How to Reason about Self-Locating Belief

David Shulman

Department of Mathematics

Delaware Valley College

July 5, 2011

Abstract

When reasoning about self-locating belief, one should reason as if one were a randomly selected bit of information. This principle can be considered to be an application of Bostrom's Strong Self-Sampling Assumption(SSSA)[2] according to which one should reason as if one were a randomly selected element of some suitable reference class of observer-moments. The reference class is the class of all observer-moments. In order to randomly select an observer-moment from the reference class, one first randomly chooses a possible world w and then selects an observer-moment z from world w. The probability that one selects z given that one has chosen w should be proportional to the amount of information that z is capable of representing. There are both wagering arguments and relative frequency arguments that support our theory of anthropic reasoning. Our theory works best when the amount of information represented is finite. The infinite case is represented as a limit of a finite cases. We can learn from experience how best to represent the infinite case as a limit of finite cases and also learn from experience whether our theory or some other theory is the superior theory of anthropic reasoning. In order to test which theory is best, we use standard Bayesian methodology: We just need prior probabilities for the theories that are being tested and then we only have to use Bayes' rule.

1 Reasoning about Self-Locating Belief

In this paper, we describe a general, Bayesian theory of anthropic reasoning. We use the expressions "anthropic reasoning" and "reasoning about self-locating belief" to refer to reasoning about our identity or our temporal location. If we are suffering from amnesia, we need to figure out who we are. If we have an appointment to meet someone at a certain time, we might wish to know the current time. Sometimes, we do not really care about our identity or temporal location except for the fact that reasoning about identity or temporal location might help us figure out which worlds are most likely to be actual.

Anthropic reasoning is ubiquitous in both science and everyday life, but there are many difficult to analyze puzzles involving probabilistic anthropic reasoning such as Sleeping Beauty[8], Doomsday[13, 7], Lazy Adam[2], and Absent-Minded Driver[19]. We need a general theory that will allow us to analyze any anthropic reasoning problem.

There already exists a well-known general rule for reasoning about self-locating belief. According to the Self-Sampling Assumption (SSA)[2], one should reason as if one were a observer randomly selected from some suitable reference class of observers. Here, by an observer is meant something that is capable of reasoning about self-locating belief. This is a very crude and imprecise definition of what it means to be an observer, but it will suffice for now. However we make the definition more precise, we want to be able to model normal adult humans as observers.

The SSA is almost good enough but because observers can believe and desire different things at different times or make different choices at different times, we need the Strong Self-Sampling Assumption (SSSA)[2] according to which we should reason as if we were a randomly selected element of some suitable class of observer-moments. By an observer-moment we mean an ordered pair consisting of an observer and a time interval (and that time interval might just be a single point in time). We would use observer-moments to represent a time-slice of an observer. But there might be other relevant ways to split an observer into parts with each part being modelled as being capable of having its own coherent beliefs and desires and making its own coherent decisions. In that case, we would consider each of these parts to be a separate observer-moment. For example, if someone is suffering from multiple personality disorder and thus might be modelled as consisting of several different personalities that exist at the same time, we might consider temporal slices of these personalities to be observer-moments.

The SSSA is a perfectly fine general principle, but it is too underspecified. We need to know what is and what is not an observer. The random selection from the reference class presupposes the existence of a prior probability distribution on that reference class and the question arises what probability distribution should be used[1]. We do not expect a theory of anthropic reasoning to tell us how to distribute prior probability among the many different possible worlds[2], but we do expect the theory to tell us how the prior probability that is given to a possible world is apportioned to the several observer-moments that inhabit that world.

[1] And we, of course, also need to know how to choose reasonable reference classes.

[2] However, it might be unrealistic to believe that we should separately analyze the nonanthropic and anthropic aspects of the problem of selecting a prior.

We need a precise theory that will tell us which prior probability distributions should be used by observer-moments. Almost any reasonable theory of anthropic reasoning can be interpreted as an application of the SSSA or can be interpreted to be equivalent to the SSSA and the SSSA does allow us to apply standard techniques for reasoning about probabilities. We are not saying very much if we just say that we should use the SSSA. Even in the nonanthropic case, it is of limited use to just say that we are going to apply Bayesian methodology or that we will be doing Bayesian statistics if we do not say anything about how we choose our priors; we could (in the nonanthropic case) use reference priors[1], we could try to use a universal prior[10], we could use empirical Bayes[6], or we could do something else that would allow us to choose priors in a systematic way. In the anthropic case, we also need to be precise about how we generate our priors.

The rest of this paper is organized as follows: In section 2 we introduce our centered possible worlds formalism and some notation that will make it easier to analyze wagers. In section 3 we justify a limited indifference principle[9]; if we know which world is actual, but we do not know which of two observer-moments we are and those observer-moments are in identical subjective psychological states, we should believe ourselves equally likely to be either observer-moment. This limited indifference principle is surprisingly difficult to justify. In section 4 we generalize our indifference principle to the case where the two-observer-moments are not in the same subjective psychological state but they still live in the same world. In this case, there is no question of our not knowing which of these two observer-moments we are. But we might still ask what we should believe if we do not take into account any anthropic information when estimating probability. This is a question about prior probability or about the random selection process used in the SSSA. We argue that if x and y live in the same possible world and have the same capacity for representing information and they both belong to the relevant reference class, then they should have equal probability of being chosen by the random selection process.

Section 5 analyzes exactly what is and what is not an observer and then shows how we might derive a certain popular assumption, the Self-Indication Assumption[5] if we add enough ghost observer-moments to each possible world so that all worlds have the same number of observer-moments according to some reasonable way of counting observer-moments. Section 6 discusses the SIA in more detail and explains why we do not favor the SIA.

In section 7, we discuss the issue of choosing reference classes. Once we see that if we use maximal reference classes, counterintuitive results can be derived both with and without using the SIA, we might notice that we can avoid some of our problems if we use minimal reference classes. But then, as section 7.1 points out, we will not be making full use of anthropic evidence that we actually do have. In 7.2 we show that if minimal reference classes are used, we might be vulnerible to a collective Dutch Book. 7.2 also presents several arguments in favor of the use of almost maximal reference classes. One argument is a Dutch Book argument and another argument is a relative frequency argument. None of these arguments is incontrovertible and we have not really described how to handle the infinite case. We have to see which theory of anthropic reasoning works best in practice and which method for analyzing infinite scenarios works best.

The testing of anthropic reasoning theories can be represented using standard Bayesian methodology(Section 8). But we have to be careful about $P(E|T)$, the probability that we would observe the evidence E that we actually observe given that the theory T of anthropic reasoning is correct. If the theory T is allowed to make use of knowledge of E in order to predict E, the fact that T can predict we observe E is not very surprising.

Section 9 briefly describes why the mere fact that we seem to be atypical is not a reason to say that anthropic reasoning conflicts with observation.

Section 10 relates our theory to the idea that updating is communication[16].

Section 11 discusses an application of anthropic reasoning to a puzzle (Lazy Adam) in which the decisions of one observer (Adam) determine whether other observers exist.

Section 12 contains our conclusion and summary.

2 Formalism and Notation

In this section, we first introduce our centered world formalism (2.1), then discuss when certain information might be considered irrelevant to estimating a posterior probability (2.2) and finally describe a notation for representing wagers (2.3). We might judge a theory of probabilistic reasoning by how successful rational agents are if they use the theory in order to determine how to solve decision problems; many decision problems can be represented as problems involving wagers.

2.1 The Centered Possible Worlds Formalism

We let W represent the class of possible worlds of interest to us. W, in general, will not represent all worlds, only the interesting worlds, and if $w \in W$, w might not actually be a possible world, but an equivalence class of worlds. If the differences between two possible worlds are irrelevant[3] , we consider them equivalent

[3] To say that the difference between worlds v and w is irrelevant is to say the difference is neither directly relevant or indirectly relevant. The difference would be directly relevant if we cared about whether w is more likely to be actual than v. But even

and refuse to distinguish between them.

All observer-moments of interest to us will be assumed able to reason coherently about which of the worlds in W is actual. But observer-moments do not have to be perfectly rational; some reasoning problems might be too hard for some of the observer-moments living in W. There might be some larger set $V \supset W$ of worlds such that some observer-moments living in W cannot reason coherently about which of the worlds in V is actual.

Observer-moments of interest do have to be able to do more than just reason coherently about how likely they think various worlds in W are to be actual. They also have to be able to reason coherently while using only some of the knowledge that they actually have. Thus if an observer-moment actually knows K, and $\bar{K}$ is only part of her knowledge K and $X \subset W$, she should be able to reason as if she only knew $\bar{K}$ and reason coherently about how likely it is that the actual world lies in the set X. We might not impose this requirement for all possible $\bar{K}$, but we should require it for certain important $\bar{K}$. One important knowledge set is the set of propositions known to be true by all observer-moments who live in worlds in W.

But it is not just (bare) possible worlds that are of interest to us. We are really more concerned with centered worlds[14]. A centered world is just an ordered pair consisting of a (bare) world w and a center c. In general, many different kinds of center are possible, but in the centered worlds (w, c) that we analyze, c will be an observer-moment that inhabits world w.

We let $W*$ represent the class of centered possible worlds of interest to us. We assume that every $z \in W*$ is obtained by enriching a world $w \in W$ with a center c so that $z = (w, c)$ and that for every $w \in W$, there exists at least one c such that $(w, c) \in W*$. All $z \in W*$ are assumed to know[4] that they belong to $W*$ and to be able to reason probabilistically about who they are among the elements of $W*$. They are assumed to know how to apply the SSSA.

if the difference is not directly relevant, it might be indirectly relevant. If we are suffering from amnesia and want to know whether we are named George or Bill, we might not really be very interested in whether we live in a world v in which there are more people named George than Bill or a world w in which more are named Bill than George except for the fact that learning v rather than w is actual might tend to cause us to increase our estimate for how likely it is that we are named George. A precise definition of irrelevance will be provided in section 2.2.

[4] When we refer to $z \in W*$ knowing or believing or getting utility from something, what we mean is the following: z is an ordered pair (w, c) where c is an observer-moment in world w. So we should be refering to the knowledge or beliefs or utility of c in w when we refer to something being known or believed by or the utility of z.

A problem arises when we refer to what z believes or knows or to the utility of z; z might change her beliefs or desires. She might acquire knowledge or forget. Thus we cannot necessarily say about some proposition p that z believes that p or z does not believe that p. Both might be true but at different times. We also have the problem of in-between-believing[20]. At certain points in time, there might be no fact of the matter as to whether z believes that p. From some perspectives, she might be said to believe and from others, she might be said not to believe. We shall temporarily ignore these problems and assume that z has definite beliefs, desires, and knowledge.

We also have the problem alluded to by Weatherson[21] of vague belief states. We can either just assume that observer-moments have crisp and not vague beliefs or we can say that if it seems that it is vague whether some observer-moment z living in world $w \in W$ believes p and does not believe q or vice versa, what is really happening is that w is really two different worlds in one of which z believes p and in the other of which z believes q.

We assume that all $z \in W*$ agree on a nonanthropic prior P. If $w \in W$, then $P(w)$ represents the prior probability that world w is actual. It would be more precise to say that $P(w)$ represents the conditional probability that w is actual given that the actual world lies in W (and not taking into account any knowledge that one observer-moment in $W*$ has and another does not have). Observer-moments might give nonzero prior probability to the possibility that some world outside of W is actual, but they might be uninterested in such a possibility or have difficulty reasoning about worlds outside of W. Thus strictly speaking observer-moments in $W*$ do not necessarily know they belong to $W*$, but they reason as if they knew they belonged to $W*$.

In order to analyze the set W, it is helpful to study $W*$; in order to analyze some subset $V \subset W$, it will be helpful to study a certain set $V*$ of centered possible worlds. We shall generalize the star notation so that if $V \subset W$, $z \in V*$ if and only if there exists $v \in V$ such that there is a center c with $z = (v, c) \in W*$. Thus $V*$ is the set of centered worlds in $W*$ that can be obtained by enriching a world in V with a center. In the case where $V = \{v\}$ is a singleton set, we write $v*$ and not $\{v\}*$.

The star notation lets us go from bare worlds to centered worlds; in order to go in the reverse direction, if $Z \subset W*$, we define $\hat{Z} \subset W$ as the set of $w \in W$ such that there exists c with $(w, c) \in Z$. If we think of (w, c) as part of w, then $\hat{Z}$ might be thought of as the set of bare worlds which contain elements of Z. If $Z = \{z\}$ is a singleton set, we write $\hat{z}$ rather than $\hat{Z}$. In that case, $\hat{z}$ is a singleton set consisting of just a single world; we shall be careless about distinguishing between the singleton set and the one world it contains.

What a $z \in W*$ needs to do is the following: Given an interesting subset $V \subset W*$, compute the posterior probability that she belongs to V. First she needs a prior probability $P_z(V \cap R_z)$ where $R_z \subset W*$ is the reference class used by z. The observer-moment is reasoning as if she were a random element of R_z and the prior distribution P_z used to do the random selection should satisfy a certain constraint that relates the nonanthropic prior P to the anthropic prior P_z. We require for all $A \subset W$ that $P_z(A * \cap R_z) = P(A|\hat{R_z})$. We want the probability $P(A|\hat{R_z})$ to be obtainable[5] by summing the probabilities of the observer-moments (in the reference class) that belong to worlds in A.

