*Sleeping Beauty: why violations of probability laws are ‘Deal’ breakers*

Beauty (see [3]) is a rational agent participating in an experiment in which a coin is tossed on Sunday night. She is awakened Monday morning, asked her credence in *heads*, told what day it is, asked again her credence in *heads*, and then debriefed. If the coin landed heads, that’s the end of the experiment; she’s free to go. If *tails*, however, Beauty is given a drug that erases all memory of the experiences she has had since awakening and which puts her back to sleep. Then on Tuesday morning she is again awakened, asked her credence in *heads*, told what day it is, asked again her credence in *heads*, and then debriefed. That’s the end of the experiment and she’s free to go. Beauty knows all of this in advance; the problem is what her initial personal credence in *heads* should be on Monday morning. A *halfer* says one-half. A *thirder* says one-third. For halfers, there is a second issue: what Beauty’s credence in *heads* should be after learning *Monday.*

Philosophers have gotten fairly worked up about Sleeping Beauty, and have spilled more than ink in so doing. Most curious are those who have proposed sacrificing laws of probability, as applied to credences, on her behalf. Jacob Ross for example claims in [13] that there is a tension between the one-third solution and countable additivity of credences. Several halfers ([4], [11] and [16]) meanwhile have advocated violations of the law of total probability for credences.

I see no reason to take serious issue with the naive and majority view (thirding; see [3], [7], [12], [15], etc.). However, there do seem to be fairly natural situations in which an agent can maximize utility by acting in concert with the minority view (halfing; see [4], [6], [9], [10], [11], [16], etc.). So while I will give a brief argument for thirding, that isn’t my focus. My primary task is rather to defend the laws of probability as applied to *both solutions.* My targets here are not lacking in philosophical subtlety; however, I believe they’ve gotten their mathematics wrong.

## *1. Thirding and countable additivity.*

I will begin this section with a (merely representative) positive argument for thirding, then address some recently raised doubts. Most halfers, including David Lewis [9], have maintained that Beauty gains no information from Sunday night to Monday morning. Many thirders, for example Horgan [7], think that Beauty does gain information, and that this is the basis for her change in credence. I hold with those who maintain that Beauty *loses* information. Beauty hasn’t merely lost track of the time; that much is normative in rational agency. Rather, she’s lost track of whether she’s done all of this before, in exactly this way. That’s not normative, and it’s relevant to *heads.*

According to the bounded martingale stopping theorem, if a non-information losing rational agent has expected credence $r$ in $A$ at a random, almost surely finite future time depending only on gathered evidence (i.e. a time at which the agent has it in their power to say *stop*), then they have credence $r$ in $A$ *now.* When time and information are discrete, this can be viewed as a continuous version of the:

> *Law of total probability.* If $A$ is a measurable event and $B_1, B_2$ partition the sample space then $P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2)$.

A vaguer version with fewer (too few; see [1]) hypotheses was formerly popularized in philosophical circles by Bass van Fraassen under the name *reflection principle.* I propose here an:

> *Inverse reflection principle.* If a non-information gaining rational agent has lost evidence expressible as conditionalization on a countable partition $\mathcal{E}$, their credence in $A$ now is equal to the expected value $E(A|\mathcal{E})$ of their prior credence in $A$.

Regarding the non-information gaining hypothesis, what is important is that the agent's gained information $G$ yields no information about $A$ not already given by conditionalization on $\mathcal{E}$. That is, $P(A|E \wedge G) = P(A|E)$ for every $E \in \mathcal{E}$. This is important below, as otherwise the argument given for thirding appears to depend crucially on whether or not Beauty is debriefed on Monday before her memory is wiped. (It should not.)

The argument now runs as follows. Let $x$ be Beauty's credence in *heads* during an awakening. I will assume that Beauty accepts Elga's "restricted principle of indifference" [3], which says that $P(\textit{Monday tails})$ and $P(\textit{Tuesday tails})$ share a common value, here $\frac{1}{2}(1-x)$. She's lost information potentially relevant to *heads.* By the inverse reflection principle, her credence in *heads* should be equal to its expected prior value. If it's Monday, the prior value is that of Sunday night–namely $\frac{1}{2}$. If it's Tuesday, the prior value is that of moments before her memory was erased on Monday–namely zero. Beauty's credence in *heads* now should therefore be the expected value of these prior credences, given by the weighted average $\frac{1}{2} \cdot P(\textit{Monday}) + 0 \cdot P(\textit{Tuesday}) = \frac{1}{2}(\frac{1}{2}(1+x)) = \frac{1}{4} + \frac{1}{4}x$. On the other hand her credence in *heads* is $x$, so we get $x = \frac{1}{4} + \frac{1}{4}x$, which has solution $x = \frac{1}{3}$.

