How To Infer Probabilities

Eissa Haydar*

August 15, 2024

Abstract

This paper presents a formal system of deductive reasoning on probability spaces. Using this, we present a solution to the Sleeping Beauty problem and analyze other attempts. We then take our learnings and deflate anthropic paradoxes such as the Doomsday argument.

1 Introduction

Suppose we had some discrete probability spaces, events, and correspondence relationships among them. What more can we infer? That is, how can we combine the information we know in order to derive new probability spaces? And, what probabilities can we conclude? We will generally refer to events and probability spaces together as *outcomes*. Suppose we had a probability distribution on outcomes A_1, A_2, and so on, each with respective probability p_1, p_2, etc. We denote this probability space with $\{A_1^{p_1}, A_2^{p_2}, \ldots\}$. Thus, a fair coin flip might be represented as $\{\text{Heads}^{1/2}, \text{Tails}^{1/2}\}$.

The main motivating example in this paper will be the Sleeping Beauty problem:

*Thanks to Matthew Adelstein for telling me about the problem, and to Ammar Ahmad, Gordon Belot, Ibrahim Haydar, Musa Haydar, Jim Joyce, Siddarth Namachivayam, Abigail Peacock, Juan Ruiz, and Andrew Vincent for helpful comments on drafts.

> SLEEPING: Sleeping Beauty goes to bed on Sunday knowing that a fair coin is flipped with the following effect. If it lands heads, she will awaken on Monday, and if it lands tails, she will awaken on Monday *and* Tuesday. Furthermore, she knows that upon awakening she will have no memory of ever having awoken before; thus each wake up is indistinguishable. Upon waking up, what credence (degree of confidence, represented as a probability) should she assign to the coin being heads?

It turns out that the confusion around this problem stems from confusion about the methodology that can be employed to infer the probability that Beauty sets her credence to be. In this paper, we dissolve this confusion by formally specifying how to infer probability spaces from the rules of the game, including what can be concluded from an application of the principle of indifference.

Let's denote the coin being heads H, the coin being tails T, the day of the week being Monday m, and the day of the week being Tuesday t. Beauty is thus trying to locate herself among (H, m), (T, m), and (T, t). But, to begin with, the only probability distribution she knows about is $\{H^{1/2}, T^{1/2}\}$. From the rules of the game alone, she cannot immediately answer.

Thus, in §2, we will set up a formal system of deduction which can be used to infer Beauty's credence. This will give us an answer of $1/2$. Importantly, we will distinguish between actual and local probabilities, where an actual probability of an outcome is the probability that it will occur, whereas a local probability is its subjective likelihood taking applications of the principle of indifference into consideration. In §3, we will show that Adam Elga (Elga 2000) was only able to prove that Beauty's credence should be $1/3$ by conflating actual and local probabilities. Others, such as Cian Dorr (Dorr 2002) and David Lewis (Lewis 2001) have also concluded wrongly due to this conflation, as we will demonstrate. In §4, we will consider the appeal to iteration that is often used to justify an answer of $1/3$, and show that it is actually an appeal to a different game. To conclude, we will spend §5 moving

beyond the Sleeping Beauty problem with our learnings. It turns out that all so-called 'anthropic paradoxes', such as the Doomsday argument, arise from uncareful treatments of the probability spaces under consideration.

2 Operating on Probability Spaces

2.1 Formalities

Here is the full set up of our system. A probability space with outcomes $A_1, A_2, \ldots$, with respective probabilities $p_1, p_2, \ldots$, is denoted $\{A_1^{p_1}, A_2^{p_2}, \ldots\}$. The order we write the outcomes makes no difference. If outcomes X and Y each occur exactly when the other occurs, we write $X \longleftrightarrow Y$ and say that X and Y *coincide*. If the same outcome exists twice in the same probability space, we may simply combine them, adding the two probabilities. That is, it is a rule in our system that $\{X^{p_1}, X^{p_2}, A_1^{q_1}, \ldots\} \longleftrightarrow \{X^{p_1+p_2}, A_1^{q_1}, \ldots\}$. We can also eliminate outcomes with probability 0: $\{X^0, A_1^{q_1}, \ldots\} \longleftrightarrow \{A_1^{q_1}, \ldots\}$. Take outcomes $X, Y, A_1, \ldots, B_1, \ldots$ with $p + q_1 + q_2 + \cdots = 1$, $t_1 + t_2 + \cdots = 1$ and $0 < p, q_1, q_2, \ldots, t_1, t_2, \ldots$. Our first rule of inference is *injection*:

(INJ) $X \longleftrightarrow Y, \{X^p, A_1^{q_1}, A_2^{q_2}, \ldots\} \vdash \{Y^p, A_1^{q_1}, A_2^{q_2}, \ldots\}$