Once z has a prior probability for V, she can compute her posterior $P_z(V \cap R_z|K_z)$ where K_z represents what z knows. The set K_z is a set such that $y \in K_z$ if and only if z would say to herself that for all she knows she might be y.

[5] when A contains only a finite number of observer-moments

2.2 Irrelevance

We know that $W*$ might not actually represent possible centered worlds but instead represent equivalence classes of centered worlds and that if the difference between two worlds is irrelevant we might consider them equivalent. We should be more precise about what irrelevance means.

So let us start with an actual set W of (bare) possible worlds and an actual set $W*$ of centered worlds rather than equivalence classes of worlds and centered worlds. We consider some observer-moment $z \in W*$ who wants to estimate $P_z(A|K_z)$ for some $A \subset W*$ and ask ourselves when it is acceptable for z to consider certain (bare) worlds and consider certain centered worlds equivalent.

Let $\equiv *$ represent an equivalence relation on $W*$. Let $\equiv$ represent an equivalence relation on W. We would normally expect there to exist some relationship between these two equivalence relations. If an observer-moment i can exist in two different possible worlds, v and w, we might consider (v, i) and (w, i) to be counterparts of each other and then if $v \equiv w$, we would expect $(v, i) \equiv *(w, i)$; more generally we might define a natural counterpart relationship between observer-moments living in different (bare) worlds and if i living in v is a counterpart of j living in w and $v \equiv w$, we would expect $(i, v) \equiv *(j, w)$.

If $V \subset W$, then we write $V_\equiv$ to represent the closure of V under $\equiv$. Thus if w is a (bare) world, $w \in V_\equiv$ if and only if there exists a $v \in V$ with $v \equiv w$. If $Z \subset W*$, we write $Z_{\equiv *}$ to represent the closure of Z under $\equiv *$. Thus $z \in Z_{\equiv *}$ if and only if there exists $y \in Z$ with $y \equiv *z$. If C is any set and R is an equivence relation on C, then we can generate the set G of equivalence classes of C modulo R. We have $g \in G$ if and only if there exists a $c \in C$ such that g is the set of elements $d \in C$ with $d R c$. If $B \subset C$, then $B \mod R$ refers to the set of $g \in G$ with $g \cap B \neq \emptyset$.

We shall make statements about closures $Z_{\equiv *}$ and $V_\equiv$, but it should be easy to translate these statements into statements about equivalence classes $Z \mod \equiv *$ and $V \mod \equiv$. For example, in order to evaluate a probability for $Z \mod \equiv *$, it suffices to evaluate the probability of $Z_{\equiv *}$. It just simplifies the exposition to use closures rather than equivalence classes.

In order to work with closures, we need to insure that P_z is defined on all of $W*$ not just on R_z, but that is simple enough if we just specify that for all $X \subset W*$ with $X \cap R_z = \emptyset$, $P_z(X) = 0$. We might assume that the set A whose posterior probability z wants to estimate is closed under $\equiv *$. Then what z really cares about is whether

$$P_z(A|K_z) = P_z(A|(K_z)_{\equiv *})$$

and thus we care whether

$$\frac{P_z(A \cap K_z)}{P_z(K_z)} = \frac{P_z(A \cap (K_z)_{\equiv *})}{P_z((K_z)_{\equiv *})}.$$

Assuming that $P_z(A|K_z) \neq 0$, that means that we care about whether

$$\frac{P_z(K_z)}{P_z((K_z)_{\equiv *})} = \frac{P_z(A \cap K_z)}{P_z(A \cap (K_z)_{\equiv *})}.$$

In words, this is requiring that the probability that a random element of the closure $(K_z)_{\equiv *}$ of the knowledge set is actually an element of the knowledge set K_z be independent of whether that random element is also an element of A. If this condition is true, than z can refuse to distinguish between equivalent bare worlds and equivalent centered worlds when estimating the probability of A.

Of course in the previous paragraph, z is using the fact that the closure of her knowledge state is $(K_z)_{\equiv *}$ and there might exist an observer-moment $y \equiv *z$ such that $(K_y)_{\equiv *} \neq (K_z)_{\equiv *}$; in that case the difference between equivalent observer-moments is not totally irrelevant even if it is basically irrelevant for z who only cares about the posterior probability of A. If the only subsets of $W*$ about which z cares are sets B that are closed under $\equiv *$ and the differences between equivalent uncentered worlds and equivalent centered worlds is basically irrelevant with respect to any such set B and for all $z \equiv *y$, $(K_z)_{\equiv *} = (K_y)_{\equiv *}$, then from z's point of view, the differences between equivalent worlds and equivalent centered worlds is totally irrelevant.

2.3 Wagers

We also need some special notation to represent wagers. A wager is represented by a function f from $W*$ to $\mathcal{R}$ where $\mathcal{R}$ represents the real numbers. If $z \in W*$, $f(z)$ is the amount of utility that z gains by accepting rather than rejecting the wager. We assume that the amount gained by z is independent of how other $y \in W*$ decide when offered the wager f and also independent of any other wagers that might be offered. We let f have domain all of $W*$ so that no useful information about one's location within $W*$ can be gained from the fact that one is offered a certain wager. The return from a wager is a utility value rather than a monetary value because utility is not necessarily linear in monetary return. We assume all $z \in W*$ try to maximize their expected utility $\sum_{y \in K_z} P_z(y)f(y)$[6].

We shall have occasion to add the utilities of two y and z such that $K_y \neq K_z$. If $K_y = K_z$, then z might have to compute an expected utility and thus might have to be able to compare y utility and z utility. She

[6] We are assuming that there is no overlap: If $x \neq y$ are two observer-moments in $W*$, then it is not possible to be both x and y. Thus if x and y represent time-slices of the same observer in the same possible world, the time-intervals during which x and y live are disjoint.

will have to be able to measure y utility and z utility on a common scale so that one unit of y utility is as valuable as one unit of z utility. But if $K_y \neq K_z$, no one actually needs to measure y utility and z utility on a common scale in order to decide what to do. But we can ask how an observer-moment would make decisions if she did not know whether she was y or z but knew that she had a certain probability of being y and a certain probability of being z. This really means that she has a certainly probability of being someone just like y and a certain probability of being someone just like z[7] because anyone who actually might be y knows she is not z and vice versa. But in any case it should be possible to perform an act of imagination and imagine that one does not know whether one is y or z and then try to figure out how one would make decisions when not sure whether one is y or z. Even if it really does not make sense to compare y utility and z utility if $K_y \neq K_z$, it is not impossible that there might be some preferred way to scale the utilities of y and z so that one unit of y utility and one unit of z utility are equally valuable. Theories of anthropic reasoning should work if it turns out that it makes sense to add the utilities of observer-moments y and z with $K_y \neq K_z$ and it will make sense to add the utilities if it is possible to measure the utilities of y and z using a common scale.

3 A Limited Indifference Principle

In this section, we argue for a limited indifference principle that if $x, y \in W*$ with $\hat{x} = \hat{y}$ and $K_x = K_y$, then $P_x(x) = P_x(y)$[9][8]. This is a very limited principle. It only says that if there are two observer-moments who live in the same possible world and who are in the same subjective psychogical state and for all we know we might be one of these observer-moments, then we are no more likely to be one of them than to be the other one.

We shall present several possible arguments for this principle. Some of these arguments are more convincing than others but we do need a principle that will enable us to compute $\frac{P_x(x)}{P_x(y)}$.

When trying to construct an argument for our limited indifference principle, we must keep in mind that the argument should not be too easily generalizable. We do not expect a human observer-moment that lasts ten million seconds to have the same prior probability as one that lasts one second; these two observer-moments will not be in the same (relevant) subjective psychological state (or sequence of states).

We also need to be very clear about how limited our principle is. It is certainly not true that if $K_x = K_y$ and x and y inhabit different worlds that they necessarily have the same prior probability. Nor do we have

[7] Someone who has preferences and experiences that are very similar to the preferences and experiences of z.

[8] Elga's[9] discussion refers to observers, rather than observer-moments, but his principle is in essence the same as ours.

the principle that if x, y, z live in the same world and are in the same subjective psychological state, then $P_z(\{x, y\}) = 2P_z(z)$. The problem is that there might be overlap. In the worst case x and y might be the same observer-moment. Or they might share some of their computational hardware. Or x and y might not be making independent observations of their environment so there might be a sense in which if I am x, then I am also in part y. This happens with human observer-moments if x and y are observer-moments that belong to the same observer[9] and y occurs immediately after x and thus regardless of how radically and how quickly the external environment might be changing, it takes time to correct the obsolete information y has received from x.

The first argument for the limited indifference principle is that it is a simple principle that is not obviously absurd and it is hard to think of another principle that is equally simple and that is workable. If $K_x = K_y$ and $\hat{x} = \hat{y}$ and $\frac{P_x(x)}{P_x(y)} \neq 1$, what is the ratio to be? We might say that $P_x(y)$ and $P_x(x)$ should depend on the lengths of the briefest descriptions of y and x in some canonical language. But then we have the added complexity of discovering the ideal canonical language.

The problem with this argument is that it is too easily generalizable to the case where x and y live in different possible worlds. We certainly do not want to say that if $K_x = K_y$ with $\hat{x} = v \neq w = \hat{y}$ and w is a world in which with a few exceptions every observer is a brain in a vat but v is a world more like the actual world, that $P_x(x)$ should equal $P_x(y)$. We are not all that likely to be a brain in a vat.

The second argument is a wagering argument. We can without loss of generality restrict to the case where both x and y know they both belong to a certain world w and they both know they are either x or y but they do not know which one and we assume there is no overlap. If there were overlap, we could just find some reasonable way of modifying our P_x so that $P_x(x|x \text{ or } y) + P_x(y|x \text{ or } y) = 1$

Because $K_x = K_y$, we must have $P_x = P_y$. Assume that $P_x(y) = rP_x(x)$ with $r > 1$. So $P_y(y) = rP_y(x)$. If there is a wager f such that $f(y) = 1$ and $f(x) = -s$ with $1 < s < r$ while $f(z) = 0$ unless z is either x or y, then both x and y will compute an expected value of $\frac{r}{r+1} - \frac{s}{r+1} > 0$ and accept the wager. But then it is inevitable that the total return is $1 - s < 0$. So it seems it would have been better if both x and y had rejected the offer. And since x and y are in the same subjective psychological state, they would either both accept or both reject. Thus the best option is for them to both reject and that appears to show that if $P_x(y) = rP_x(x)$, then r should not be greater than 1. A very similar argument would show that r should not be less than 1. Thus $P_x(y) = P_x(x)$.

This wagering argument is too similar to an argument that in the Prisoners dilemma scenario it is rational

[9] We say that $x, y \in W*$ belong to the same observer if they are temporal slices of the same observer in the same possible world.

for both prisoners to cooperate because it is better that they both cooperate than that they both defect. However, each individual prisoner is better off if she defects regardless of what the other prisoner does. The mere fact that one prisoner cooperates does not force the other to cooperate.

In our scenario both x and y can say: "I am much more likely to be y than x. So it makes sense for me to accept a wager f with $f(y) = 1$ and $f(x) = -2$. I know the other guy (x if I am y and y if I am x) thinks the same thing but she is mistaken. It is not desirable that the other guy accept, but that should not stop me from accepting because I am more likely to be y."

It is certainly consistent to say that x and y should have equal prior probability but it is also consistent to say that y should have ten times as much prior probability. Yet it seems intuitively reasonable if we have to choose between two possible solutions to the problem of choosing priors that we prefer the one that results in greater total utility and greater average utility. This argument about what is intuitively reasonable does not generalize to the case where x and y live in different worlds; we really only care about the total or average utility of the observer-moments who actually exist. If x and y live in the same world we can assume that we have scaled the utility functions of x and y in such a way that it makes sense to compute a simple sum or average of the utilities of x and y. If $\hat{x} \neq \hat{y}$, then we would be more interested in a weighted sum or average of the utilities of x and y where the weights of x and y are proportional to $P(\hat{x})$ and $P(\hat{y})$ respectively.

Our third argument is that there is a sense in which it is not consistent to say that y should have greater (for example, three times as much) prior probability, than x. If I am x (or y) and I say that "The other guy has only a probability of $\frac{1}{4}$ of being y even though the other guy is in the same subjective psychological state as I am and I believe I have a probability of $\frac{3}{4}$ of being y", there is no obvious justification for the difference between what I believe is my probability of being y and what I believe is the other guy's probability. The third argument does not generalize to the case where only one of the observer-moments x, y exists. In that case, there would be one big difference between me and the other guy: I exist and she doesn't.

The fourth argument is a simplified version of an argument of Elga's[9]. The argument involves comparing three different scenarios and is based on the assumption that similar scenarios should be analyzed similarly. The argument is not quite convincing but it is still worth analyzing.

In the first scenario, there is only one possible world and in that world, there are only two observers, Al and AlDup (a duplicate of Al). Originally, there was just one observer, but then at a certain point t_0 in time, a duplicate of Al appeared. At any point in time subsequent to t_0, Al and AlDup are in the exact same subjective psychological state. At some time t_1 after t_0, Al wants to estimate the probability that he is Al rather than AlDup. (Of course, AlDup also wants to estimate the probability that he is Al.) We might analyze the scenario as containing two relevant observer-moments, Al at time t_1[10] and AlDup at time t_1[11]. We would like to show that Al at time t_1 should believe the probability that he is Al to be .5. If we could demonstrate this, it should not be too hard to show that our limited indifference principle that $P_x(x) = P_x(y)$ if $K_x = K_y$ and $\hat{x} = \hat{y}$ should be true in general.