So much for the positive argument; I move now to the defense. Ross in [13] claims that there is a tension between the one-third solution and the principle that rational credences should be countably additive. His argument has some plausible features, but unfortunately involves an experimental model that cannot be faithfully implemented at any nomologically accessible world. At our world, in particular, the model (and hence the argument) isn't any good, and, moreover, for reasons long familiar to practitioners of stochastic analysis.

Ross begins by defining a general "Sleeping Beauty problem" to be "a problem in which a fully rational agent, Beauty, will undergo one or more mutually indistinguishable awakenings..." where the number of such awakenings is a function of a discrete random variable into a set $S$ of hypotheses. He then claims that thirders are committed to the following "indifference principle":

> *Finitistic Sleeping Beauty Indifference (FSBI).* In any Sleeping Beauty problem, for any hypothesis $h$ in $S$, if the number of times Beauty awakens conditional on $h$ is finite, then upon first awakening, Beauty should have equal credence in each of the awakening possibilities associated with $h$.

*FSBI*, together with some additional premisses (details omitted), leads to the following:

> *Generalized Thirder Principle (GTP).* In any Sleeping Beauty problem, upon first awakening, Beauty's credence in any given hypothesis in $S$ must be proportional to the product of the hypothesis' objective chance and the number of times Beauty will awaken conditional on this hypothesis.

The pathological example to follow shows that *GTP* is in conflict with:

> *Countable Additivity (CA).* For any set of countably many centered or uncentered propositions, any two of which are incompatible, rationality requires that one's credences in the propositions in this set sum to one's credence in their disjunction.

Here is the example.[2]

> *Sleeping Beauty in St. Petersburg (SBSP).* Let $S = \mathbf{N}$ and suppose that Beauty awakens $2^X$ times, where $X$ is a random variable with $P(X = n) = 2^{-n}$, $n \in \mathbf{N}$.

If Beauty subscribes to *GTP*, then in *SBSP* it would appear that she must assign equal credences to the exhaustive and mutually exclusive assertions $X = n$, which violates *CA*.

It's easy to see how this argument could become popular–it's counterintuituve and seemingly plausible at worlds where a literal reading of *FSBI*, indeed an apparent premiss of thirders, can be rationally defended. At our world, however, thirders can't rationally defend such a reading, as they know that with non-zero probability, Beauty will die (for example) between tails awakenings. That thirders have failed to be explicit about mortal contingencies (which would rightly be perceived as tedious) is clearly innocent. Neglecting such remote possibilities simplifies the thirder model, and of course their effect on Beauty's credence function vanishes with increasing remoteness. It's well understood, however, that

[2] This is a nontechnical exposition, but for those who know of such things, the original problem corresponds to a positive recurrent Markov chain, the one-third solution deriving from its stationary probability measure, while *SBSP* is a null recurrent chain, for which all stationary distributions are infinite. That null recurrent chains are essentially useless as models for nomological phenomena is generally accepted; see [5], in particular Example 6.1.1 and Section 6.9 (chapter summary). A representative passage: "There is not much engineering significance to null recurrence; it is highly sensitive to modeling details over the entire infinite set of states. One usually uses countably infinite chains to simplify models; for example, if a buffer is very large and we don't expect it to overflow, we assume it is infinite. Finding out, then, that the chain is transient or null recurrent simply means that the modeling assumption is not very good."

to neglect mortality in the class of examples to which *SBSP* belongs puts the model out of touch with reality. Indeed, *SBSP* is so sensitive to conditions of implementation that acknowledging any fairly time-stationary prospect for mortality whatsoever (that Beauty might be transformed into a marble bust of Pascal by unfortunate quantum effects will do) eliminates any perceived conflict with *CA*.

