Aaron Brown

Betting With Sleeping Beauty

Waking up to the probabilistic fairy tales we tell ourselves

Guess it came up 'tails'...
Ever thought about shifting to Saturday?
WIC WIT
SLEEPING BEAUTY
RUMPLESTILTSKIN
G E

Monday night poker at some guy's pad

The Sleeping Beauty problem is a paradox in probability theory, originally proposed by philosopher Arnold Zuboff. Sleeping Beauty is in a lab experiment on Sunday. A fair coin is flipped, but the result is not shown to SB. Instead, she is given a drug that puts her to sleep until Monday morning. If the coin was heads, the experiment ends Monday night. If the coin was tails, SB is given another sleeping drug that puts her to sleep until Tuesday morning, and wipes out her memory of Monday.

The question is, when SB awakens, what is her subjective probability that the coin flip on Sunday night was heads?

One argument says it is one-half. Sunday night, that was clearly the correct answer. SB has no new information when she awakens, since she knew she was going to wake up, so the probability must remain one-half.

Or you can argue the answer is one-third. There are four equally likely possibilities: the coin was heads and this is Monday, the coin was heads and this is Tuesday, the coin was tails and this is Monday, and the coin was tails and this is Tuesday. We can rule out the second, because if the coin had been heads SB would not wake up in the lab on Tuesday. Therefore, in only one-third of the remaining equally likely, exhaustive, and mutually exclusive possibilities was the coin flip heads.

The theory of everything

Before going into this more deeply, there is a practical application of this problem. Suppose two physicists propose different theories to explain the origin of the universe. In the first model, all the physical constants we observe today are explained. In the second model, there is a point early in the universe in which it was equally likely to take on the physical constants we observe today or a set of physical constants incompatible with matter forming. Assuming the two models are otherwise equal in parsimony, empirical evidence, and other factors, do we prefer the first model?

On first thought, the answer is "yes," since the first model explains more. But the Anthropic principle argues the models are equally good. If the second model was true and physical constants had taken different values, we would not have been around to notice. Since we know we exist, we know physical constants are compatible with human life, so whether a model explains this or not is irrelevant to its plausibility.

This is the same dispute underlying the Sleeping Beauty problem. The two models are equally plausible before consideration of values taken by physical constants. We can treat this like the coin flip on Sunday. Before any other evidence, we think there is a one-half chance the first model is correct. Observing the universe today is like SB awakening. If the second model is correct, there was only a one-half chance we would be here, so the person who argues for one-third in the Sleeping Beauty problem should argue that the fact we exist means the first model now has twice the plausibility of the second model. The person arguing for one-half says the models remain equally plausible.

Of course this precise example is artificial, but the philosophic issue is real. When deciding how much plausibility to attach to a physical model of the universe, do we consider how well it explains the universe in general, or how well it explains the universe conditional on us being around to observe it? This is not a purely theoretical dispute, one approach or the other could be a more reliable guide to inquiry, could lead to better progress in physics. This is, in principle, an empirical question. For example, we have discovered empirically that assuming simple mathematical laws can lead to productive guesses about the universe that are confirmed later by experiment. We can argue over why that's true or what it means, "Is God a mathematician?," but it is true (so far, anyway). The Anthropic principle has neither aided nor impeded inquiry to date, but its utility will ultimately be decided by practicing physicists, not philosophers.

I don't like the Sleeping Beauty problem myself, because it involves memory-erasing drugs. If we allow that, why not probability-distorting drugs that make people think fair coin flips have one chance in three of landing heads? Once you start monkeying with the brain, it becomes difficult to define rational belief.

The captain's paradise

The statistician and computer scientist Radford Neal has a better version, the Sailor's Child. A sailor has girlfriends in two ports, both of whom want to have children by him. He flips a fair coin to determine whether to impregnate one or both of the women. If the coin is heads, he flips again to determine which is to be the lucky girlfriend. If the coin is tails, both girlfriends get lucky. We assume the sailor carries out his plan successfully as directed by the coin flips, and to keep it simple neither he nor either of his girlfriends has any other children.