In order to help us analyze this first scenario, we consider a second scenario in which there are two possible worlds, H and T. These worlds differ only in how an unfair coin tossed by a robot lands. If the coin lands heads, H is actual (probability .1) and if the coin lands tails, T is actual (probability .9). In both worlds Al and AlDup are the only observers and in both worlds at any time after time t_0, Al and AlDup are in the exact same subjective psychological state. The duplication and coin-tossing are assumed to be completely independent processes. The coin-tossing does not affect the subjective psychological state of either Al or AlDup. We might analyze this scenario as containing four relevant observer-moment Hal (Al at time t_1 in the heads world), Tal (Al at time t_1 in the tails world), HalDup (AlDup at time t_1 in the heads world) and TalDup (AlDup at time t_1 in the tails world).

Since the coin-tossing is entirely independent of the duplication, in order to show that in the first scenario Al should believe the probability he is Al equals .5, it suffices to show that in the second scenario Hal (as well as Tal, HalDup, and TalDup) should believe that the probability that he is Al (i.e. that he is either Hal or Tal) to be .5.

We consider a third scenario, that is exactly like the second scenario except that at a certain time t_2 later than t_1 one of the two observers goes into a coma. In H, it is AlDup who goes into a coma and in T, it is Al. At some time t_3 later than t_2, Al in the world H wants to estimate the probability that he is Al. Al at time t_3 in H is in the same subjective psychological state as AlDup at time t_3 in T. So AlDup at t_3 in T also wants to estimate the probability that he is Al.

We might analyze the third scenario as containing six relevant observer-moments. We have the same four relevant observer-moments that we had in the second scenario: Hal, Tal, HalDup, and TalDup. These are all in the same subjective psychological state. We also have Hal2 (Al at time t_3 in H) and TalDup2 (AlDup at time t_3 in T). Hal2 and TalDup2 are in the same subjective psychological state but their subjective psychological state is different than the state of Hal, Tal, HalDup, and TalDup.

If we assume that Hal2 uses $\{\text{Hal2}, \text{TalDup2}\}$ as his reference class, clearly Hal2 should believe that the probability that he is Hal2 is .1. The other reasonable choice of reference class is the class consisting

[10] t_1 might be a short time-interval rather than just a point in time.

[11] The reason we might consider these two observer-moments to be the only relevant observer-moments is that nothing essential changes if we assume that both Al and AlDup are unconscious except at t_1.

of all six observer-moments. Because the coin tossing is assumed irrelevant, we would like to assert that $P_{\text{Hal2}}(\text{Hal2}|H) = P_{\text{Hal2}}(\text{TalDup2}|T)$ because that would allow us to conclude that even if the larger reference class is used, Hal2's posterior probability estimate for his being Hal2 should still be .1. However, unfortunately, it is perfectly consistent to give Hal2 and Hal twice the prior probability of HalDup and give Tal twice the prior probability of TalDup and TalDup2. (Regardless of how the coin falls originals have twice the prior probability of duplicates.)

Let us just assume that Hal2 should conclude that the posterior probability of H is .1. If we could assume that $P_{\text{Hal}}(\text{Hal}|\text{Hal or TalDup}) = P_{\text{Hal2}}(\text{Hal2}|\text{Hal2 or TalDup2}) = .1$, simple algebra would show that $P_{\text{Hal}}(\text{Hal}|\text{Hal or HalDup}) = .5$. And then given the assumption that similar scenarios should have similar analyses, we would have the result that in the first scenario, Al should believe that he is as likely to be Al as AlDup. Unfortunately, Hal and Hal2 are different and it need not be the case that $P_{\text{Hal}}(\text{Hal}|\text{Hal or TalDup}) = P_{\text{Hal2}}(\text{Hal2}|\text{Hal2 or TalDup2})$. It would be the case if a certain continuity assumption is true, but it is not clear that the continuity assumption in question is any more obvious and any less in need of proof than our limited indifference principle.

Assuming that our four arguments demonstrate the truth of the indifference principle that if $K_x = K_y$ with $\hat{x} = \hat{y}$, then $P_x(x) = P_y(x)$, we find it natural to believe that if $K_x = K_y$ with $\hat{x} = \hat{y}$, then $P_z(x) = P_z(y)$, for any $z \in W*$ such that $x, y \in R_z$. A general argument for this result is that $\frac{P_z(x)}{P_z(x)+P_z(y)}$ is z's estimate of how likely she should think it is that she is x given that she knows that she is either x or y[12]. But someone who knows that she is either x or y would be in knowledge state K_x and thus would believe it as likely that she be x as that she be y. Therefore, we should have $\frac{P_z(x)}{P_z(x)+P_z(y)} = .5$ and $P_z(x) = P_z(y)$.

We would also like to say something about the ratio of the prior probabilities of x and y when x and y live in the same possible world but are not in the same subjective psychological state.

4 Assuming $\hat{x} = \hat{y}$ what should be the value of $\frac{P_z(y)}{P_z(x)}$ when y, $x \in R_z$?

If x and y are in different knowledge states ($K_x \neq K_y$), then there is no question of an observer-moment not knowing whether she is x or she is y, but the ratio $\frac{P_z(y)}{P_z(x)}$ still matters. To see why, consider the following simple scenario.

The Fundamental Scenario:

There are only two possible worlds, v and w. If they ignore anthropic information, the observer-

[12] We are still assuming that overlap is not a problem.

> moments in v and w would have no reason to think one of these worlds more likely to be actual than the other world. But observer-moments do have anthropic information available. In both worlds, there are exactly two possible subjective psychological states, A and B. We can represent v as containing c observer-moments in state A and d observer-moments in state B. The world w consists of e observer-moments in state A and f in state B. The numbers c, d, e, f are all finite. We have no problem with overlap. All the different observer-moments in each world are genuinely distinct observer-moments and thus for example, the c observer-moments in world v who are in state A are all distinct and do not share resources or inhibit each other's capacity for believing and desiring and accepting or rejecting offers to wager. Every observer-moment needs to estimate the probability the actual world is v taking into account the anthropic information she actually does have.

We call this scenario fundamental because if we know how to analyze this scenario, we know how to analyze most scenarios that can be represented as having only a finite number of possible worlds with observers and in which every world has no more than a finite number of observer-moments.

In this scenario, all observer-moments are assumed to know all the details of the scenario and to know which subjective psychological state they are in, but in general, they do not know which world they inhabit and have only limited information about their identity and temporal location: They just know that they are in state A or they know they are in state B.

Assume all observer-moments in the same possible world have equal prior probability and let z be an observer-moment who is in state A. We assume that z uses a reference class consisting of all $c + d + e + f$ observer-moments. Not taking into account her anthropic information, z would say that the prior probability that both v is actual and she is in state A is $P_z(A \text{ and } v) = (\frac{c}{c+d})(\frac{1}{2})$. The prior probability that she is in state A and that w is actual is $(\frac{e}{e+f})(\frac{1}{2})$. After taking into account her knowledge that she is in state A, we obtain a posterior probability of $\frac{\frac{c}{c+d}}{\frac{c}{c+d}+\frac{e}{e+f}}$ for v being actual. This simplifies to $\frac{c(e+f)}{c(e+f)+e(c+d)}$. If, however, z had assumed that observer-moments in state A had many times more prior probability than observer-moments in state B, she would arrive at a very different posterior probability. If she assumed that observer-moments in state A had infinitely many times as much prior probability as those in state B, she would arrive at a posterior probability of .5.

A natural hypothesis about $\frac{P_z(x)}{P_z(y)}$ in the case when $\hat{x} = \hat{y}$ and $x, y \in R_z$ is that prior probability should be proportional to cognitive complexity. This is implicitly assumed in, for example, [2]. Since we want $P_z(x) = P_z(y)$ if $K_x = K_y$ and $\hat{x} = \hat{y}$, we shall assume that observer-moments in the same subjective psychological state have the same amount of cognitive complexity. We do not have to make the assumption, but if we did not make the assumption, instead of discussing the cognitive complexity $I(x)$ of a certain observer-moment x, we would have to discuss the average complexity of an observer-moment in world $\hat{x}$ and subjective psychological state K_x. So instead of saying that $P_z(x) = P(\hat{x})\frac{I(x)}{I(\hat{x})}$ where $I(x)$ represents the complexity of x and $I(\hat{x})$ represents the cognitive complexity of the whole set of observer-moments living in the same world as x and we assume that the total complexity is finite, we could use the same formula but $I(x)$ would have to represent the average complexity of the observer-moments in $\hat{x} \cap K_x$. Our arguments and our exposition can be simplified if we assume that if $K_x = K_y$, then $I(x) = I(y)$.

One measure of the cognitive complexity of an observer-moment z is the amount of (relevant) information $I(z)$ that she is capable of representing. This suggests that we should have $\frac{P_z(x)}{P_z(y)} = \frac{I(x)}{I(y)}$ in the case where $I(x)$ and $I(y)$ are both finite.

The basic reason is that we believe our information theoretic rule to be correct is that it is simple and intuitively appealing. We also have a wagering argument for our rule and an argument based on the concept of indecomposable or atomic moment and the idea that we should be able to represent $W*$ as a union of independent atomic moments.

We first give our wagering argument. We assume that W and $W*$ are really equivalence classes of worlds and centered worlds and that if we just specify $z \in W*$, we have not specified certain potentially important information about z; we have not specified which decision problems[13] z needs to solve. Maybe she does not need to explicitly represent which problems she is trying to solve but she does need to devote cognitive resources to solving these problems. Our key assumption is that the total amount of cognitive resources that an observer-moment can devote to solving decision problems is proportional to the amount of information she is capable of representing about who she is among the observer-moments in $W*$. Or we might just assume that the amount of information that z can store about which problems she is trying to solve is proportional to the amount of information she is capable of representing about who she is among the elements of $W*$[14].

In any case, we are assuming that computational resources that can be used to solve decision problems are scarce. In order to solve a certain decision problem, an observer-moment might need to make use of her posterior probability estimate for how likely it is that she belongs to some set $Z \subset W*$. Even if $P_z(Z|K_z)$ is trivial to compute, some computational resources will be spent when z computes the probability and then uses the probability to optimize her decision-making.

[13] Wagering problems are decision problem. We might also regard probability estimation problems as decision problems in which an observer-moment is trying to maximize some kind of epistemic utility.

[14] Thus we are treating information storage space as the only scare resource.

Given that scarcity of computational resources exists, we will explain why the scarcity matters. We will use a wagering argument. Not every decision problem is a wagering problem but our argument can be generalized to apply to decision problems that are not wagering problems. We are interested in P_z, a prior that supposedly does not take into account any information that z has and other observer-moments in $W*$ do not have. Thus we shall not take into account which wagering problem z is trying to solve. We just assume it is some random wagering problem. We assume because computational resources are scarce and the fact that z is a rational agent is only demonstrated when z is trying to solve some decision problem and because there are so many possible decision problems that she might need to solve, that the probability that z will have the computational resources available to solve a given randomly chosen decision problem is small[15] and we assume the probability is proportional to the amount of information she is capable of representing about her location among the observer-moments in $W*$. This assumption is most reasonable for simple, easy to describe decision problems but more complex problems can be represented as a sequence of simpler problems.

We want the observer-moments in $W*$ to choose their prior probabilities in such a way as to optimize expected utility when faced with a random wagering problem D. When computing this expected utility, we will ignore those observer-moments who do not have the resources available to solve the wagering problem. If we could restrict to the case where in each $w \in W$, there is at most one $y \in w*$ who has the resources to solve D, then the probability that a given z will actually be trying to solve D is the product of $P(\hat{z})$, the probability that z lives in the actual world, and $\frac{I(z)}{I(\hat{z})}$. And that would be a reason for prior probabilities to be proportional to amount of information represented.

In reality, there might be some worlds where many observer-moments have the resources to solve D, but we might pretend that D comes in several different variants. There is no essential difference between the variants, between the different ways that a decision problem might be formulated. We might describe our situation as one in which all an observer-moment knows is that she is dealing with a random variant of a random decision problem. If there are enough variants, then it is quite likely to be true that in each world only at most one observer-moment will have the resources available to solve a specific random variant of D. And we have to analyze each variant differently since each variant is a different problem and for each variant there is a different set of observer-moments who have the available cognitive resources to deal with the variant. Introducing these imaginary variants should not affect which prior P_z should be used. But with the help of these variants we could see why prior probability should be proportional to cognitive complexity.

[15] We are assuming that which other decision problems need to be solved by z is determined by some stochastic process.

This argument might seem to similar to an argument that in the fundamental scenario if $c = e = 1$ (in both v and w, there is one observer-moment in state A) and $f = 10^{100}$, then an observer-moment z in state A should believe the two possible worlds equally probable. But this argument for $P_z(v) = P_z(w)$ would be ignoring the fact that observer-moments in state B exist in world w.

In our wagering argument, we treat observer-moments who do not have the resources to handle a random problem as if they were not observer-moments. We are treating them as if they do not exist[16]. We are only using our wagering argument to determine a prior. We need some way of constraining our prior. We are following the general philosophy of making our prior as uninformed as possible, taking into account as little as possible. If we actually have more information, we can conditionalize. Thus we will have wagering arguments constraining what P_z is, but we will not try to provide a wagering argument to constrain the posterior $P_z(\ |K_z)$.