Injection basically says that within a probability space, an outcome can be substituted for an outcome it coincides with. Take $\{A^{1/2}, B^{1/2}\}$ together with $B \longleftrightarrow C$. This can be read as half the time A occurs and half the time B occurs, and B coincides with C. This quite obviously implies that half the time A occurs while C occurs the other half, exactly when B occurs. This is what injection formalizes and generalizes: $\{A^{1/2}, B^{1/2}\}, B \longleftrightarrow C \vdash \{A^{1/2}, C^{1/2}\}$. Our second rule of inference is *scaling*:

(SCL) $\{\{B_1^{t_1}, B_2^{t_2}, \ldots\}^p, A_1^{q_1}, A_2^{q_2}, \ldots\} \vdash \{B_1^{t_1 p}, B_2^{t_2 p}, \ldots, A_1^{q_1}, A_2^{q_2}, \ldots\}$

Scaling says that if a probability space B has probability p in probability space A, then the outcomes within B can replace B in A if scaled by p. As an example, take the following

probability space: $\{A^{1/2}, \{B^{1/2}, C^{1/2}\}^{1/2}\}$. This can be read as half the time A occurs and half the time one among B or C occurs with equal probabilities. And, of course, this implies that half the time A occurs and B and C each occur one quarter of the time. Scaling formalizes and generalizes this: $\{A^{1/2}, \{B^{1/2}, C^{1/2}\}^{1/2}\} \vdash \{A^{1/2}, B^{1/4}, C^{1/4}\}$.

The formulas in our system are outcomes and coincidences between them. Call an outcome that is a premise or conclusion *actual*. We define the *actual probability* of some outcome X, or $P(X)$, to be the probability of X in an actual probability space which, up to injection and scaling, contains no more copies of X. If X has probability p in some actual probability space, then it is at least the case that $P(X) \geq p$. If we can derive two different actual probabilities for the same event, we call this a *contradiction*. To understand the sort of situation in which this system of deduction is useful, we will consider a few examples.

First, take that if a fair coin lands tails then another fair coin is flipped. This tells us two things. Firstly, we may begin with an actual probability space of a first fair coin flip. Letting subscripts denote the order of the coin flips, we denote this initial probability space $\{H_1^{1/2}, T_1^{1/2}\}$. We also know that if the coin lands tails, then another coin flip probability space will be actual. Thus, we have $T_1 \longleftrightarrow \{H_2^{1/2}, T_2^{1/2}\}$, and these two are our premises:

(P1) $\{H_1^{1/2}, T_1^{1/2}\}$

(P2) $T_1 \longleftrightarrow \{H_2^{1/2}, T_2^{1/2}\}$

Now suppose we ask for the probability that at least one coin lands heads. By basic probability, this is asking $P(H_1 \vee H_2) = P(H_1) + P(H_2) - P(H_1 \wedge H_2)$, and we have that H_1 and H_2 are exclusive, so the third probability is zero. Thus, it suffices to find the actual probabilities of H_1 and H_2. We can immediately deduce this for H_1 from our first premise, knowing that there are no more copies of H_1 to worry about in nested probability spaces (if we did not know this, we would use $\geq$ in place of $=$, but H_1 and T_1 are exclusive):

(C1) $P(H_1) = 1/2$ (P1)

We don't immediately have H_2 in an actual probability space. But we have T_1, which we know coincides with a probability space containing H_2. Thus, we inject:

(C2) $\{H_1^{1/2}, \{H_2^{1/2}, T_2^{1/2}\}^{1/2}\}$ (P1, C1, INJ)

And, finally, to get H_2 in an actual probability space, we can simply scale its probability by the probability of the space it occurs in:

(C3) $\{H_1^{1/2}, H_2^{1/4}, T_2^{1/4}\}$ (C2, SCL)

(C4) $P(H_2) = 1/4$ (C3)

And so we have that the probability of at least one coin landing heads is $P(H_1) + P(H_2) = 1/2 + 1/4 = 3/4$. This was, of course, very obvious from the set up of the problem—injection and scaling are very natural rules.

As a second example, take $\{A^{1/4}, \{A^{1/2}, B^{1/2}\}^{3/4}\}$ as our premise. This represents a case where there is a $1/4$ chance A is defaulted to, otherwise A and B are equally likely. Even though A is in an actual probability space, A is also a possible outcome of the nested probability space (A is not exclusive with $\{A^{1/2}, B^{1/2}\}$) and so we can only immediately conclude:

(P1) $\{A^{1/4}, \{A^{1/2}, B^{1/2}\}^{3/4}\}$

(C1) $P(A) \geq 1/4$ (P1)

But we can simply combine the primary and secondary chances of A by scaling:

(C2) $\{A^{1/4}, A^{3/8}, B^{3/8}\}$ (P1, SCL)

(C3) $\{A^{5/8}, B^{3/8}\}$ (C2)

(C4) $P(A) = 5/8$ (C3)