Years later, a child of this sailor knows only this information. He knows nothing about possible events in the other port. What should his subjective probability be that he has a half-sibling in the other port? We can argue that the question was determined by the first coin the sailor flipped, and it was equally likely to be heads or tails, so the answer is one-half. Or we can argue that there were four possible outcomes of the two coin flips, one of which the child knows is impossible (heads followed by a flip in favor of the girlfriend in the other port). In two of the three cases he has a half-sibling so the probability is two-thirds.

There is a hidden assumption in this argument. Why is it natural to assume SB bets only $1 when the coin comes up heads, but a total of $2 when it comes up tails?

One argument that many thirders find convincing is that SB or the sailor's child will lose money betting at even odds. Suppose every time SB awakens, she bets one dollar that the coin flip was heads. When it is heads, she wins $1 on Monday. When it is tails, she loses $1 on Monday and $1 on Tuesday. In the long run she loses unless she insists on 2-to-1 payout, implying her subjective probability that the coin was heads is one-third. Or suppose many sailors father children in the manner described above. Half of them will have children in both ports, one-quarter of them will have children only in port A. There will be twice as many sailors' children in port A with half-siblings in port B as sailors' children in port A without half-siblings. If I bet $1 at even odds with all the sailors' children in port A that they have a half-sibling in port B, I will win twice as many bets as I lose.

There is a hidden assumption in this argument. Why is it natural to assume SB bets only $1 when the coin comes up heads, but a total of $2 when it comes up tails? Suppose SB bets one chip on each awakening. If the coin flip on Sunday was heads, the chip is worth $1. If the coin flip on Sunday was tails, the chip is worth $0.50. Now she breaks even betting at 1-to-1 payout. In the first example we held the amount bet each day constant, in the second example we hold the total amount bet constant. If you want to use this argument to bolster the thirder position, you have to explain why it's correct to assume equal betting amounts each day.

In the sailor's child variant, suppose the Navy pays $10,000 to illegitimate children of sailors when they reach age 21, but the amount is split among all children of a single father. You approach a sailor's child and offer to bet her double or nothing on her Naval paternity payment, that her father's initial coin flip was tails. You will win two-thirds of your bets, but since you win $5,000 when you win and lose $10,000 when you lose, you will break even. From your point of view, of course, the probability of winning is two-thirds. But from the standpoint of a sailor's child, who bets only once and cannot establish probability by long-run frequency, the revealed subjective probability appears to be one-half, since she's willing to bet double or nothing and breaks even doing so.

Bruno's take

It was the great Italian mathematician Bruno de Finetti who made the betting odds a person would accept the fundamental definition of probability (this is the foundation of the Bayesian or subjectivist theory of probability). His favorite example was establishing the probability that life existed on Mars a billion years ago. It's hard to know what that probability is, or even what it means. But suppose an expedition will determine the answer tomorrow, and there is a security that pays $10 if the statement is true. There is some price (de Finetti claimed) at which you are indifferent between buying and selling that security. If it is $0.10, your revealed subjective probability that there was life on Mars a billion years ago is 1 percent. Saying you don't know a probability is saying you don't know what you think.

> **This is not a hair-splitting academic point. Outside of textbooks, important risk decisions involve potential outcomes that cannot be reduced to a single numeraire**

Suppose, however, that the expedition above financed itself by selling bonds denominated in Mars Expeditionary Currency (MEC), the currency that the colonists will use. One MEC is worth one dollar today. But if life is discovered to have existed on Mars one billion years ago, one MEC will be worth ten dollars. The existence of life improves the chance of valuable scientific and cultural discoveries, and also the chance that Mars can be made suitable for life. If you would pay $0.10 for the security that pays $10, you should be willing to pay 0.10 MEC for a security that pays one MEC. That has to be true because 0.10 MEC is worth $0.10 today, and one MEC if the security pays off will be worth $10. So in MEC, you think there is a 10 percent probability that life existed on Mars a billion years ago.