We also have another justification of our information-theoretic rule based on indecomposability (i.e. atomicity).

To understand why we care about indecomposability, reflect about the fact that some observer-moments live too long to be analyzed as unified observer-moments and should really be decomposed into sets of shorter-lived observer-moments. There is a sense in which it is difficult to conceive of a human observer-moment z that lasts ten thousand seconds as having definite beliefs or making definite decisions or being in a definite subjective psychological state. In ten thousand seconds, the external environment can change drastically. There can be justifiable drastic changes of relevant belief during those ten thousand seconds. It might be quite misleading to reason about z as if she had a definite knowledge state K_z. Wagering or decision-theoretic arguments based on the assumption that observer-moments have definite beliefs and make definite decisions based on those beliefs and a computation of the action that leads to greatest expected utility might not really be applicable (even approximately) to an observer-moment like z. A wager offered to z during the last three hundred seconds of her life might be something about which she was ignorant for most of her life. There is a sense in which we should not try to model z as if she were a single rational agent with definite desires and beliefs.

It might be true just by happenstance that the long-lived observer-moment z does have definite desires and beliefs, but if z lives long enough, it could easily be the case that one part of z is exposed to different evidence than another part of z and thus different parts of z have different beliefs. If it makes sense to split

[16] This is very different from saying we want to compute the posterior probability of some specific set A and then ignoring observer-moments who do not need to know this probability. When we know that we need to know the probability of A and other observer-moments do not need to know this probability, we know some information that is not common knowledge to all $z \in W*$. But we can picture all observer-moments as knowing that they are trying to solve some random decision problem.

z into a set of smaller observer-moments and these smaller observer-moments have the capacity to disagree about some relevant topic, then we cannot consider z to be indecomposable.

But if z is sufficiently short-lived, it might not be possible to meaningfully view z as a union of two independent observer-moments z_1 and z_2 and thus we might not be able to split up z any further. We might be able to represent z as a union of z_1 and z_2, but it might be the case that z_1 and z_2 are human observer-moments who only live one ten-thousandth of a second and who belong to the same observer in the same world with z_2 starting her existence at the exact time z_1 ceases to exist. In this situation, there is no meaningful sense in which we could really think of z_1 and z_2 as being completely distinct agents who are free to make their own decisions, arrive at their own beliefs, and independently observe their environment. The observer-moment z_2 just does not have any time to respond to new evidence that she has access to and z_1 does not have access to. It takes time to process information and formulate new beliefs and make new decisions. We cannot reasonably represent z_2 and z_1 as independent rational agents.

Of course, there is always some dependence between two observer-moments y and x that belong to the same human observer. But it can be convenient to model $W*$ as consisting of a set of independent, indecomposable observer-moments. This modelling is an idealization, but sometimes it can be a useful approximate description.

We have explained what indecomposability means; we need to say a little more about what independence means. If y is just another name for z, then certainly y is not independent of z. But if z and y live in the same world and have virtually identical beliefs and would make virtually identical decisions if faced with the same decision problems, that does not necessarily mean that z and y are extremely dependent. The similarity between y and z might just arise because y and z are exposed to very similar external environments. If y and z were exposed to very different environments, they would have different beliefs; in that case y and z could be quite independent. But if the very fact that y has a certain belief or chooses to accept or reject a certain offer to wager forces it to be the case that z has the same belief or makes the same decision about accepting or rejecting the wager, then we have substantial dependence. There could also be complete dependence between y and z even if K_y and K_z are very different: There would be complete dependence if once we know K_y, we could predict what K_z would be without knowledge of the environment to which z has been exposed. That means that even if z were exposed to a very different environment than y and even if that environment were very different from the environment that any observer-moment in $W*$ is exposed to, we could still predict K_z knowing just K_y.

A few more worlds about independence. The test for independence involves counterfactuals, but these counterfactuals should not involve worlds that are too different from the actual world. We are interested, for example, in what human observer-moments who last only one thousandth of a second actually are capable of; we are not interested in what they could do if they could think ten thousand times more quickly than they actually can. In the case of two observer-moments y and z both of whom live for .0001 seconds and one of whom begins life as soon as the other dies, we are interested in the actual relationship between the cognitive capacities of y and z, we are interested in the fact that they actually do share hardware because thinking and learning takes time and we are not interested in some alternative world where y and z do not share hardware.

Another point is that independence is really a relationship involving sets of observer-moments. The observer-moment z might be capable of representing much information that is not represented by x and capable of representing much information not represented by y, but if one knows both K_x and K_y, one might be able to accurately predict K_z without knowing the evidence to which z has been exposed. In this case z is highly dependent on the set of observer-moments $\{x, y\}$.

A further point is that whether z is independent of $X \subset \hat{z}$ might be determined by a stochastic process. Some random process A in the brain of z might determine whether we have independence. Technically, if A can have different possible results I and II, then $\hat{z}$ is not really just one possible world, but an equivalence class of worlds: A world in which A has a result I is different from one in which it has result II, but we are not distinguishing between these two different worlds.

The stochastic element that determines whether z is independent of $X \subset \hat{z}$ might not be a random process A in the brain of z; it might be some other element of the environment that is incompletely modelled when we choose to use W as the set of all possible worlds. For example, W might not fully take into account which questions observer-moments in W find it interesting to answer.

There can be much more said about the nature of independence, but at this point we shall just assume that it can be reasonable to regard $W*$ as a set of independent, indecomposable (i.e. atomic) moments and we assume that if two atomic moments belong to the same possible world, they should be given the same prior probability. The rationale for this is that prior probability should represent what we would believe if we did not take into account any knowledge we might have that is not also known to every other observer-moment in $W*$. If we do not take into account any such knowledge, then we can treat atomic moments x and y as if they were in the same knowledge state (as if $K_x = K_y$). Then if $\hat{x} = \hat{y}$, we would apply our limited indifference principle and conclude that x and y should have equal prior probability. Our argument should not be generalized to nonatomic moments because nonatomic moments do not necessarily have definite knowledge states.

Our assumption about atomic moments can be translated into an information-theoretic criterion. We necessarily have for all $z \in W*$ that $z \in K_z$. But aside from having to satisfy this one simple constraint, the knowledge state of z can be any subset of $W*$ independent of the knowledge states of all other elements of $W*$. Thus each atomic observer-moment can represent $|W*|-1$ bits of information where $|W*|$ represents the number of elements in $W*$. And in fact our information theoretic criterion is verified: $\frac{P_z(x)}{P_z(y)} = \frac{I(x)}{I(y)}$ if $|W*|$ is finite. This criterion would also be verified for observer-moments x and y that could be represented as finite unions of atomic moments.

The version of the atomic moments procedure that we have just described is unrealistic in several different ways, but it still might be useful to reason as if it were true. One primary reason that modelling $W*$ as consisting of independent atomic moments is unrealistic is that in any realistic scenario in which all the possibly relevant details are taken into account, $|W*|$ is huge. It is not reasonable to conceive of the $z \in W*$ as independently representing information: $|W*|$ is too many bits for one indecomposable rational agent to know. If we work with huge equivalence classes of observer-moments so that each $z \in W*$ is really a huge equivalence class and $|W*|$ is very small, then the idea that $W*$ consists of independent atomic moments is more plausible.

Another reason that our modelling of $W*$ as consisting of independent atomic moments might be considered unrealistic is that it takes time for an observer-moment to make inferences and while in the process of making inferences, an observer changes her subjective psychological state. We make the idealization that observers can take as much time as they want to apply the rules of logic and probability theory to to their knowledge about who they are among the observer-moments in $W*$. All an observer-moment $z \in W*$ has to do is specify which subset $K_z \subset W*$ represents her state of knowledge; it might be some other observer-moment who makes inferences based on that state of knowledge. Or we might model our observer-moments as being logically omniscient and able to instaneously derive any needed conclusions[17]. In nonanthropic Bayesian reasoning, we also presuppose a certain amount of logical omniscience so it should be no surprise that we might need to model observer-moments as being perfect at logic.

We now describe a second version of our modelling of $W*$ as consisting of a set of atomic moments and in this version, we do not assume independence. It is still the case that the $z \in W*$ cannot usefully be split into smaller observer-moments. But here the underlying assumption is that it takes a certain amount of

[17] or more precisely they are able to make decisions as if they could instaneously derive any needed conclusion.

time to create a new belief, a new intention, a new decision. So excessively small observer-moments are not meaningful. Atomic observer-moments are, by definition, moments that are large enough to be meaningful, but not so large that they are capable of having subparts with contradictory beliefs. In this version of the definition, an atomic observer-moment is supposed to represent a minimal unit of consciousness.

The observer-moments in $W*$ are not assumed independent but for any $z \in W*$, we can still ask how much information is z capable of representing that is not represented by other observer-moments (i.e. how much new information, how much unique information can z represent? Not how much is actually new but how much could be new.). We assume that all atomic moments living in the same possible world are capable of representing the same amount of new (relevant) information.

But even if we did not make the assumption about new information, we still might argue that atomic moments living in the same world should have the same prior probability. If we really are reasoning as if we did not have any anthropic knowledge about who we are among the observer-moments in world w and x and y are atomic moments in w, we should think ourselves as likely to be x as to be y. This argument does not generalize to the case of nonatomic moments because if x or y is nonatomic, she might have no definite probabilistic belief about who she is among the observer-moments in $W*$ (there is no unique K_x or no unique K_y.) or no definite belief about which decision problem she has to solve.

5 The SIA and What is an Observer?

We now address the issue of exactly what is and what is not an observer or observer-moment. Our first comment is that an observer has to be able not only to represent information but also to use that information in a sensible manner. If faced with a decision problem, an observer needs to make sensible decisions and these decisions should be based on a computation of expected utilities. The computation of expected utilities requires posterior probabilities to be estimated based on the information available. There might be other reasons an observer might want to estimate posterior probabilities. In any case, an observer, actually an observer-moment, must be capable of having coherent probabilistic beliefs about who she is among the observer-moments in $W*$ and making coherent decisions based on those beliefs or acting as if she were making coherent decisions based on those beliefs[18],[19].

[18] Of course if z is an observer-moment, it might take time for her compute the necessary posterior probabilities and expected utilities and the actual computation of probabilities and utilities might be done by some other observer-moment based on the information available to z.

[19] In section 4, we discussed the possibility that observer-moments may not have the computational resources to solve all the problems they need to solve, but we can still model observers as being able to solve any given (simple) problem; we might model observer-moments as being born with partial solutions to many problems (they might have learned from previous observer-moments who belong to the same observer, for example), but they have difficulty handling a large amount of new information

As specified so far, there is a certain amount of circularity in our explanation of what an observer-moment is. A certain set W of worlds is interesting and then we enrich worlds in W with centers in order to obtain a set $W*$. This set is a set of observer-moments if each element of the set $W*$ is capable of reasoning coherently about who she is among the elements of $W*$. If we had chosen a different set X of possible worlds to start with, then we would also choose a different set of centered worlds $X*$ and there might be elements in $X * \cap W*$ that can be rational when reasoning about $X*$ but not when reasoning about $W*$. Thus the concept of observer-moment is not absolute.

Aside from a few ways in which we can easily take into account lack of logical perfection such as the fact that observer-moments in $W*$ do not have to be able to reason well about worlds outside of W, we are modelling observer-moments as being perfectly rational in making use of the information they have. Since ours is a prescriptive rather than a descriptive theory, observer-moments do not actually have to be rational in the way we model them as being rational. They just have to have the capacity to be rational. Even modelling them as having the capacity to be rational might be a bit of an idealization, but it still might be a useful idealization that can result in helpful advice being given to observer-moments about how they should estimate posterior probabilities.

We just want to make sure that we are not idealizing excessively. We would like to model normal adult humans as observers. But humans are very bad at estimating exact prior probabilities and even well-educated people with mathematical and scientific training can easily make bad mistakes in logic or fail to notice an argument that uses a long and complicated proof. That is true enough. But when the stakes are high enough, humans can and will make reasonable approximate probability and utility estimates especially if they can use artificial aids such as computers or obtain advice from expert observer-moments whom they trust.

A human $z \in W*$ might not be able to articulately describe a good prior probability distribution P for W, but that does not mean she does not have some implicit intuitions about the ratios of likelihoods of worlds in W and even intuitions about what these ratios would be if she had different information available than the information she actually does have. And these intuitions are not ad hoc or random but based on some guiding principles of which she might only have implicit knowledge. She might not know how either implicitly or explicitly to convert intuitive principles into prior and posterior probability distributions, but she could do so if she had sufficient help from experts or computational devices and sufficient time. There could, of course, be apparent contradictions in her intuitions but then she would just have to regard her intuitions as firstly only approximate and secondly only reliable with a certain probability that is less than

about which problems they are supposed to solve.

one. It still might be possible to arrive at reasonable prior and posterior probabilities at least for certain $W*$.

In any case, nonanthropic Bayesian reasoning if interpreted as giving advice to human decision-makers implicitly assumes that humans have a kind of rationality they really do not possess. But nonanthropic Bayesian reasoning is useful. In an anthropic context, it might also be useful to model humans as rational when dealing with certain issues.