Note that our coincidence relation goes both ways. Is there ever a situation in which we might want to express some one-directional equivalent? Here is an example: consider a case

where events A and B occur with equal probability, $\{A^{1/2}, B^{1/2}\}$. Suppose that whenever B occurs, A gets thrown in for free anyways. Thus, A always happens, but B only happens half the time. How can we represent this? If we used $B \longleftrightarrow A$, we would be able to conclude both that $P(A) = 1$, which is correct, and $P(B) = 1$, which is incorrect. Thus, we may desire to say something like $B \longrightarrow A$. This would work, but it is not necessary. The two As occur under different conditions, and so it suffices to denote the first A_1, the second A_2, define A as $A_1 \vee A_2$, and proceed as in our other example with an or above. That is, our premises would be $\{A_1^{1/2}, B^{1/2}\}$ and $B \longleftrightarrow A_2$. This would allow us to conclude that $P(A_1) = 1/2$, $P(B) = 1/2$, and $P(A_2) = 1/2$, and thus that $P(A) = 1$.

2.2 Principle of Indifference

We must make one other formalization before we can properly discuss Sleeping. When Beauty tries to reason about which day of the week it is, she is not given a probability space. The view of time we will take in this paper is that it can be treated as a line (*timeline*), where points on the timeline are identified with the set of propositions which are true at them. Thus, an accurate clock exactly pins down the moment in time, and viewing a clock known to be accurate makes locating oneself on the timeline oneself easy. In general, one may only know so many relevant propositions at a given moment in time, and it may be the case that two or more points on the timeline fulfill all known restrictions and have no known probabilistic relationship to each other. At this point, one can do no better than thinking them both equally likely.

In Sleeping Beauty's case, if she we wakes up in the game and knows it to be tails, then she considers either wake up equally likely because they are equally good candidates: she has the same experience in either case. If, instead, the Monday wake up were four minutes, Tuesday were five, and Beauty had an accurate mental stopwatch, then, knowing tails, she would consider Monday and Tuesday equally likely for the first four minutes before considering Tuesday certain. Call the number of minutes Beauty knows to have passed m.

This difference in time distinguishes the wake ups to Beauty, and she is able to narrow her location on the timeline down to Tuesday once $m > 4$, as the proposition 'the wake up lasts at least m minutes' will only apply to Tuesday from then on. If Beauty knows that the coin is heads, she can locate herself to the Monday wake up, as that is the only candidate. She does not need to apply the principle of indifference in that case.

Naively, we may attempt to immediately apply the principle of indifference to the experience of waking up: there are three distinct and indistinguishable possibilities, (H, m), (T, m), and (T, t). Why not think them all equally likely? For two reasons. Firstly, these three events do not all occur on the same timeline, and so our principle of indifference on *time* does not apply. Secondly, we have relevant probabilistic information, namely about the coincidence between the wake ups that will be experienced and the result of the coin, and so we cannot apply a principle of indifference over the set of wake ups directly.

We will represent a principle of indifference as a probability space with the events labeled with the subscript 'ind'. It will be important to keep track of whether events in our space originated from a principle of indifference or not, as an application of the principle of indifference allows placing two events which coincide into the same probability space. If we conclude a probability of some outcome X from a principle of indifference, i.e. from X_{ind}, we call this the *local probability* of X. The local probability will depend on the other events under consideration, so if we concluded the local probability of X from a probability space containing outcomes, say, Y and Z as well, we would denote this $P_l^{X,Y,Z}(X)$. This can be read as the local probability of X among X, Y, and Z, and it will be equivalent to $P(X|X \vee Y \vee Z)$. Since we can always impose a trivial application of the principle of indifference in cases where we can locate ourselves with certainty, we will allow concluding local probabilities when we could have instead concluded an actual probability. In Beauty's case, the principle of indifference yields $T \longleftrightarrow \{(T, m)^{1/2}_{\text{ind}}, (T, t)^{1/2}_{\text{ind}}\}$.

Thus, we may now properly consider Sleeping, from the rules of which we have:

(P1) $\{H^{1/2}, T^{1/2}\}$

Furthermore, we know that the coin is heads exactly when the heads wake up occurs and the coin is tails exactly when both tails wake ups occur:

(P2) $H \longleftrightarrow (H, m)$

(P3) $T \longleftrightarrow (T, m)$

(P4) $T \longleftrightarrow (T, t)$

Finally, we have our principle of indifference:

(P5) $T \longleftrightarrow \{(T, m)^{1/2}_{\text{ind}}, (T, t)^{1/2}_{\text{ind}}\}$

Upon waking up, how likely should Beauty consider (H, m)? In order to answer this, we must deduce a probability space containing each wake up, and compare their relative probabilities. Since we are asking how likely Beauty should consider the possibility, we are asking about the local probability (that is, we are asking about how likely (H, m) is among, or given, $(H, m) \vee (T, m) \vee (T, t)$):

(C1) $\{(H, m)^{1/2}, T^{1/2}\}$ (P1, P2, INJ)