Once you realize that Bayesian probability depends on the numeraire used for the bet, there is no contradiction between SB having a one-half and a one-third subjective probability for the same event, depending on what's at stake. The dispute between halfers and thirders comes down to the numeraire used, and you can make a case for any probability at all by selecting the appropriate numeraire. For some reason, it seems obvious to some people that the numeraire has to be defined to have the same value every time SB wakes up, while it seems equally obvious to other people that the numeraire has to be defined to have the same value every time the experiment is run. In fact, neither position is obvious and many other positions are equally defensible. The key error in most analyses of the Sleeping Beauty problem is assuming that the answer has to be a single probability.

This is not a hair-splitting academic point. Outside of textbooks, important risk decisions involve potential outcomes that cannot be reduced to a single numeraire. De Finetti can't compute a probability if you tell him you will pay one apple for a security that pays one orange if there was life on Mars a billion years ago. He has to convert payment and payoff to the same units in order to divide to get a probability. Moreover, he needs the ratio of values to be the same in all possible outcomes of the bet. In short, he needs a universal absolute numeraire, something that doesn't exist any more than there is a universal absolute frame of reference in physics. What numeraire can give relative values to money and human life, to honor and sex, to excitement and God, to pleasure today versus happiness of your great-great grandchildren long after you are dead? What numeraire assigns the same value ratios to these and other things whether you are alive or dead, rich or poor, healthy or sick, loved or hated?

Options, numeraires, and alcoholics

This idea is well known in finance. The Black–Scholes option pricing formula works by a change of numeraire. Instead of trying to price an option in dollars, we price it in units of the underlying stock. In this numeraire it is locally riskless, so we can compute (under some assumptions) the probability of exercise from the option price. However, we label this the "risk-neutral probability" and know that it is distinct from what we call the "actual probability" as determined by long-run frequency of exercise. These are by no means the only two probabilities that can be attached to the event; we can use anything as a numeraire, and each numeraire will result in a different probability assignment.

This is not intended to be an attack on Bayesian probability theory, we'll see later that the same problem exists in a different form under frequentist theory. My claim (which I will not argue in detail in this article but have discussed elsewhere) is that if you want to use probability theory for real decisions outside a casino or textbook, you have to give careful thought to your numeraire. More important, you have to recognize that your numeraire will not cover all aspects of the problem and will not give identical value ratios in all possible outcomes. This has significant practical effect on actual risk decisions. Analyses that ignore the numeraire issue are usually, not just sometimes, deeply misleading.

Before leaving de Finetti, I'll mention another issue. Suppose SB is an alcoholic who never turns down a drink, but regrets drinking afterward. Her utility for a drink now is +8, but her utility for having a drink at any other time is negative the number of drinks squared. Consistent with that she will drink if alcohol is available now, but take steps to make alcohol unavailable in the future, say by never having it in her home and putting herself on the do-not-sell list at local bars and liquor stores. This is by no means unusual behavior; in fact, this sort of thing is far more common than preferences that remain constant. We offer this SB a choice between (1) one drink for sure, or (2) H drinks if the coin flip on Sunday was heads, zero drinks otherwise, or (3) T drinks if the coin flip on Sunday was tails, zero drinks otherwise.

SB reasons that (1) is worth +8 if the coin flip was heads, and net +7 if the coin flip was tails because then she will get a drink today for +8 but also a drink at the other awakening for −1. Expected utility is 7.5. (2) is worth +8H if the coin flip was heads and 0 otherwise. This has the same expected value as (1) if H = 15/8. (3) is worth 0 if the coin flip was heads and $+8T - T^2$ if the coin flip was tails (+T today but $-T^2$ for the other awakening). This has the same expected value as (1) if T = 3 or T = 5. Therefore, if we establish the break-even H and T for SB, we'll decide based on H = 15/8 that she thinks there is 8/15 chance of heads. Based on T we'll decide that SB thinks there is both 1/3 and 1/5 chance of tails, and in neither case will SB's probabilities of heads and tails add up to one.