So then, yes, humans are observers. What about superhuman intelligences? They are also observers although they might not be part of the same reference classes as humans. What about chimpanzees? Probably not for most $W*$ of interest to us, but there is much that we still do not understand about chimpanzees. What about Neanderthals? They are very much like humans and thus should probably be considered to be observers.

Certainly stones or even mosquitos are not observers for any interesting $W*$. But if we adopt a very liberal definition of what it means to be an observer or an observer-moment, we can reach an interesting conclusion.

Assume we have a set W of possible worlds that contains a finite set of actually minimally rational observers. We also imagine that each world contains a huge quantity of observers that are not minimally rational. Assume that if we include these extra observers, then all worlds have the same number of observers. Let us also assume that all observers, including not minimally rational observers, in the same possible world have the same prior probability. Now let us calculate a posterior probability in stages. We want the posterior probability that world $w \in W$ is actual. We start with a prior probabiiity $P(w)$. This is just the nonanthropic prior probability of w. Since we actually are a minimally rational observer, let us just conditionalize on the information that we are minimally rational. After conditionalization on this information, we obtain a semiprior probability $Q(w) = \frac{N_w P(w)}{\sum_{v \in W} N_v P(v)}$ where for any $v \in W$, N_v is the number of minimally rational observers in v. This follows because each observer in world w has prior probability $\frac{P(w)}{M}$ where M is the total number of observers in a possible world. Thus the prior probability of being a minimally rational observer in world w is $\frac{N_w P(w)}{M}$ and for any world v, the prior probability of being a minimally rational observer in world v is $\frac{N_v P(v)}{M}$. Thus $Q(w)$ is the conditional probability of w being actual given that we are minimally rational. The semiprior gives more probability, other things being equal, to worlds with more observers. Thus in essence we have made the Self-Indication Assumption[5] (SIA) according to which we need to first adjust our nonanthropic prior probability distribution to take into account the fact that we exist and are at least minimally rational. Then we can conditionalize on any more specific information that we have. In essence it is as if we are using Q rather than P as our nonathropic prior probability distribution since we really do not have to account for the imaginary not even minimally rational observers and whatever they might know and believe and decide.

It is amusing that we can derive the SIA as a special case of the SSA and we could also obtain a version of the SIA as a special case of the SSSA. We would just need to have a measure on the observer-moments in $W*$. For example, we might use the information theoretic measure. Then we would let N_v equal the total number of bits of information represented by the observer-moments in world v rather than the total number of minimally rational observers and apply the same formula as in the previous paragraph for computing a semiprior given a nonanthropic prior. But this SIA seems ill-motivated and unjustified. It really is ultimately an assumption that we should reject, but it has many advocates and there are some appealing arguments in favor of the SIA which do not involve vague references to taking into account the mere fact that we exist as conscious rational beings and do not involve artifical nonrational observers.

6 The SIA

The SIA does start to seem appealing if we consider a scenario where there are only two possible worlds $v, w \in W$ and no two observer-moments in the same possible world are in the same subjective psychological state. We assume that $P(v) = P(w)$ and that there is some observer-moment z in v who does not know which world is actual. For all she knows she might be y who lives in world w. Given that one cannot have two different observer-moments in the same subjective psychological state in the same world, that means $K_z = K_y = \{y, z\}$. If we apply the SIA, then we obtain that $Q(v) = \frac{N_v}{N_v+N_w}$. But since here we are dealing with observer-moments rather than observers, N_v and N_w should represent the total amounts of information that are represented by observer-moments in worlds v and w respectively. Let a be the amount of information that y and hence z can represent. We have then $P_z(z|K_z) = (\frac{a}{N_v})\frac{N_v}{N_v+N_w} = \frac{a}{N_v+N_w}$ and a similar computation shows that $P_z(y|K_y) = \frac{a}{N_v+N_w} = P_z(z|K_z)$. This seems satisfying: The worlds v and w are equally likely. That means y and z are equally likely to exist. So if all an observer-moment knows is that she is either y or z, should she not believe both options equally likely? If we did not apply the SIA and $N_v \neq N_w$, then z would derive from the fact that observer-moments in knowledge state K_z are not equally atypical in the two worlds the conclusion that the two worlds are not equally likely.

The SIA also has unfortunate consequences[20]. If there are only two possible worlds v and w and both

[20] The argument against the SIA presented in this paragraph is similar to an argument (the Presumptuous Philosopher's Scenario) presented in [5].

worlds would be considered equally likely if we did not take into account any information we have that other observer-moments do not have and the capacity of an average observer to represent information is the approximately the same in the two worlds but w has $10^{10^{100}}$ times as many observers as v, the SIA would make it very difficult for us to learn that v is the actual world. Before we acquire any evidence, we would think w virtually certain to be actual and given the very real possibility (probability at least 10^{-1000}) of error in observation or interpretation of any sequence of observations, it is difficult to see how that initial bias in favor of w could be overcome. Therefore, a Presumptuous Philosopher might tell us that it is really unnecessary to expend any effort collecting evidence; the actual world is w. Given the fact that there are viable cosmological theories that disagree about how many observers there are but do not disagree radically about how atypical we are, the Presumptuous Philosopher Scenario is a real problem for the SIA.

But we might have the problem that it is difficult for additional evidence to overcome a strong initial bias even if we do not make the Self-Indication Assumption. We might consider again the example where there are only two possible worlds and no two observer-moments in the same possible world can be in the same subjective psychological state. If we are observer-moment z in world v or observer-moment y in world w with $K_z = K_y$, then our posterior probability estimate for v will be $\frac{\frac{a}{2N_v}}{\frac{a}{2N_v}+\frac{a}{2N_w}}$ if we do not apply the SIA. But we might have $N_w = 10^{10^{100}} N_v$. In that case, v would seem virtually certain to be actual. Instead of the nonanthropic evidence giving about equal support to the theory that v is actual and the theory that w is actual, we might have strong nonanthropic evidence in favor of w being actual. But it is very hard to counteract the effect of the huge size of $\frac{N_w}{N_v}$.

We might note that if the number of observers in the actual world is huge, the assumption that no other observer-moment is in the same subjective psychological state as we are becomes rather doubtful. Our argument in the previous paragraph depended on our being very atypical if world w is the actual world. It is only reasonable to think that we might be all that atypical in world w if in w there is a huge number of subjective psychological states that are actually instantiated by some observer-moment in w (huge compared with the number of instantiated states in v). Should one try to maximize number of observers (as the SIA would tend to lead us to do) or minimize subjective psychological states (as we would tend to do without the SIA)? The latter seems more reasonable and more reminiscent of Occam's razor.

Or we might take a different approach to the problem of untestability. Perhaps there is something wrong with how we choose our supposedly nonanthropic priors when huge numbers are involved. There are problems both with and without the SIA. Or maybe it is not our anthropic reasoning theories that are responsible for our difficulties; there are nonanthropic reasoning theories that tells us that (on entirely nonanthropic grounds) one hypothesis is 10^{100} times more likely than another hypothesis. Here again we will run into problems because it will be difficult to use empirical evidence to overcome that factor of 10^{100}.

To judge the SIA, we need to return to fundamentals rather than examine what happens when we apply the SIA to scenarios involving huge numbers. And the transition from the prior P to the semiprior Q really is inadequately motivated. The strongest argument for the SIA is the argument involving the scenario with the two worlds v and w and two observer-moments z and y and the fact that if we apply the SIA, we do get $P_z(w|K_z) = P_z(v|K_z)$. And thus z (as well as y) should consider herself as likely to be y as to be z.

But this argument is not totally convincing. It is true that objectively it would be better if both y and z agree that that the posterior probability of v being actual is .5 rather than agreeing on some other number and since y and z are in the same subjective psychological state, they will agree on some number. To see that it is better if they choose .5 rather than some other number, we just have to consider wagers f with $f(x) = 0$ for all observer-moments except y and z. If we use the objective, nonanthropic prior probability of .5 for v, we see that if $.5(f(y) + f(z)) > 0$, it is better that both observer-moments accept the wager. If $.5(f(y) + f(z)) < 0$, it is better if both reject. This seems to suggests that y and z should use a posterior probability of .5 for v. Both y and z can even agree that it would be better if they both used the posterior probability of .5 rather than both using some other value but y and z do not get to have a dialogue about what to believe. They make separate decisions. If either y or z applies the SSSA and not the SIA and N_w is much bigger than N_v, then she (y or z) will conclude that v is much more probable than w and hence she is much more likely to be z than y.

Both y and z could say, “If I were to believe that I am as likely to be z as to be y, that would not force the other guy (z if I am y and y if I am z) to believe the same thing and in any case, I am more likely to be z than y. I know that means that the other guy is more likely to be y than z. But the other guy is in the same subjective psychological state as me and will think herself to be much more likely to be z than y and she will be wrong. But so what? She does not exist.”

We might also consider applying the SIA to a scenario where there are two possible worlds, v and w, $P(v) = P(w)$, $N_v = 1$, $N_w = 10^{100}$ and all observer-moments in both worlds are in the same subjective psychological state. If there is a wager f such that $f(z) = -10^9$ if $z \in v$ and $f(z) = 1$ if $z \in w$, then if all observer-moments make the Self-Indication Assumption, they will all accept the wager and the expected total return is $.5(-10^9)+.5(10^{100})$ which is huge and positive. Thus it would seem, given that all observer-moments care about expected total rather than expected average utility, that it is better that all observer-moments accept rather than all reject. All observer-moments will accept if they believe the SIA and thus believe it vritually certain w is actual. They will all reject if the SIA is not used. But even if it is better that all accept rather than all reject, we should keep in mind that decisions are not made by a committee of all observers. If in both worlds all observer-moments except one accept and that one observer-moment rejects, than the total expected return is $.5(10^{100} - 1) > .5(10^{100} - 10^9)$ and that would be a reason for each individual observer-moment to reject if she believes all other observer-moment (if there are other observer-moments) will accept. Each individual observer-moment would reject if she did not use the SIA. Of course, if every observer-moment rejects, inferior results are obtained, but each individual observer-moment makes her own decisions.

In the previous paragraph, we used the nonanthropic prior probability $P(v) = P(w) = .5$ to compute expected total return. If we really believe the SIA, we might suggest that when computing expected total return, we should consider w much more likely than v; after all if all observer-moments believe the SIA, they will all believe w much more likely. But if we took that suggestion, we would run into other problems. We might consider a scenario where v and w would be considered equally likely if we ignored the SIA, but $N_w = 10N_v$ and all observer-moments are in the same knowledge state. We have a wager f such that $f(z) = -50$ for z living in v and $f(z) = 1$ for z living in w. Even if the SIA is applied, w will not be considered to be even 50 times as likely as v to be actual and all observer-moments will reject f. But if we compute an expected total return in which w is ten times as likely as v because there are ten times as many observers (or ten times as much information capacity to store information), then the expected total return if all observer-moments accept f is positive: It is $\frac{-50N_v}{11} + \frac{10(10N_v)}{11}$. This suggests that if a wagering argument based on the wager of the previous paragraph is a reason to believe the SIA, then this paragraph provides a reason to believe a strong form of the SIA which would give w 100 times as much prior probability as v rather than just 10 times. And then we could easily enough construct an argument for a superstrong SIA.

7 Choosing A Reference Class

7.1 Should $R_z = K_z$?

We have analyzed the application of the SSA and SSSA both with and without the SIA. In both cases, there were difficulties. If we use the SIA in a version of the fundamental scenario in which $c = e = 1$ (one observer-moment in state A in both worlds),$f > 1$ and $d = 0$ (observer-moments in state B only exist in world w), observer-moments in state A will be believe v and w equally likely and that might be what we want. Without the SIA, v and w would not be believed equally likely.

If we apply the SIA to a scenario in which there are only two possible worlds, v and w with $P(v) = P(w)$, $K_x = K_y$ for any $x, y \in \{v, w\}*$ and $N_w = 10^9 N_v$, then (as in the Presumptuous Philosopher Scenario) w will be thought much more likely than v. If we do not apply the SIA, we can see that observer-moments will realize that v and w are equally likely.

But all this assumed the use of maximal reference class. If we do not use maximal reference classes $R_z = W*$ but instead use minimal reference classes $R_z = K_z$, then we could have a probability estimate of .5 in all these scenario. The use of minimal reference classes, at least in certain scenarios (but which ones?) has been advocated by [4].

There are three main problems with minimal reference classes. The first is that if for all $z \in W*$ and all $w \in W$, $K_z \cap w* \neq \emptyset$, then no observer-moment will ever learn very much about which world is actual. All observer-moments have in common a certain amount of knowledge. That is why they can agree on a nonanthropic prior P. But no observer-moment will be able to learn very much from the fact that she has some observational evidence that other observer-moments do not have. If z has evidence E, then in every world in W, there is some observer-moment who also has evidence E. Thus she cannot exclude any world w as impossible just because she has observed E. She will compute $P_z(w * | K_z) = \frac{P_z(w*\cap K_z)}{P_z(K_z)}$ but $P_z(w * \cap K_z)$ is just $P_z(w*)$ because there are no observer-moments in world w who are in the reference class but not in K_z. Also $P_z(K_z) = 1$ because K_z is the whole reference class. Thus the posterior probability of $w*$ is the same as the prior probabilty. It is impossible to learn from observation. We might in fact be faced with a situation similar to this in cosmology. There might be many plausible cosmological theories which do not differ in their predictions as to which subjective psychological states will be experienced by at least one observer-moment[3].