(C2) $\{(H, m)^{1/2}, \{(T, m)^{1/2}_{\text{ind}}, (T, t)^{1/2}_{\text{ind}}\}^{1/2}\}$ (C1, P5, INJ)

(C3) $\{(H, m)^{1/2}, (T, m)^{1/4}_{\text{ind}}, (T, t)^{1/4}_{\text{ind}}\}$ (C2, SCL)

(C4) $P_l^{(H,m),(T,m),(T,t)}(H, m) = 1/2$ (C3)

What if, instead, Beauty woke up and found out it was Monday? How likely should she think heads is now? Borrowing from above:

(C6) $\{(H, m)^{1/2}, (T, m)^{1/2}\}$ (C1, P3, INJ)

(C7) $P_l^{(H,m),(T,m)}(H, m) = 1/2$ (C6)

3 Thirders

3.1 Elga

At this point, we shall go through the proof that Beauty's credence is 1/3 in (Elga 2000).

Elga proves two things. I will denote Elga's probability as p. First, he proves that $p(T, m) = p(T, t)$ and then that $p(H, m) = p(T, m)$. He reasons that since these are exclusive and exhaustive as candidates for the wake up Beauty experiences, they must sum to 1, and thus each has a value of 1/3.

To prove that $p(T, m) = p(T, t)$, Elga appeals to the principle of indifference. They have equal probability given tails, and tails referring to the same event as 'the current wake up is a one of the two tails wake ups' means the probabilities must be the same. More formally, $p((T, m)|T) = p((T, t)|T)$ is the same as saying that $p((T, m)|(T, m) \vee (T, t)) = p((T, t)|(T, m) \vee (T, t))$ and so $p(T, m) = p(T, t)$.

To prove that $p(H, m) = p(T, m)$, Elga appeals to the fact that it does not matter when the coin flip occurs, and so Beauty may reason as if the coin flip has yet to occur. Of course, if Beauty wakes up and knows the coin has not yet flipped, she should think them equally likely. Thus, $p(H|\text{Monday}) = p((H, m)|(H, m) \vee (T, m)) = 1/2$, and so $p(H, m) = p(T, m)$.

At first glance, both of these are reasonable. But, upon closer inspection, we see that Elga was not careful. When he proves that $p(T, m) = p(T, t)$, he points out that they are both equally likely given tails. Thus, he proves only that their *local probabilities* are equal, and this is, of course, not surprising because of our principle of indifference—their local probabilities will be equal in any space containing both of them. That is, he proves that $P_l^{(T,m),(T,t)}(T, m) = P_l^{(T,m),(T,t)}(T, t)$. On the other hand, when Elga proves that $p(H, m) = p(T, m)$, he appeals to their actual probabilities being the same (i.e. that $P(H, m) = P(T, m)$), and thus proves that $P_l^{(H,m),(T,m)}(H, m) = P_l^{(H,m),(T,m)}(T, m)$. Consider (C6) and (C7) in the previous section, and compare those to Elga's proof, in which he uses that $p((H, m)|(H, m) \vee (T, m)) = 1/2$. The actual probabilities of $P(H, m)$ and $P(T, m)$ being equal does not, in general, mean

that their local probabilities will be, too, although it does mean that their local probabilities among themselves will be equal. Elga showed that $P_l^{(T,m),(T,t)}(T,m) = P_l^{(T,m),(T,t)}(T,t)$ and that $P_l^{(H,m),(T,m)}(H,m) = P_l^{(H,m),(T,m)}(T,m)$, and so he cannot simply combine results to conclude that their local probabilities are equal among (H,m), (T,m), and (T,t), or that $P_l^{(H,m),(T,m),(T,t)}(H,m) = P_l^{(H,m),(T,m),(T,t)}(T,m) = P_l^{(H,m),(T,m),(T,t)}(T,t)$.

These three wake ups are only truly exclusive and exhaustive when their local probabilities among themselves (i.e. $P_l^{(H,m),(T,m),(T,t)}$) are under consideration. And so Elga's proof does not hold. Elga conflates local and actual probabilities because he begins his assessment of the problem assuming that there is some actual probability space of the form $\{(H,m)^{p_1},(T,m)^{p_2},(T,t)^{p_3}\}$. However, no such space exists in the rules of SLEEPING, and we can only derive one by the use of the principle of indifference. On the other hand, we can immediately deduce the actual probabilities of each of these events from the rules of the game: $P(H,m) = P(T,m) = P(T,t) = 1/2$, because (T,m) and (T,t) coincide.

3.2 Dorr

Now we move to an analogy of Cian Dorr (Dorr 2002). He presents a variant case in which Beauty will awaken on Monday and Tuesday. In case of heads, Beauty's memories are merely delayed upon her Tuesday wake up. In case of tails, Beauty wakes up on Tuesday with no memory of ever having awoken before. Dorr argues as follows.