Beliefs and preferences

An ironic illustration of this kind of behavior is the experience with voluntary gambling restrictions. In some jurisdictions, casinos are required to offer bettors cards with preset loss limits, such as $100 per month. The programs are entirely voluntary for the bettors, and the bettors select the limit. Casinos are required to refuse bets from these bettors except through their cards. The bettors who sign up for these programs on their own are energetic and inventive in cheating on them, to the point that the schemes are ineffective. This makes is nearly impossible to define consistent sets of beliefs and preferences that explain casino betting behavior. If you can't describe risk-taking in a casino, with easily established objective probabilities and all bets and payoffs in cash, you can't describe it anywhere. Some people try to get around this by labeling the cheating bettors "problem gamblers" with defective decision-making capacity, but this is exactly how everyone behaves, in and out of casinos.

This does not mean I reject the idea of subjective probabilities. If I see you bet one apple against one orange on some event, it is not enough information to determine either your beliefs or preferences. If I also see you trade an orange for two apples I can state (a) you value an orange at two or fewer apples, and (b) you think you had at least one chance in three of winning the bet. I have to observe both gambling and trade to measure your beliefs and preferences, or in other words, your subjective probability distribution and your utility function. However, there are important situations in which people's behavior cannot be explained by any reasonable assignment of beliefs and preferences, and other situations in which many different assignments are consistent with observed behavior. Insisting that SB must have a single subjective probability, much less one that can be determined by objective argument, puts you very high up an ivory tower.

If you can't describe risk-taking in a casino, with easily established objective probabilities and all bets and payoffs in cash, you can't describe it anywhere

If we move from Bayesian to frequentist theory it might seem that we can dispense with all the complexities of SB's utility function and settle the issue objectively by counting results. The tricky frequentist detail is events must be embedded in a series in order to assign probabilities. When a frequentist says a fair coin has a 50 percent chance of landing on heads, she means if you give her any confidence level and error bound, she can give you a number of coin flips such that the probability of having an observed frequency within the error bound of 50 percent is greater than the confidence level. For example, if you want to be 99 percent sure of a frequency within 1 percent of 50 percent (that is, greater than 49 percent and less than 51 percent), you need to flip at least 16,510 times.

That definition is clear for coin flips, but what about non-repeatable events? SB and the sailor's child are asked about the probability of coin flips that have already happened, not the frequency of an outcome of a series of future events. In order to determine a frequentist probability, we need to embed the situation in a hypothetical series of indistinguishable repeated events.

This is easy for the halfer position. We imagine many Sleeping Beauty experiments. If SB always guesses the coin flip came up heads, she will be right for half the coin flips and wrong for half, in the frequentist sense defined above.

Exotic probabilities

The thirder wants to argue that if SB always guesses heads she is right on one-third of her awakenings and wrong on two-thirds. If I flip a biased coin with a one-third probability of landing heads, it takes 14,721 flips to be 99 percent confident that the frequency of heads will be within 1 percent of one-third. But if SB awakens 14,721 times, there is only a 97 percent chance that the frequency of heads is within 1 percent of one-third. So at the very least, SB saying the probability of heads is one-third means something different from the probability of a biased coin that everyone understands.

It's easier to see this effect with fewer awakenings. If I flip the biased coin twice there is one chance in nine I will get two heads, four chances in nine I will get one head, and four chances in nine I will get no heads. But if SB awakens twice there is one chance in four it will be heads both times (two experiments, both flip heads), one chance in four it will be heads one time (two experiments, heads the first time, tails the second), and two chances in four it will be heads neither time (one experiment, tails). There is no assignment of independent probabilities to the individual awakening events that gives this distribution of outcomes, therefore it's not clear how to define the probability of an individual event. Moreover, the expected number of heads in these two awakenings is three-quarters, not the two-thirds you would think from two repetitions with one-third probability each time. Expected values should add, even for non-independent events.