The second problem is best illustrated by a scenario in which there are only two possible worlds, v and w and only two possible subjective psychological states A and B. In both v and w, there is just one observer-moment in state A. In v, she is the only observer. In w, there are millions of observers. The prior probability of v is .5. If an observer-moment is in state B, then she has conclusive evidence against v being actual. That would lead us to think that learning that one is not in state B but rather in state A would be evidence in favor of v being actual. But if we use $R_z = K_z$ for all observer-moments, then the observer-moment who is in state A would believe the posterior probability of v to be .5. and not use the fact that she has failed to observe B as evidence.

The third problem is that in many cases, if minimal reference classes are used, a collective Dutch Book can be constructed. A collective Dutch Book can also often be constructed even if reference classes other than minimal are used. We shall analyze below why we need to use not just nonminimal but actually maximal reference classes and one of our arguments for our position will be a collective Dutch Book argument.

7.2 Why should $R_z = W*$?

We believe that almost maximal reference classes should be used: $R_z = W*$. $W*$ might not, strictly speaking be maximal because it might be the case that there are observer-moments X who do not belong to $W*$ but there are some observer-moments in $W*$ who cannot reason coherently about whether they are elements of X.

Keeping in mind that $W*$ need not be all possible centered worlds, we can present our arguments for why we should have $R_z = W*$ for all $z \in W*$. We will still assume that that $W*$ is finite. We also assume that if $x, y, z \in W*$ with $\hat{x} = \hat{y}$ and $x, y \in R_z$, then the ratio $\frac{P_z(y)}{P_z(x)}$ is equal to the ratio of the amounts of information that x and y are capable of representing. These amounts are assumed to be a finite number of bits.

The first argument is a conceptual argument. $R_z = W*$ is a simple rule. Perhaps $R_z = K_z$ is even simpler, but that choice can run into problems for reasons we discussed above. A reason to think $R_z = W*$ is especially simple is that the rule $R_z = W*$ is consistent with the philosophy that priors should represent maximal ignorance and we represent more informed knowledge states by conditionalization of the prior. This is a simple and elegant philosophy.

The second argument is a collective Dutch Book argument. We assume that there exists four observer-moments v_1, v_2, w_1, w_2 with $K_{v_1} = K_{w_1}$ and $K_{v_2} = K_{w_2} \neq K_{w_1}$ as well as $\hat{v_1} = \hat{v_2}$ (let v represent the world v_1 and v_2 both inhabit) and $\hat{w_1} = \hat{w_2}$ so we shall say that both w_1 and w_2 live in w. We shall set up a collective Dutch Book. Observer-moments v_1 and w_1 will be betting on whether they are v_1 or w_1 and observer-moments v_2 and w_2 will be betting on whether they are v_2 or w_2. So we need to know the ratios of the posterior probabilities of v_1 and w_1 and the ratio of the posterior probabilities of v_2 and w_2. To compute these ratios, we make use of the fact that for all $z \in W*$, if $y \in K_z$ the rato $\frac{P_z(z|K_z)}{P_z(y|K_z)} = \frac{P_z(z)}{P_z(y)}$ and thus we need only compute ratios of prior probabilities. First we see what happens if all observer-moments use maximal reference classes.

We make a simple computation. $\frac{P_{v_1}(v_1)}{P_{v_1}(w_1)} = \frac{P(v)(\frac{I(v_1)}{I(v*)})}{P(w)(\frac{I(v_1)}{I(w*)})}$ taking into account the fact that $I(v_1) = I(w_1)$. Thus the ratio of the prior probabilities of v_1 and w_1 is equal to $\frac{P(v)I(w*)}{P(w)I(v*)}$. This will also be the ratio of the prior probabilities of v_2 and w_2. But if we do not use maximal reference classes there is no guarantee that these probability ratios are equal. If they are not equal, we can set up a collective Dutch Book.

The Dutch Book argument requires that the utilities of different observer-moments be combined. If we need to combine the utilities of v_1 and v_2, we cannot simply add up or average their utilities. If we are some observer-moment $z \in W*$ reasoning as if we do not know who we are among the observer-moments in R_z, we would try to act in such a way as to maximize $\sum_{y \in R_z} P_z(y)U(y)$[21] where U stands for utility. This tells us that when we combine the utilities of v_1 and v_2, we need to compute a weighted average with weights being proportional to prior probabilities. But if two observer-moments in R_z both live in the same world their prior probabilities are proportional to the amount of information they can represent.

We are now ready to define a wager that sets up a Dutch Book. Assume that nonmaximal reference classes are used and as a consequence $\frac{P_{v_1}(v_1)}{P_{v_1}(w_1)} > r > s > \frac{P_{v_2}(v_2)}{P_{v_2}(w_2)}$ for some real numbers r, s. Let $\frac{I(v_1)}{I(v_2)} = q$. So we can use $qU(v_1) + U(v_2)$ and $qU(w_1) + U(w_2)$ to combine utilities of observer-moments in the same world[22]. Let $a > 1$ be chosen so $sa < r$. Finally define the wager f so that $f(v_1) = \frac{1}{q}$, $f(v_2) = -a$, $f(w_1) = -\frac{r}{q}$, $f(w_2) = as$, and $f(z) = 0$ for any observer-moment z other than v_1, v_2, w_1, w_2. Then v_1 (and w_1) will accept the wager because she thinks herself more than r times as likely to be v_1 as to be w_1. The observer-moment w_1 (and w_2) will accept the wager because she believes herself less than s times more likely to be w_1 than to be w_2. But then when we combine the utilities of the observer-moments in world v, we get $q(\frac{1}{q}) - a < 0$ and $-q(\frac{r}{q}) + as < 0$. So no matter which world is actual, we have a loss of utility.

Our third argument is a relative frequency argument. In order to test the adequacy of our theory of anthropic reasoning, we consider what happens if a scenario is repeated many times. We ask ourselves if there is an approximate equality between relative frequencies and our posterior probability estimates. Let us consider a finite scenario. Thus both W and $W*$ are finite. In order to completely describe our scenario, we would have to specify the values of $I(z)$ for all $z \in W*$ and also specify a nonanthropic prior probability distribution P on W. Now let M be some very large number and let us repeat our scenario M times. In order for the relative frequencies we compute to be most meaningful, we want the different repetitions of the same scenario to be independent repetitions.

Thus we will define a product scenario(actually a power scenario). The set of all possible worlds is W^M. Thus W^M is the set of length M sequences of worlds belonging to W. Any observer-moment living in a world $w = (w_1, w_2, w_3, \ldots, w_M) \in W^M$ is a sequence $(z_1, z_2, z_3, \ldots, z_n)$ such that for $1 \leq i \leq M$, z_i is an observer-moment belonging to w_i. We have $P(w) = \prod_{1 \leq i \leq M} P(w_i)$ if $w = (w_1, w_2, w_3, \ldots w_M)$ where $P(w_i)$ is the prior probability of w_i in the unrepeated scenario. We want $I(z) = \prod_{i \leq i \leq M} I(z_i)$ where $z = (z_1, z_2, z_3, \ldots, z_M)$. Of course the prior probability of $(w_1, w_2, w_3, \ldots, w_M)$ should be the product of the

[21] We assume that overlap is not a problem.

[22] But we can set up our Dutch Book for any $q > 0$.

prior probabilities of the w_i if we are going to be independently repeating the same scenario M times. It is not quite as intuitive why the amount of information is also computed by taking a product, but if we imagine that z_1 has six different places where it can store information and z_2 has five different places, then there are six times five ways of obtaining an ordered pair of places for storing information with the first member of the ordered pair coming from z_1 and the second from z_2. We also need to define knowledge states. We do so in the obvious way: $K_z = \prod K_{z_i}$ if $z = (z_1, z_2, z_3, \ldots, z_M)$.

For any world $v \in W$, if we choose a world w at random from W^M, we see that if M is large enough that it is virtually certain that the proportion of w_i in the M-tuple $(w_1, w_2, w_3, \ldots, w_M)$ such that $w_i = v$ is approximately equal to $P(v)$. Thus the number of indexes i such that $w_i = v$ should be approximately $MP(v)$. This is one examle of relative frequencies in the repeated scenario and probabilities in the unrepeated scenario being approximately equal. But we are more interested in probabilities that take into account anthropic knowledge.

Let us choose an observer-moment $z = (z_1, z_2, z_3, \ldots, z_M)$ who lives in the random world w. We want to analyze what z knows and that means we need to analyze what each z_i knows. We are actually not interested in all indexes. We will pick some possible knowledge state $K \subseteq W*$ and some world $v \in W$. (So we are analyzing some world of the unrepeated scenario and some possible state of knowledge of observer-moments in that scenario.) We are interested in the sets i_K and i_v where i_K is the set of i such that $K_{z_i} = K$ and i_V is the set of i such that $w_i = V$.

Let $n_{K,v}$ be the total amount of information that can be represented by the set of observer-moments z in world v who have $K_z = K$. If z is a typical observer-moment in the product scenario and M is large enough, we expect the number of indexes i that are in $i_K \cap i_v$ to be approximately equal to $P(v)M(\frac{n_{K,v}}{I(v*)})$. If $i \in i_v$, the probability that $i \in i_K$ is $\frac{n_{K,v}}{I(v*)}$ if z is a typical observer-moment in $(W*)^M$ (and hence z_i is a typical element of $w_i = v$) and if M is large, it typically will be the case that a fraction approximately equal to $\frac{n_{K,v}}{I(v*)}$ of the $i \in i_v$ will also be in i_K. The number of indexes i in i_K should be approximately equal to $M \sum_{x \in W} P(x) \frac{n_{K,x}}{I(x*)}$ if M is large enough. Thus if M is large enough the proportion of indexes in i_K that actually belong to $i_K \cap i_v$ is approximately equal to the posterior probability for world v that would be computed by an observer-moment in knowledge state K in the unrepeated scenario (assuming the observer-moment uses maximal reference classes). So if we want relative frequencies to be approximately equal to probabilities, we should use maximal reference classes.

All this assumes that z is typical. The typicality was computed using a prior probability distribution on centered worlds in the product scenario. This distribution implicitly assumed that maximal reference classes should be used but if we are discussing how unusual we are among all the observer-moments who live in W^M, there really is no natural alternative to maximal reference classes. We need a probability distribution that is defined for the whole set of observer-moments who live in w[23]. . But z could be atypical. So we have not proven our theory to be the only reasonable theory of anthropic reasoning. There are other nonabsurd theories of anthropic reasoning. And we have barely discussed the infinite case. In the next section we discuss both issues.

8 The Infinite Case and Learning from Experience

Up until this point we have focused on the case where the total number of worlds or equivalence classes of worlds is finite, the number of observers in each world is finite and each observer could only represent a finite amount of information. But the actual world might contain an infinity of observers. So we have to be able to analyze the infinite case.

Scenarios involving infinite sets can be problematic even if no anthropic reasoning is involved. There really is no magic formula for handling the infinite case. We just have to represent the infinite case as a limit of finite cases which we know how to handle. The problem is that there are many different ways of representing an infinite scenario as a limit of finite scenarios which are intended to approximate the infinite scenario.

We could approximate an infinite case by saying that all except finitely many observers in finitely many worlds are irrelevant. We could approximate an infinite case by saying that we shall work with equivalence classes of worlds and observer-moments and only deal with a finite number of equivalence classes. Thus perhaps each centered world can be represented in a canonical way as an infinite sequence of bits and we shall not distinguish between centered worlds whose first few bits are the same. Or we could combines the idea of working with equivalence classes with the idea of working with subsets.

[23] One might object to the whole idea of using a product scenario in which a typical observer-moment $z = (z_1, z_2, z_3, \cdots, z_m)$ can be represented by a list of z_i with very different reference classes. We might think that relative frequencies in such a product scenario tells us nothing about the unrepeated scenario because it mixes up observer-moments in different reference classes but if all that is different about the observer-moments is how much they know and what they know, it does not seem strange to mix up observer-moments of different kinds. I might conduct an experiment which can have several possible results and some of the information about the results might be known to some observer-moments but not other observers in the certain possible worlds. If I repeat the experiment several times, it might be quite likely that which set of observers know which of the results might change and the results themselves might also be different. If we use minimal reference classes, observer-moments z_i would use different reference classes just because they know different things and yet it seems to perfectly reasonable to mix up these different observer-moments in one product scenario. If we use the fundamental scenario as our unrepeated scenario and specify that the only reason that some observer-moments are in state A and others are in state B is that they have reached different conclusions after following the same experimental protocal and they have only reached different conclusions because they have read a different number on a dial, then it seems perfectly reasonable to mix up state A and state B observer-moments in a repeated (i.e. product) scenario.

But there are too many possible ways of representing an infinite scenario as a limit of finite scenarios. It is difficult to think of a purely theoretical and totally convincing reason for preferring one method to another method. We have to learn from experience which method works best. Even in the finite case, our arguments for our theory of anthropic reasoning are not incontrovertible. Someone else might have a different theory that is not implausible and ultimately we will rely on our experience to determine which theory is better.