Some details: when Beauty is explicit about mortality, she employs not *FSBI* but:

> *Sleeping Beauty Partiality (SBP).* If the number of times Beauty awakens is $M$, then for any hypothesis $h$ in $S$, upon first awakening, Beauty's credence in the $k$th awakening associated with $h$ should be proportional to $Ch(M \geq k|h)$.

Here $Ch(\cdot)$ is objective chance. *SBP*, together with other plausible hypotheses (first night mortality rates independent of $h$ and credences in first $h$ awakenings proportional to $Ch(h)$), implies that Beauty's absolute credence in the $k$th awakening associated with $h$ should be

$$P(h \wedge k) = \frac{Ch(h) \cdot Ch(M \geq k|h)}{\sum_{j \in S, l \in \mathbf{N}} Ch(j) \cdot Ch(M \geq l|j)} = \frac{Ch(h) \cdot Ch(M \geq k|h)}{E(M)}.$$

Summing over $k \in \mathbf{N}$, Beauty's credence in $h$ should be

$$P(h) = \frac{Ch(h) \cdot E(M|h)}{E(M)}.$$

Define the *fidelity* of an implementation to be $Ch(N = M)$, where $N$ is the number of times Beauty is told she will awaken and $M$ is the number of times Beauty does awaken. The *variation distance* between two discrete credence functions $R$ and $Q$ on a set $S$ is the quantity

$$v(R, Q) = \frac{1}{2} \sum_{h \in S} |R(h) - Q(h)|.$$

We've seen that Beauty's credence in $h$ is $P(h) = \frac{Ch(h) \cdot E(M|h)}{E(M)}$. Therefore, if $E(N) < \infty$ then as fidelity approaches 1 Beauty's credences converge in variation to the distribution $Q(h) = \frac{Ch(h) \cdot E(N|h)}{E(N)}$. On the other hand if $E(N) = \infty$ then as fidelity approaches 1 Beauty's credence in $h$ approaches zero for every $h$, and her credences diverge in variation. In the former case, the distribution $Q$ constitutes a stable solution to the problem. In the latter case, there is no stable solution, meaning that individual agents cannot avoid mortality estimates in establishing or even approximating their credences.

This is consistent with *CA* and recovers the one-third solution, modulo agreement that thirder Beauty may assign credence $\frac{1}{3} + \epsilon$ to heads for a smallish $\epsilon$. Diehard thirders who insist on $\epsilon = 0$ meanwhile may indeed run afoul of *CA* per [13]. Where rational, however, such thirders inhabit nomologically (at least) inaccessible worlds.[3]

## *2. Halfing and the law of total probability*

Lewis [9] bases his halfing scheme on the following premiss: (L1) Only new relevant evidence, centred or uncentred, produces a change in credence; and the evidence $(H1 \vee T1 \vee T2)$ (sic) is not relevant to HEADS versus TAILS. Here of course $H1$ is *Monday heads*, $T1$ is *Monday tails* and $T2$ is *Tuesday tails.* As I explained in the first section, I don't think this premiss is correct. Beauty *loses* information, that information is relevant to heads, and Beauty must, as it turns out, have credence $\frac{1}{3}$ in *heads.* However, I don't think that makes halfing uninteresting. Indeed, there are natural protocols governing the allocation of utility that call for Beauty to act as if her credence in *heads* were $\frac{1}{2}$. The distinguishing feature of these protocols is that only one *tails* awakening is significant to Beauty's well-being.

For example, Bostrom [4] proposes a thought experiment (*Beauty the high roller*) where bets are offered to Beauty on Mondays only. (Suppose fake bets are offered on Tuesdays.) Given such a protocol, Beauty should behave as a halfer in the style of Hawley [6], who assigns *Monday* probability 1 conditioned on *tails.* Less arbitrarily, if one real bet were to be offered, but on *Monday* or *Tuesday* with equal likelihood conditioned on *tails*, Beauty should act in concert with Peter J. Lewis [10]'s *quantum Sleeping Beauty* interpretation. Shaw [14] introduces (in essence) bets that Beauty can make only (and only *once*) by agreeing to them during each awakening of the experiment. Such a protocol is consistent with David Lewis [9]'s answer (familiar to any statistician) to tails world oversampling: sample weight dilution of the *tails* awakenings.