None of this proves the thirder position is incorrect from a frequentist perspective, just that the proposed one-third probability does not obey the normal frequentist rules. There's no consensus method for stating a frequentist probability that has been defined by embedding in a hypothetical series of dependent events. I think people often overlook these difficulties because they think of a series in which frequency exactly matches probability, for example two experiments with one head, one tail, and three awakenings. There's no doubt in that case that one-third of the awakenings are in an experiment in which the coin landed heads. But that's just a statement about frequencies of outcomes given frequencies of inputs, there is no randomness involved, hence no probabilities.

Despite these difficulties, it is true that for any level of confidence and any error bound around one-third, a thirder can name a number of awakenings such that the probability of the frequency of heads being within the error bounds is greater than the confidence. So it's consistent for SB to claim a frequentist probability of one-third, even if it is a slightly exotic probability. The question is whether SB is required to claim this probability, in other words, if the natural way to imagine repeating the event is to repeat awakenings rather than to repeat coin flips.

We immediately run into a problem with Sleeping Beauty because we wipe out her memory, so it's hard to come up with a practical use of the probability estimate

Practice makes perfect

I'm interested in a practical answer, so we have to discuss the purpose of estimating the probability. We immediately run into a problem with Sleeping Beauty because we wipe out her memory, so it's hard to come up with a practical use of the probability estimate. Let's use Neal's Sailor's Child instead. Suppose that the coins in the sailor's realm are all biased, half of them flip heads one-third of the time, the other half flip heads two-thirds of the time. This does not change the problem, because the unconditional probability of a coin flipping heads is still one-half (we assume the sailor chose his coin at random). We don't care about the second flip of the coin (assuming the first flip was tails) because we assume the assignment of ports to heads and tails was random (Neal makes a different assumption, but it amounts to the same thing for this purpose).

Someone tells me the story and gives me the coin. Having no other information, I clearly believe the chance of heads on a subsequent flip is one-half. I happen to know the sailor's child however, and also know that he is a thirder. I take the coin to him, and ask him what he thinks the probability of the coin landing heads on a single flip is. He answers 13/27. To see this, imagine 12 sailors, six with 1/3 coins and six with 2/3 coins. Two of the six sailors with 1/3 coins flip heads, one has a child in port A. The other four sailors with 1/3 coins all have children in port A. Four of the six sailors with 2/3 coins flip heads, two have a child in port A. The other two sailors with 2/3 coins flip tails and have children in port A. Therefore in port A (and port B), five of the nine children had fathers who flipped a 1/3 coin. So if I pick one of the nine children at random, there is a 5/9 chance his father's coin was a 1/3, and a 4/9 chance it was a 2/3. The probability that the coin will land heads on the next flip is 13/27. Although there was only one sailor and only one or two children, enumerating all the cases like this gives us the correct probability.

I have exactly the same information as the sailor's child, yet I think the probability of heads on the next flip of this coin is one-half and he thinks it's 13/27. Who is correct? It all depends on how I got the coin and how I chose him. Once I disclose these things to him, the probability is always the same for both of us. In this respect the problem is similar to the Monte Hall problem in that the correct answer depends on what might have happened in other circumstances.

Suppose the reason I got the coin has nothing to do with whether there are one or two sailor's children or if there is a single child, what port he or she is in. Perhaps I was chosen to get the coin by lottery, or perhaps all coins used in paternity decisions are shipped to me by law. In this case there is no information in my receipt of the coin to change the probability of heads from one-half.

I said I happened to know the sailor's child, which implies I know of only one sailor's child. If I know everyone in both ports, then knowing there is only one sailor's child means I know the coin flipped heads, which makes my conditional belief about its probability of heads on a subsequent flip 5/9, not one-half or 13/27. If I tell this to the sailor's child, he has the same probability estimate.

Instead suppose I know everyone in one port, and no one in the other. That means if the original flip had been heads, half the time I would not

I'm interested in a practical answer, so we have to discuss the purpose of estimating the probability have known any sailor's child. Therefore, the fact that I do know a sailor's child means there is only 1/3 chance that the original flip was heads, so now the 13/27 figure is correct. If I had not known a sailor's child, I would have computed a probability of 14/27 for the next flip.