We use standard Bayesian methodology for using experience to determine which of two theories T_1, T_2 of anthropic reasoning is superior. We judge theories by their prior probability of being the correct theory. (So we assume we know $P(T_1)$ and $P(T_2)$, the prior probabilities of the two theories. Different researchers will disagree about what these probabilities should be, but we can hope that the disagreement is not so great that there cannot be agreement on which theory has the greater posterior probability.) We also judge theories by how well they predict the available evidence E. So we need to know $P(E|T_1)$ and $P(E|T_2)$. In our intended application, if we are an observer-moment z, our available evidence is represented by our knowledge state K_z. So $P(E|T_i)$ would represent the prior probability $P_z(K_z)$ that according to theory T_i, z should give to K_z. We can compute the odds ratio:

$$\frac{P(T_1|E)}{P(T_2|E)} = \frac{P(T_1)P(E|T_1)}{P(T_2)P(E|T_2)}$$

in order to compare the adequacy of the two theories.

There are two important points that need to be made about our approach for testing theories. One point that needs to be made is that we are only comparing theories of anthropic reasoning; T_1 and T_2 are assumed to agree on the correct nonanthropic probability distribution for W. Another point is that we are assuming that both T_1 and T_2 use the SSSA (or SSA) and that our formalism makes sense when applied to T_1 and T_2 but the SSSA formalism is quite flexible. We can choose to use whatever R_z and P_z we please provided our P_z is consistent with the nonanthropic prior P and we can even add artificial observer-moments to $W*$. So $W*$ might include stones.

There are also pitfalls associated with our Bayesian methodology. Any theory T that uses minimal reference classes $R_z = K_z$ will have $P(E|T) = P_z(K_z) = 1$. Thus T gets too much credit for predicting the evidence E. One way to compensate for this is to declare that T has low prior probability. Another approach is to require of T more than just that it work in predicting what we experience; we might also require that it work when used by other observer-moments. But if we are z and we are assessing how well T worked when used by some other observer-moment y, we can assess things other than how well K_y is predicted. For example, y and z might be observer-moments belonging to the same observer o with z being the unique observer-moment who lives four hours after y in world $\hat{y} = \hat{z}$. The observer-moment y might make probabilistic predictions about what z will observe and we can assess how true those predictions are.

To formalize and generalize what is happening, we introduce the formalism of special relationships: One special relationship might be being the best friend of and another special relationship might be belonging to the same observer-moment in the same possible world but living four hours later. The formalism of special relationships allows us to refer to observer-moments using relative rather than absolute vocabulary. A special relationship S is just a relation $S \subset W * \times W*$ such that if aSb, then $\hat{a} = \hat{b}$. Thus a special relationship is just a set of pairs of centered worlds in which each pair is a pair representing observer-moments living in the same (bare) possible world. A special relationship is not necessarily a function or the inverse of a function: We can have a special relationship S and $a, b, c, \in W*$ with $a \neq c$ with both aSb and cSb and can have a special relationship with $b \neq c$, aSb and aSc.

If $y \in W*$ and y knows she is y (i.e. $K_y = \{y\}$) and S is a special relationship, y can consider $S(y) = \{z : ySz\}$. For any z in $W*$, she can compute $P_y(K_z|S(y))$, the probability that a random observer-moment in relationship S with y will be in knowledge state K_z. If y does not know that she is y, but only that she is in the larger set K_y, she can still predict $P(y, S, K_z) = \sum_{x \in K_y} P_y(x|K_y)P_x(K_z|S(x))$. This would be y's prediction of the probability that a random observer-moment in relationship S to herself is in state K_z.

But we are not the predictor y, but some other observer-moment z testing the accuracy of y's predictions. Assume that there is only one special relationship S of interest to us and that for any $x \in K_z$, there is at most one observer-moment y such that ySx. If ySx, use the notation $S^{-1}(x)$ to refer to y. In this situation, we might say that $P(E|T) = \sum P_z(x)P(S^{-1}(x), S, K_z)$ where we sum over all $x \in K_z$ such that $S^{-1}(x)$ exists. We assume that at least one such x exists. We are evaluating how likely it is that the observer-moment who is in relationship S with us will predict that the observer-moment with whom she is in relationship S will be in knowledge state K_z.

If there are several different relationships of interest to us or if there exists $x \in K_z$ such that there is more than one y with ySx, then our calculation becomes more problematic. To treat the case where there can be several different y in the relationship with a given $x \in K_z$, we redefine $S^{-1}(x)$ to refer to the set of y such that ySx and then we might write $P(E|T) = \sum P_z(x) \sum C_y P(y, S, K_z)$ where the second summation is over all $y \in S^{-1}(x)$ and the C_y are weighting coefficients such that $\sum_{y \in S^{-1}(x)} C_y = 1$. These C_y weight the relative importance of the different y. We would like to use prior probabilities to determine the weights, but which prior probabilities? We cannot assume that according to theory T, any of the y are in R_z. We cannot assume that there is any $v \in W*$ such that $S^{-1}(x) \subset R_v$. We might just have to say that the C_y are some measure of how much we care about each y; perhaps the C_y have to be obtained by using some other theory than T to compute prior probabilities. If there were two different special relationships S_1, S_2 but both $S_1^{-1}(x)$ and $S_2^{-1}(x)$ never contained more than one element, we would have to compute $\sum P_z(x)(C_1 P(S_1^{-1}(x), S_1, K_z) + C_2 P(S_2^{-1}(x), S_2, K_z))$ where $C_1,\ C_2$ are weights that have to be determined. These weights might represent how much we care about the different special relationships.

Even if there is only one special relationship S of interest to us and $S^{-1}(x)$ never has more than one element for $x \in K_z$, our test might not be very adequate if we might be too similar to the $S^{-1}(x)$. If there is not very much that we know and the $S^{-1}(x)$ do not know, then there will not be much actual predicting for us to test.

9 Has Anthropic Reasoning Already Been Disconfirmed by Observation?

It has been claimed that anthropic reasoning has already been disconfirmed by experience[18]. But that is impossible. If we have to estimate posterior probabilities and we have strictly anthropic knowledge, then we have to use some theory of anthropic reasoning even if our theory is that we should use minimal reference classes, which means that we are ignoring our anthropic information if we only want to predict nonanthropic facts. Although it cannot be true that anthropic reasoning is mistaken; it can be true that there exists a theory of anthropic reasoning is that is superior to the one we advocate or that no one has yet successfully constructed a plausible and usable general theory of anthropic reasoning that actually works and allows us to make useful predictions.

There are several reasons one might think that anthropic reasoning has already been disconfirmed. For example, if we use anthropic reasoning of the kind advocated in this paper, we might find it remarkable that we are observer-moments who are living in a civilization with a relatively small capacity for representing information[24]. But these arguments against anthropic reasoning depend on assumptions about nonanthropic priors. We might be using the wrong nonanthropoic priors. Another possibility is that our nonanthropic priors might not be prior enough. If we really did not know whether we inhabited a region of space-time with civilizations which currently can represent amounts of information that are much greater than what our

[24] That is how we would translate into our terminology [17]'s concern that there might be a conflict between anthropic reasoning and observation because we are part of a civilization with only a few observers.

civilization can represent, we might use very different priors. Another possibility is that our information-theoretic criterion is incorrect. Or perhaps we are incapable of reasoning coherently about most of the observer-moments who live in other civilizations and in that case anthropic reasoning really would not apply.

10 On Ignoring Anthropic Information and Updating As Communication

A point that needs to be made is that for most of this paper, we have not assumed that there is any special relationship between observer-moments who belong to the same observer. So there is no reason that anything like Meacham's Learning Principle[15] be valid.[25] But in practice, we usually update by a procedure that is fairly close to conditionalization of a chronologically prior distribubtion (a distribution believed in by a previous observer-moment who is part of the same observer as we) and we often ignore anthropic information when computing probabilities if all we care about is the posterior probabilities of uncentered possible worlds. And we might wonder why this is often acceptable. It is often acceptable for the reason that we are allowed to conditionalize on the knowledge K_z in stages, rather than do all the conditionalization in one step.

If we are $z \in W*$, we might divide our knowledge K_z into parts K_i such that $\cap_{i \in I} K_i = K_z$ and split the process of conditionalization on K_z into a set of processes of conditionalizing on each K_i. Some of these K_i might represent messages we receive from other observer-moments and some might represent other kinds of knowledge (i.e. they might represent what we are currently observing now). So we might first conditionalize on the knowledge we receive from other observer-moments and then conditionalize on the knowledge we receive from our current observation. So we are viewing updating as communication[16].

All any other observer-moment y can truthfully tell us is that she (y) belongs to some superset K of K_y[26] (i.e. $K \supset K_y$). She will not necessarily direct that message only to us or direct the message to all other observer-moments who live in the same world; instead we assume that the message goes to all observer-moments x such that yRx where R is some special relationship. So we might use the notation $M_{K,R}$ to refer to messages we think we have received. If we think we have received $M_{K,R}$ that means we

[25] According to the Learning Principle, "A sequential updating rule R should be such that the subject's current de dicto credences lie in the span of the credences R prescribes to her extended doxastic epistemic successors." In Meacham's formulation, a doxastic successor of observer-moment z is a later time-slice of the observer to which z belongs. We need to allow for extended successors because observers might die or become unconscious and in that case R should just assign reasonable credences to imaginary observer-moments that would have existed if only certain observers were not dead or unconscious during certain time-intervals. But regardless of how we fill in the details of what it means to be an extended successor, in our formulation, there is no special relationship between observer-moments merely because they belong to the same observer.

[26] She might try to tell us something about her posterior probability distribution $P_y(\ |K_y)$ but since her anthropic prior P_y should be the same as ours, the most she can really tell us is that she belongs to K_y.

believe that someone in relationship R to us has sent us a message that they know that they belong to K. It is, possible, that we have misread or misinterpreted the message. It is possible that we are just imagining that we have received a message when we really have not received any such message. It is possible that the observer-moment who sent us the message might be intentionally or unintentionally deceiving us. Or that when we receive a message $M_{K,R}$, the message really means more than just that some observer-moment in relationship R to us knows K; it means more because certain observer-moments who know K and are in relationship R to us do not send us the message $M_{K,R}$ because they are inarticulate. All this is possible but under favorable circumstances, we should be able to compute reasonable posterior probabilities by a simple procedure that makes use of the information in our messages.

Assume that we do not have to worry about misreading or misinterpreting messages or observer-moments lying about what they know or observer-moments being inarticulate. Thus we might consider an observer-moment in world w who actually receives the message $M_{K,R}$ as being equivalent to any observer-moment $y \in w*$ such that there exists $x \in w*$ with xRy and $K_x \subset K$. Assume also that we might consider observer-moments in relationship R to each other to be equivalent. Finally assume that the difference between equivalent observer-moments is basically irrelevant. All these assumptions might be true if R represents the relationship of belonging to the same observer in the same world but living one moment earlier and we do not care about our temporal location. Under these assumption we might reason as if K contained part of our relevant knowledge; actually we might know that we do not belong to K (an earlier time-stage of us belonged to K) but if we ignore irrelevant details, we might regard ourselves as belonging to K. We can use the relevant knowledge in K to define a probability distribution on $W*$ that we might regard as a prior and then conditionalize on the additional relevant knowledge that we have.

11 Lazy Adam

At this point, we would like to discuss a particularly perplexing scenario the Lazy Adam scenario[2]. What makes the scenario particularly perplexing is that it involves one observer-moment making decisions that affect the total number of observers that exist (and the total amount of information that can be represented). One observer, Adam, can make a decision that will affect whether other observers exists. But certainly observer-moments can affect whether other observer-moments exist: People can choose to or choose not to reproduce and they can choose to take or not to take potentially suicidal risks.

We shall introduce the Lazy Adam scenario (actually a streamlined version of Bostrom's scenario) by first describing simpler scenarios and then gradually modifying them until we obtain a Lazy Adam sceanario. We start with a version of the Fundamental Scenario in which $c = e = 1$ while $d = 0$ and $f = 10^{100}$. Thus world v has exactly one observer-moment and that observer-moment is in knowledge state A. The world w also has one observer-moment in state A as well 10^{100} in state B. We might assume that there is some stochastic cosmological process h that took place early in the history of the universe (before there were any observers) that is responsible for the actual world being v or w and that h is a typical instance of a well-understand class H of processes and based on what every observer-moment knows about H, there is no reason to think v more likely to be actual than w, but once an observer-moment z who is in state A takes into account that she is in state A, she will believe it virtually certain that the actual world is v.

Nothing essential changes if instead of observer-moments just knowing whether they are in state A or state B, they also know some irrelevant information. We assume that every observer consists of just one observer-moment and that all observers know their birth rank; an observer has birth rank i if there were exactly $i - 1$ observers who were born before her. But what really matters is whether an observer is low rank (rank 1) or high rank (rank greater than 1). Once one knows whether one is low or high rank, then it is irrelevant what one's exact rank is.

Nothing essential changes if we allow each observer to consist of a large number of atomic observer-moments representing different time-slices of an observer's existence. Every observer is composed of the same number of atomic moments and every atomic moment knows not only her birth rank but also her moment rank. An atomic observer-moment z belonging to observer o is of moment rank i if exactly $i - 1$ atomic moments belonging to o lived before z started her life. The scenario we are describing in this paragraph is a simplified version of the Doomsday Argument scenario. A low birth rank observer-moment will think it virtually certain that the actual world is v.

Many researchers believe the conclusion of the Doomsday Argument (that world v is virtually certain to be actual) to be highly problematic. We believe the Doomsday Argument to be valid because the natural ways to avoid the Doomsday Argument conclusion do not work. The SIA is not an assumption we wish to make. If we use minimal reference classes, we can be Dutch Booked. There are good arguments for using universal reference classes. But in this section, we are not concerned primarily with whether the Doomsday Argument is valid but with whether the Lazy Adam scenario is really any more paradoxical than the Doomsday Argument scenario and our answer is no.