Upon waking up and before regaining her memories, Beauty's credence should be evenly distributed among the four possibilities. Suppose further that Beauty is able to know after, say, a minute that she has not had a 'flooding-back' of memories, and can thus rule out (H,t). Having no information which distinguishes between the remaining options, Dorr has her evenly redistribute her credence that (H,t) among the remaining options and ends with each at 1/3. Is Beauty correct to do this? We hold that Beauty does, in fact, have information which distinguishes between the remaining options in the form of the probabilistic structure of the game which prevented us from, for instance, simply applying the principle

of indifference over all wake ups in SLEEPING.

Dorr's case yields the following premises:

(P1) $\{H^{1/2}, T^{1/2}\}$

(P2) $H \longleftrightarrow (H, m)$

(P3) $H \longleftrightarrow (H, t)$

(P4) $T \longleftrightarrow (T, m)$

(P5) $T \longleftrightarrow (T, t)$

(P6) $T \longleftrightarrow \{(T, m)^{1/2}_{\text{ind}}, (T, t)^{1/2}_{\text{ind}}\}$

Note that though there are two wake ups given heads, Beauty can exactly locate herself after the flooding-back, and thus does not need a principle of indifference. We can now deduce the local probability when experiencing one of the three indistinguishable wake ups:

(C1) $\{(H, m)^{1/2}, T^{1/2}\}$ (P1, P2, INJ)

(C2) $\{(H, m)^{1/2}, \{(T, m)^{1/2}_{\text{ind}}, (T, t)^{1/2}_{\text{ind}}\}^{1/2}\}$ (C1, P6, INJ)

(C3) $\{(H, m)^{1/2}, (T, m)^{1/4}_{\text{ind}}, (T, t)^{1/4}_{\text{ind}}\}$ (C2, SCL)

(C4) $P_l^{(H,m),(T,m),(T,t)}(H, m) = 1/2$ (C3)

Prior to the flooding-back, there is also an application of the principle of indifference to the heads wake ups. Thus, the four evenly split credences Dorr mentions are local probabilities. To assume Beauty can simply redistribute over the remaining options is incorrect. But Dorr makes this assumption because he, like Elga, presupposes a probability space of the form $\{(H, m)^{p_1}, (H, t)^{p_2}, (T, m)^{p_3}, (T, t)^{p_4}\}$. If such a probability space was actual and we ruled out one of the four options, we truly could redistribute as Dorr suggests.

3.3 Lewis

Though David Lewis correctly held that the solution to SLEEPING is 1/2, he still considered Monday to be evidence. Though Elga correctly reasoned about it not mattering when the coin is flipped, Elga agrees with the claim that Monday is evidence (shifting Beauty's credence from 1/3 to 1/2), and so the goal of this subsection is to deflate this claim. In particular, the intuition for Monday as evidence is as follows. Given heads, Monday is certain, whereas given tails, Monday only has probability 1/2. Thus, by an application of Bayes' theorem, Monday is evidence for heads. The problem with this is that Monday does not, in fact, have a probability of 1/2 given tails. The actual probability of Monday is 1 given tails, whereas the local probability is 1/2 if we apply the principle of indifference.

Lewis holds that the solution to SLEEPING should be 1/2 because it was 1/2 prior to going to sleep, and Beauty knew she would certainly sleep and wake up. Thus, sleeping and waking up cannot be evidence for tails. But did Beauty not know, furthermore, that she would sleep and wake up *on Monday* in either case? This is why the distinction between local and actual probabilities is so important. If we reason using local probabilities here, we are purporting that the Monday wake up might have instead been a Tuesday wake up in case of tails. But this is nonsensical: the Monday wake up was certainly going to happen. If Beauty wakes up and is told it's Monday, her evidence is merely that she is experiencing a Monday wake up, which would have happened in either case. Furthermore, it was always going to happen that she would experience the Monday wake up then, on Monday. There is nothing chancy here, the Monday wake up never occurs at the exclusion of the Tuesday wake up, but reasoning using the local probabilities entails this.

4 Iteration

When introducing the intuition behind his position, Elga says: 'in the long run, about 1/3 of the wakings would be Heads-wakings' (Elga 2000). So, does this mean that we ought to

think Beauty's credence is 1/3? Consider the following case, for any natural number $n \geq 1$:

> n-Sleeping: n-Beauty goes to bed knowing that n fair coins are flipped. For every coin that lands heads, she will experience one 'heads' wake up, and for every coin that lands tails, she will experience two 'tails' wake ups. Of course, n-Beauty has no memories upon waking up, and each wake up is indistinguishable. Upon waking up, what credence should n-Beauty assign to the coin being heads?

