Series*. Vol. 462. 1. IOP Publishing, p. 012050. arXiv: `1004.3816v1 [hep-th]`.

Steinhardt, Paul (2011). 'The Inflation Debate'. *Scientific American*, pp. 36–43.

Tegmark, Max (2005). 'What Does Inflation Really Predict?' *Journal of Cosmology and Astroparticle Physics* 2005.

— (2014). *Our Mathematical Universe*. New York: Knopf.

Titelbaum, Michael G. (2013). 'Ten Reasons to Care About the Sleeping Beauty Problem'. *Philosophy Compass* 8.11, pp. 1003–1017.

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