It does not really matter if it is a stochastic cosmological process or a stochastic process in the brain of some observer (the observer with birth rank one and let us call him Adam) that determines whether the actual world is v or w. The stochastic process in Adam's brain might determine whether or not he pushes a certain button of a cloning machine at a certain time t. If the button is pressed at time t, $10^{100} - 1$ clones of Adam will be constructed. Adam knows enough about biology and physics to know that the cloning machine works. If at time t, the button is not pressed, Adam will be the only observer. Adam, and in fact all observer-moments, know all this. They also all know that not taking into account any knowledge that one observer-moment has and another observer does not have but just taking into account objective physics and biology including neurology, they should believe that the probability that the button is pressed is $1 - 2^{-100}$.

Our calculations will be affected by the fact that the nonanthropic probability of v is 2^{-100} rather than .5. In order to explain our calculations, we also need to say more about the knowledge states of observer-moments. Time t is the kth moment of Adam's life. Every observer other than Adam knows at any time during his life which world is actual. The first $k - 1$ moments of Adam only know their birth and moment rank[27], but do not know which world is actual. All other moments of Adam do know which world is actual. Assume z is one of the first $k - 1$ moments of Adam. Because z is so much more atypical (10^{100} times as atypical) in world w than world v, z will believe v virtually certain to be actual. The factor of 10^{100} is much greater than the ratio $\frac{P(w)}{P(v)} = 2^{100} - 1$ of the nonanthropic probabilities of the two possible worlds.

We might have $\frac{P(w)}{P(v)} = 2^{100} - 1$ if Adam is under a compulsion to press the button at time t if a certain fair coin does not land heads 100 times in a row when it is tossed. We assume that the results of the one hundred tosses are independent and that Adam (and all other observers) know enough about physics and about the coin in question to know that the coin is in fact fair and the results of the tosses independent. What that means is that Adam before time t will predict that it is almost certain the coin in question will land heads one hundred times in a row. If Adam does not take into account the fact that he knows he is Adam or if all observers had a phase of their lives when they are totally ignorant of their birth rank and Adam were currently in that phase, then Adam will conclude that the coin almost certainly will not land heads one hundred times in a row. But if Adam does know he is Adam, he should use that information, but then he will predict that the coin will almost certainly land heads one hundred times in a row. If there were no connection between the coin toss sequence and Adam's pressing the button, he would predict the coin almost certainly would not land heads one hundred times in a row. This might seem like a spooky correlation (it seems that there is a spooky correlation between whether Adam has made a resolution to press the button and whether the coin will land heads one hundred times in a row) but the basic anomaly is the difference between anthropic (taking into account the anthropic knowledge Adam has that he is Adam)

[27] Adam does not see the actual tosses on a one by one basis; he just knows at moment k whether they all landed heads or not

and nonanthropic probability estimates for the probability that w is actual. Once there is an anomaly, it will usually not be too hard to play around with the anomaly and make it seem very striking. That there will sometimes exist a difference (a difference we might view as anomolous) between anthropic and nonanthropic probabilities is inevitable if we make it a policy to take into account anthropic information when estimating the probabilities of (bare) possible worlds.

We might now modify our scenario so that Adam is not necessarily under a compulsion to press the button if the coin does not land heads one hundred times in a row. He could cure himself of the compulsion if he wants to. If he cures himself, there is no cloning. So then Adam is really making a decision whether to make an irrevocable conditional resolution. This is a resolution to press the button if and only if the coin lands tails at least once during the one hundred toss sequence. The question is: Should Adam make the resolution?

We might first ask ourselves why Adam might want to make the resolution. One reason might be that Adam really wants that coin to land heads one hundred times in a row and he thinks by making that resolution, he can prevent the coin from landing tails even one out of a hundred times. If we are skeptical that Adam really cares that much about how the coin falls, we might imagine some other objectively improbable event like a wounded deer wandering into Adam's backyard.

But if Adam really wants that coin to land heads one hundred times in a row and there is no cost to making the resolution, why shouldn't he make the resolution? It might actually work. So we have to assume that there is a cost because making the resolution is inevitably part of a larger strategy. Adam will rely on the strategy working and thus bet on the improbable event occurring. In the case where the improbable event is a wounded dear walking into Adam's yard, the bet would mean that Adam would stay home and not go out hunting and gathering because he is confident the improbable event will happen. If it does not happen, he will be hungry. The bet is not worth making unless it is highly likely that the objectively improbable event occurs.

Should Adam make the resolution? If we apply causal decision theory correctly, we see that the answer is no. Assume Adam does not make the resolution. And the coin did not land heads one hundred times in a row. Then it would not have landed heads one hundred times in a row if he had made the resolution. There is no objective causal relationship between coin tosses and making the resolution and Adam knows that[28]. Another way to see what is going on is to notice that the decision maker is Adam regardless of

[28] Adam knows that if he does not take into account the fact that he knows he is Adam, he would conclude that even if Adam makes the resolution, the probability that there will one hundred heads in a row is very small. If Adam does not makes the resolution and just throws the coin one hundred times and does not take into account the fact that he is Adam and only draws conclusions from what he knows about the physics of coin-tossing, he will also conclude that the coin is very unlikely to land

which world is actual. Once Adam has made a computation of how likely v is given the fact that he knows that in the actual world he is Adam, he does not get another chance to draw inferences from the fact that it is remarkable that he is Adam. If in the actual world, he is Adam, he will inevitably be Adam in the counterfactual world in which he made a different decision than the decision actually made.

Thus if Adam decides not to make the resolution and the coin did not land head one hundred times in a row, it would not have landed heads one hundred times in a row if he had not made the resolution and if Adam decides to make the resolution and the coin lands heads one hundred times in a row, it would still have landed one heads one hundred times in a row.

12 Conclusion

In this section we summarize our theory and mention some topics that need further investigation.

A simple elegant theory of anthropic reasoning can be developed (at least in the finite case) by making precise Bostrom's SSSA. An observer-moment z living a certain possible world wants to know the probability that a certain proposition is true. This proposition will be about which world is actual and about her identity within the actual world. Thus what z really wants to know is the probability that she belongs to a certain set A of centered worlds (where centered worlds are pairs consisting of a world w and an observer-moment o. If z is observer-moment o living in w, then we say z is the centered world (w, o).). She will begin with a nonathropic prior probability distribution P on a set W of possible worlds and then our theory will enable her to use P to construct a prior probability distribution on some subset R_z of the set $W*$ of possible centered worlds.

The assumption that we should first generate a nonanthropic prior based on information common to all observer-moments and then use a theory of anthropic reason to construct P_z might be doubted. Perhaps the prior P_z should be constructed in one step; maybe we should just require that if $y \in W*$, $P_z(y)$ should be proportional to the complexity of the shortest description of y in some canonical language; something similar to this is suggested by Hutter[11]. The problem is that it is difficult to know which language should be chosen as the canonical language. If a minimal description length approach is properly implemented, perhaps we might arrive at posterior probability estimates not that different from those suggested by our theory.

Assuming we are going to separate the nonanthropic and anthropic parts of the task of generating P_z,

heads one hundred times in a row.

we still have to decide before we worry about how the probability $P(w)$ that is given to a world should be split up among the observer-moments belonging to that world, whether we should use the SIA to revise P to give more probability to worlds with more observers or observer-moments. But the SIA is unmotivated. Wagering arguments that might be used in favor of the SIA do not work because they ignore the fact that decisions are made by individual observer-moments belonging to particular worlds. Decisions are not made by a committee of observer-moments who might belong to different worlds.

We next have to chose a reference class on a non ad-hoc basis and the simplest choice is $R_z = W*$. Both relative frequency and Dutch Book arguments can be given in favor of this $R_z = W*$. An other possible simple rule is $R_z = K_z$ but that rule does not allow observer-moments to use strictly anthropic information to revise their probabilities for (bare) possible worlds and if we are to link cosmology with observation, we will sometimes have to let strictly anthropic information affect judgements about which possible world is most likely to be actual. There is no obvious alternative to $R_z = W*$ if we want a simple rule that analyzes simple scenarios similarly and that arrives at reasonable results.

Next we have to specify a rule for constructing P_z on R_z. We have the limited indifference principle that if x and y are two observer-moments in the same subjective psychological state that live in the same world w, $P_z(x) = P_z(y)$. There are several arguments for this intuively natural principle, but the basic justification aside from the simplicity of the rule is that if we knew that w is actual, then both x and y exist and if x believes she is more likely to be x than y, then since x and y are in the same knowledge state, y will also believe that she is more likely to be x than y. But x and y have the same information available and they cannot both be more likely to be x than y.

It is a little more difficult to know how to allocate probability in the case where x and y know different things. We contend that if x and y belong to the same world, then $\frac{P_z(x)}{P_z(y)}$ should equal $\frac{I(x)}{I(y)}$ where I stands for the amount of information that an observer-moment can represent. We might also use another criterion: Let x and y be indecomposable observer-moments, which cannot meaningfully be represented as a union of smaller observer-moments. (If, for example, it takes a certain amount of time to acquire new information or a new belief or because conscious awareness requires a certain minimal amount of complexity, it might not be meaningful to split certain observer-moments into smaller parts.) If x and y live in the same possible world, then because they are both atomic, they should have the same prior probability.

Still other criteria might be explored. For example, we might suggest that the ratio of the prior probabilities of x and y should equal the ratio of the complexities of the shortest descriptions of the subjective psychological states of x and y.

Ultimately what matters is what works in practice and we can use fairly standard Bayesian methodology to test theories of anthropic reasoning and that includes testing rules for representing infinite scenarios as limits of finite scenarios. A big problem with testing theories of anthropic reasoning is determining prior probabilities that a theory is correct. More work needs to be done on that issue.

We might also evaluate theories of anthropic reasoning by how well they maximize expected epistemic utility [12]. But that computation requires combining the utilities of different observer-moments who live in the same possible world and may not be in the same subjective psychological state. It is not clear that when combining these utilities, we should use a simple summation or averaging rather than a weighted sum or average where weights are proportional to prior probabilities of observer-moments (however it is these very probabilities that we have difficulty estimating). Another issue is that we want a simple theory that applies to several different possible $W*$ and that means that we might have to combine the utilities of observer-moments who live in several different $W*$ and the question arises how that should be done.

In any case, our theory is a simple theory. Like any theory of anthropic reasoning, it will have some unfortunate consequences: We have to accept the validity of the Doomsday Argument. We also have to accept strange coincidences in the Lazy Adam scenario. We do not have to advocate Adam making an intuitively implausible decision, but we do have to accept a strange coincidence. But in some scenarios, our theory will lead observer-moments to make decisions that would have worse consequences than if all observer-moments agreed on a different theory of anthropic reasoning. However, it needs to be reiterated that decisions are made by individual observer-moments and not by committees of observer-moments (not even by a committee consisting of all the observer-moments that belong to a certain observer in a certain possible world).

Our theory does have some problematic consequences, but it is simple and natural and avoids many of the problems (such as vulneribility to Dutch Books) of some other theories.

References

[1] James 0. Berger, Josè M. Bernardo, and Dongchu Sun. The formal definition of reference priors. *Annals of Statistics*, 37(2):905–938, 2009.

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

[3] Nick Bostrom. Self-locating belief in big world: Cosmology's missing link with observation. *Journal of Philosophy*, 99(12), 2002.

[4] Nick Bostrom. Sleeping Beauty and self-location: A hybrid model. *Synthese*, 157(1), 2007.

[5] Nick Bostrom and Milan Cirkovic. The doomsday argument and the self-indication assumption:reply to Olum. *Philosophical Quarterly*, 53(210):83–91, 2003.

[6] Bradley P. Carlin and Thomas A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman, 2000.

[7] Brandon Carter. The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society of London A*, 310:347–363, 1983.

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

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

[10] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability.* Springer, 2004.

[11] Marcus Hutter. A complete theory of everything will be subjective. *Algorithms*, 3(4):329–350, 2010.

[12] Brian Kierland and Bradley Monton. Minimizing inaccuracy for self-locating belief. *Philosophy and Phenomenological Research*, 70(2):384–395, 2005.

[13] J. Leslie. Doomsday revisited. *Philosophical Quarterly*, 42(166):85–87, 1992.

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

[15] Christopher J. G. Meacham. Unravelling the tangled web: Continuity, internalism, uniqueness and self-locating belief. In Tamar Szabo Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology Volume 3*. Oxford, 2010.

[16] Sarah Moss. Updating as communication. *Philosophy and Phenomenological Research*, forthcoming.

[17] Ken Olum. The doomsday argument and the number of possible observers. *Philosophical Quarterly*, 52(207):164–184, 2002.

[18] Ken Olum. Conflict between anthropic reasoning and observation. *Analysis*, 64(1):1–8, 2004.

[19] Michele Piccione and Ariel Rubinstein. The absent-minded driver's paradox: Synthesis and responses. *Games and Economic Behavior*, 20(1):121–130, 1997.

[20] Eric Schwitzgebel. A phenomenal, dispositional account of belief. *Nous*, 36:249–275, 2002.

[21] Brian Weatherson. Should we respond to Evil with indifference? *Philosophy and Phenomological Research*, 70(3), 2005.