Thus, 1-Sleeping is the same game as Sleeping. 2-Sleeping can be represented as follows, denoting the wake ups by whether they are heads wake ups or tails wake ups (H and T) and by their order, and taking a principle of indifference for each result of the coin flips:

(P1) $\{HH^{1/4}, HT^{1/4}, TH^{1/4}, TT^{1/4}\}$

(P2) $HH \longleftrightarrow \{(H,1)^{1/2}_{\text{ind}}, (H,2)^{1/2}_{\text{ind}}\}$

(P3) $HT \longleftrightarrow \{(H,1)^{1/3}_{\text{ind}}, (T,2)^{1/3}_{\text{ind}}, (T,3)^{1/3}_{\text{ind}}\}$

(P4) $TH \longleftrightarrow \{(T,1)^{1/3}_{\text{ind}}, (T,2)^{1/3}_{\text{ind}}, (H,3)^{1/3}_{\text{ind}}\}$

(P5) $TT \longleftrightarrow \{(T,1)^{1/4}_{\text{ind}}, (T,2)^{1/4}_{\text{ind}}, (T,3)^{1/4}_{\text{ind}}, (T,4)^{1/4}_{\text{ind}}\}$

Letting H represent that the current wake up is a heads wake up, we have that $P(H) = P(H,1) + P(H,2) + P(H,3)$. From our premises, we can derive that the local probability of heads among all of the wake ups is 5/12. Thirders would gloss this game as the same as 1-Sleeping, since, overall, 1/3 of the wake ups are heads wake ups.

A closed form for n-Beauty's credence is:

$$\frac{1}{2^n} \sum_{i=0}^{n} \binom{n}{i} \frac{n-i}{n+i}$$

The $1/2^n$ represents the probability of any particular result of the series of n coin flips. Then, we sum from 0 to n because there are $n+1$ possible proportions of heads wake ups to tails wake ups. For instance, in 2-Sleeping, we have three possibilities: all of the wake ups are

heads, all of the wake ups are tails, or 1/3 are heads and 2/3 are tails. For each possible proportion of wake ups, the binomial coefficient gives the number of times this proportion occurs among the 2^n possible results of the coin flips, and the fraction gives the proportion of these wake ups that are heads wake ups. So, for $n = 2$, we have:

$$\frac{1}{4}\left(\binom{2}{0}\frac{2}{2} + \binom{2}{1}\frac{1}{3} + \binom{2}{2}\frac{0}{4}\right) = \frac{1}{4}\left(1 \cdot \frac{2}{2} + 2 \cdot \frac{1}{3} + 1 \cdot \frac{0}{4}\right) = 5/12$$

Looking to the middle term, there is one outcome, HH, where 2/2 wake ups are heads wake ups, there are two outcomes, HT and TH, where 1/3 of wake ups are heads wake ups, and there is one outcome, TT, where 0/4 wake ups are heads wake ups. And each outcome has probability 1/4, which we distribute. We can prove that:[1]

$$\lim_{n\to\infty} \frac{1}{2^n} \sum_{i=0}^{n} \binom{n}{i} \frac{n-i}{n+i} = \frac{1}{3}$$

In other words, the local credence should actually be 1/3 in the case of infinite iteration, so perhaps this is why thirders appeal to iteration. But since the credence changes for each n, we know that for any $n \neq m$, n-Sleeping and m-Sleeping are disanalogous, and so the appeal fails.

We can explain the convergence of the credence more intuitively. Imagine Beauty is to walk along a tightrope, starting 0.5m before it. The tightrope starts with no length, but each time a coin is flipped, 1m is added if heads, and 2m if tails. Say that she takes 1m steps and that a coin is flipped at the beginning and at the moment each step touches down. Thus, no matter how the coin flips, the rope grows faster than she can walk along it. Let us ask her credence that she is currently on a heads rope at each step. As she takes more steps, the number of steps she has taken is less information. The first step is 50/50, and from then

[1] A full proof is beyond the scope of the paper. As a sketch, for any ϵ, one can take n so that the value is within ϵ of 1/3 by taking enough terms around the middle so that the sum of the binomial coefficient over those terms is at least $(1-\epsilon)2^n$ (for instance, by using tail bounds on the binomial distribution) while still taking few enough terms that $n - i/n + i$ term is close to 1/3 for each of them. Thanks to Matthew Harrison-Trainor for suggesting this approach.

on she only gets less certain, and therefore closer to 1/3, which is the best Beauty can do in Sleeping if she were only aware of the outcomes of the game on play, and not the structure of the game. In ∞-Sleeping, Beauty is missing an important piece of information: that she is currently in the game and will at some point no longer be in it. This allows her to reason about the fact that half of the time she plays Sleeping, the coin happened to land heads, even if more of the wake ups, over many plays of the game, are tails wake ups.