Another possibility is that I know the first-born child of every sailor, but no subsequent children. In this case there is no information in my knowing the sailor's child, I will always know exactly one sailor's child, so the answer is one-half. Of course, I can make other assumptions about my knowledge of potential sailors' children and get other answers. Also, if I'm using the probability for some reason other than betting on a subsequent flip of the coin, or if I make some other assumption about the distribution of probabilities of heads, I can get still other answers.

Too many answers spoil the broth

Given that there are lots of frequentist answers for the probability that the sailor's child has a half-sibling in the other port, depending on what hypothetical series of other events we embed the coin flip in and also what the probability will be used for, is any one of them a natural or neutral answer that does not depend on specific assumptions? I think the answer is "no." Once you go beyond coin flips and similar constructions, there are always multiple reasonable ways to set frequentist probabilities.

> Is any one of them a natural or neutral answer that does not depend on specific assumptions? I think the answer is "no."

This does not mean frequentist probabilities are meaningless. One key to evaluating frequentist methods is how often they are correct in the long run. For example, if a frequentist statistician sets many 95 percent confidence intervals in her career, the true parameter should be in the interval close to 95 percent of the time. In practice, however, this criterion is necessary but not sufficient. What if the 95 times she's right it's about things we knew already or were unimportant, while the 5 she got wrong were things we didn't know and were crucial? Just as a Bayesian needs a numeraire to define a probability, a frequentist needs some kind of weighting scheme on predictions. In the frequentist case it's not necessary for the mathematical definition of a probability, but it is necessary to turn a frequentist probability into a useful input to a decision problem. When choosing among different frequentist probability estimates, all of which can be correct, you should consider the long-term accuracy of the method weighted by utility rather than a raw accuracy score. Bayesians require a numeraire and a prior, after which they can define a unique probability. Frequentists can assign many probabilities to the same event, they require numeraires and priors to pick which probability to use.

> The utility of a frequentist probability claim depends on the vigor and sincerity of the falsification efforts

Consider three frequentist statisticians asked if a certain levee will be breached within the subsequent 12 months. The first statistician does an extensive study using engineering models and historical data and rejects at the 1 percent level the hypothesis that the levee will fail. The second statistician notes that fewer than one levee in 100 fails in any given year using data over all levees in the country over the last century. He also rejects at the 1 percent level the hypothesis that the levee will fail. The third statistician puts 99 true statements into a hat, and also the statement that the levee will hold. He draws a statement at random and gets the levee one. He notes that the probability of drawing a false statement out of a hat with at least 99 percent true statements is at most 1 percent, so he also rejects the hypothesis that the levee will fail at the 1 percent level.

All three of these methods are equally correct in theory. All of them will lead to long-term error rates, when a rejected hypothesis turns out to be correct, under the stated 1 percent level. Actually, it's worse than that. The first will have the correct error rate only if the statistician is highly competent, the second requires minimal talent, while the third will work regardless of the statistician's skill. But in terms of utility, the first approach is useful, the second has some small value, and the third is worthless. The utility of a frequentist probability claim depends on the vigor and sincerity of the falsification efforts, not on the significance level, yet only the latter is routinely required for journal articles. Putting things in other terms, a Bayesian does all the work before she gets the answer, a good frequentist does all the work after he gets the answers.

After coming all this way, what should we tell SB? Clearly, not to sign up for dangerous and pointless medical experiments (how did it ever get past the human experimentation committee in the first place?). Another useful piece of advice is not to bet with people who give you memory-erasing drugs. That could be a very expensive practice. In terms of probabilities, listen to Buddha. "Do not dwell in the past, do not dream of the future, concentrate the mind on the present moment." All we have to add is especially don't worry about hypothetical states that might have occurred in the past, but if so your memory was wiped out, or might occur in the future, in which case your current memories will be wiped out. Then the decision is easy, the probability doesn't matter.