The appeal to iteration is often paired with an appeal to betting strategies, but such arguments do not work for Sleeping. Asking Beauty her credence that the coin landed heads *on this play of the game*, the answer is 1/2. Let us instead ask her to choose, and she must choose, one of heads or tails ('How did the coin flip?'). Is either answer better? That depends on our criterion of success. If she cares about the proportion of games she answers completely correctly on, over many runs of the game, then it does not matter what her default answer is. Indeed, if she always answers heads, then she will either be correct upon each wake up (on heads games) or incorrect upon each wake up (on tails games), and the same goes for always answering tails—in either case she will be correct on about half of the games. This gives us a good sense that her credence should be split when asked about how the coin flipped for the current run of the game. Similarly, if upon each wake up she were to stipulate that 'this is my first tails wake up', she would be correct for half the wake ups on half of the games in the long run, corresponding to the 1/4 local probability that we can derive. If Beauty instead cared about the proportion of wake ups during which she answers correctly, over all wake ups, over many runs of the game (i.e. she does not care about the current run of the game in particular), then of course she should answer tails (this will give her a 'score' of 2/3 on average). This would involve using the actual probabilities of the events along with the fact that T_1 and T_2 always coincide. But this is not what is being asked about in Sleeping.

When asked her credence that the coin landed heads, she only cares about locating herself within this particular game, and she thus defers to the local probabilities. In the case of

infinitely many coin flips, she can no longer reason about the particular game, as she will never leave it, and she is basically experiencing a wake up sampled randomly from the set of possible wake ups; so, the local probability of a heads wake up *is* 1/3 in that case. But we know that in SLEEPING the tails wake ups have a privileged relationship to each other—none can be experienced without the rest also being experienced—and none of the tail wake ups have this relationship to the heads wake up; the wake ups are thus not independent, and we should not think of them as being in the same probability space by default, as Elga does, when locating ourselves. We should not treat them as if one occurs to the exclusion of the other, unless discussing the entire set of heads or tails wake ups.

We can also illustrate why an appeal to the proportion of wake ups on play is wrong by taking an urn and placing green balls during heads wake ups and red balls during tails wake ups. The fact that red will have likelihood around 2/3 on a draw from the urn after many plays will not affect our credence that on any single run of the game we will be *placing* a green ball (as opposed to a set of red balls) *into* the urn. If we instead kept flipping coins as we drew more and more balls, then the case becomes like the tight-rope case above. Flipping one coin and drawing one ball, we should be 50/50 as to its color. From then on, we can only become less certain about the ball we draw.

5 Beyond Sleeping Beauty

Let us begin this final section by considering an example of Nick Bostrom. We quote his game exactly:

> INCUBATOR: Stage (a): In an otherwise empty world, a machine called "the incubator" kicks into action. It starts by tossing a fair coin. If the coin falls tails then it creates one room and a man with a black beard inside it. If the coin falls heads then it creates two rooms, one with a black-bearded man and one with a white-bearded man. As the rooms are completely dark, nobody knows his beard

color. Everybody who's been created is informed about all of the above. You find yourself in one of the rooms. Question: What should be your credence that the coin fell tails?

Stage (b): A little later, the lights are switched on, and you discover that you have a black beard. Question: What should your credence in Tails be now? (Bostrom 2013)

We will reason about this using the framework we present in this paper. Then we will consider the assumptions Bostrom makes in his analysis.

What probability spaces do we have? Call heads H, tails T, having a tails black beard B', having a heads black beard B, and having a white beard W. Firstly, the rules of the game give us $\{H^{1/2}, T^{1/2}\}$, $H \longleftrightarrow B$, $H \longleftrightarrow W$, and $T \longleftrightarrow B'$. If we know the coin landed tails, we should think either color equally likely by a principle of indifference. Thus, we have $H \longleftrightarrow \{B^{1/2}_{\text{ind}}, W^{1/2}_{\text{ind}}\}$. Note that we can conclude from these premises that the probability of having a black beard is 3/4. We were asked about the coin flip. Conditioning on being alive is analogous to conditioning on it being Monday in Sleeping—our evidence is only that something which would have occurred in either case occurred, and so we answer 1/2. The same reasoning goes for stage (b) in which we know our beard to be black. This was not, *a priori*, the most probable outcome, but it would have happened either way, and it is all that we know happened.

Bostrom hardly takes this possibility, which he calls 'naive', seriously, before presenting two solutions he wishes to take seriously and compare. They are both 'sampling' assumptions, which involve thinking oneself distributed over observers.[2] That is, like Elga's approach to the Sleeping Beauty problem, Bostrom presupposes a probability distribution of the form $\{(H, W)^{p_1}, (H, B)^{p_2}, (T, B)^{p_3}\}$. But to do so is incorrect. We can only derive such a space from the rules of the game (the coin flip) combined with the principle of indifference, and

[2]Bostrom technically only dubs one a sampling assumption and the other an indication assumption, but they both hold that observers should think of themselves as randomly selected from some set of observers.

thus we can only derive local probabilities from such a set. Bostrom wants to begin with such a distribution, and take actual probabilities. So, the only option is, truly, to assume oneself sampled, somehow, over observers, as he suggests. But this would involve thinking that we *literally* could have been someone else, that this probability space was actual, or that the rules of life, as compared to the rules of a game, entail such a space, independently even of applying the principle of indifference. If such a sampling assumption held, there must actually be something which corresponds to the probability space containing all of the outcomes. When we derive it, it comes from whatever we do know along the principles of indifference. When we assume it, we assume that there is something real about it—that there is something common to, or selecting, or discovering itself as, the observer—and moreso that this something predates the observer. This is the only way that our experience is to the exclusion of another experience. Instead, just like SLEEPING, our experience exists in a set, and it is this set which exists at the exclusion of another set of experiences, all of which occur.

The sort of reasoning Bostrom wants to study is where the fact of observation is relevant. But he does not understand the scope in which this applies. Consider the following case. Suppose we were to awake knowing ourselves to live in one among Universe 0, which has no people, Universe 1, which has person A, Universe 2, which has persons A and B. But, we do not know who we are. Bostrom would at least agree with us that Universe 0 is impossible. This is the scope that we hold the 'anthropic principle', or the relevance of the fact of our existence as observers, applies. We obviously cannot live in a world that has no people. But still, God is not distributing souls among probable bodies. Here, we have no probability spaces, except from our principle of indifference. Firstly, we consider any universe we can't distinguish from ours to be equally likely; $\{1^{1/2}_{\text{ind}}, 2^{1/2}_{\text{ind}}\}$. Furthermore, we should think that if we were in Universe 1, we'd certainly be person A, and if we were in Universe 2, we'd be one of persons A or B. But can't tell them apart. Thus, $1 \longleftrightarrow A$ and $2 \longleftrightarrow \{A^{1/2}_{\text{ind}}, B^{1/2}_{\text{ind}}\}$. These, and what we can conclude from them, are all we can derive. Thus, we may think A *a priori*

more likely, but it occurs with probability 1 regardless of the distribution, if there even is one, on Universes 1 and 2 (since it occurs for each of these). We can only derive Bostrom's conclusions with the sort of erroneous reasoning thirders use in the Sleeping Beauty problem, conflating actual and local probabilities. And the same goes for the 'paradoxes' Bostrom presents in general.

A view of experience which centers the 'self', as Bostrom's sampling assumptions do, is presumptious at best. We hold that the correct view is that experience is secondary. We exist, first and foremost, as complex physical beings, and our experience arises from it. Thus, the only thing we may conclude from our experience is that 'we' exist. Even with a strict 'reference class' (as Bostrom terms it; the set of other obervers we could have been) of other possible selves, times, or experiences, the idea that *we* could have been someone else is strange. What is the we that carries over between selves? Even Sleeping Beauty, who knows that she will experience a Tuesday wake up if the coin lands tails, should not think, knowing it to be Monday, that it somehow *could* have instead been Tuesday. The Monday experience, which occurs with probability 1 for heads and tails, necessarily involves Beauty experiencing in it. How strange would it to be for her to think: 'I am experiencing this, but I might have not been'. If she were not, then no Monday wake up would occur. If a Monday wake up occurs, then she will experience it.

This perspective also damns the Doomsday argument. We quote John Leslie's formulation:

> One might at first expect the human race to survive, no doubt in evolutionarily much modified form, for millions or even billions of years, perhaps just on Earth but, more plausibly, in huge colonies scattered through the galaxy and maybe even through many galaxies. Contemplating the entire history of the race - future as well as past history - I should in that case see myself as a very unusually early human. I might well be among the first 0.00001 per cent to live their lives. But what if the race is instead about to die out? I am then a fairly typical human

(Leslie 1990).

Leslie notes that 'the argument's underlying principle [is] that one should, all else being equal, take one's position to be fairly typical rather than very untypical' (Leslie 1990). Yes, with a principle of indifference, one should expect their position to be typical. This is why we should think that there is a 3/4 chance we have a black beard in INCUBATOR. But, upon finding out that we are within the first n of years of human history, we should think only that we know the universe must be so as to contain at least n people. Any universe with at least n people must have an initial n people. Finding out that we are within the initial n, we simply think: 'Ah yes, this was certain to happen in any universe with at least n people'. But we do not think any of them actually more likely.

References

Bostrom, Nick (2013), *Anthropic bias: Observation selection effects in science and philosophy*, Routledge.

Dorr, Cian (2002), "Sleeping Beauty: In Defence of Elga", *Analysis*, 62, 4, pp. 292-296, DOI: 10.1111/1467-8284.00371.

Elga, Adam (2000), "Self-Locating Belief and the Sleeping Beauty Problem", *Analysis*, 60, 2, pp. 143-147, DOI: 10.1111/1467-8284.00215.

Leslie, John (1990), "Is the End of the World Nigh?", *The Philosophical Quarterly (1950-)*, 40, 158, pp. 65-72, ISSN: 00318094, 14679213, http://www.jstor.org/stable/2219967 (visited on 08/11/2024).

Lewis, David (2001), "Sleeping Beauty: Reply to Elga", *Analysis*, 61, 3, pp. 171-76, DOI: 10.1111/1467-8284.00291.