[3] Infinite expectation (or something similar) appears to play a hidden role in some other thought experiments against thirding, for example, Bostrom [4]'s *Presumptuous Philosopher* and Meacham [10]'s *Many Brains Argument.* In Bostrom's example (Meacham's is similar) we are asked to imagine learning that the number of observers $X$ in the universe is either $M$ or $N >> M$, and are told that it would be "presumptuous" to claim that the latter was $\frac{N}{M}$ times likelier based solely on thirder reasoning. Such a claim depends on the hypothetical prior probability of $N$ observers being equal to that of $M$ observers, and where this artificial condition is satisfied the case for presumption is weak (if there are such worlds in equal numbers then "presumptuous" observers are vindicated in precise proportion to their "presumption"). In less artificial cases, the hypothetical prior probabilities $(h_k)$ of $k$ observers should meet the finite expectation requirement $\sum_k kh_k < \infty$, so that typically $h_k$ decays faster than $\frac{1}{k}$ and the philosopher doesn't even have the attitude deemed presumptuous, indeed asymptotically favors the smaller value $M$. Some philosophers have done a good job of tracing paradox to the use of models involving infinite expectations (e.g. David Chalmers [2]'s outstanding analysis of the Two Envelopes problem), but have been hesitant to characterize such use as *misuse.* However, if sincerely practicing scientists never require models assigning infinite expectation to physical quantities, the burden of proving their relevance to rational agency plainly lies with those whose appeals do.

Of course, one can't just pick any protocol whatsoever. (This could spawn fourthers, fifthers, $\frac{e}{\pi}$thers, etc.) True, what bears utility for Beauty isn't specified; the idea, presumably, is to use what *is* specified to identify natural candidates. One thing that is specified is the number of awakenings. That, it could be argued, is all we've got to go on. Hence thirding. But we are given something else. Beauty is a rational agent. Rational agents traffick in information for its own sake. Being right about how things are is intrinsically pleasing; being wrong is intrinsically displeasing. The question is: does this (dis)pleasure more resemble the pleasure of those who like to travel, or of those who like having travelled? Asked another way: with respect to information consumption, do rational agents live to eat, or eat to live? Thirders, it would appear, live to eat. Halfers eat to live. And if in fact there's an ambiguity here, Beauty's personal credence in *heads* may be as underspecified as what a bulimic had for lunch.

I'm going to just take it that this sort of hand-waving about how information isn't necessarily subject to double counting makes plausible the assumption of a protocol friendly to a one-half solution. This vindicates (if aphoristically) the one-half solutions in [6], [9], and [10]. However, it does *not* vindicate halfing schemes that update propositional credences in light of *de se* evidence by conditioning on the proposition corresponding to the set of worlds consistent with the evidence. I'll call this sort of scheme ([4], [11] and [16] are my intended targets) *Passepartoutian* halfing (after the valet of Phileas Fogg). Passepartout, figuring perhaps that London time is the *one true time*, is reluctant to advance his pocket watch when entering a new time zone; if only he can ride out his master's travels and get back to London, his watch will be back in line with those of locals and all will be well. The Passepartoutian halfer, meanwhile, perhaps figuring that objective chance is the *one true probability*, is reluctant to change her propositional credences in light of merely *de se* evidence; if only she can ride out the scenarios involving indiscriminable collocated alternatives, her credences will be back in line with objective chance (and all will be well).

Bostrom refers to his brand of Passepartoutian halfing as a "hybrid model". Indeed it is, and is the worse for it: it consistently reflects neither Beauty's expectations (minimization of real time surprisal), nor her best course of action under any natural protocol. Like all of the schemes I am criticizing, all it has going for it are two plausible yet incompatible intuitions: that (1) Lewis's lemma (L1) is correct, and (2) if Beauty learns *Monday* (thus eliminating multiplication of collocated alternatives), probabilities should revert to objective chance. The reason that these intuitions are incompatible is that acting in accord with (1) involves an implicit presumption that there is just one net significant *tails* awakening (otherwise Beauty's lost certainty as to which this was would be relevant to *heads*). But in that case, for all Beauty knows upon learning *Monday*, she is experiencing an awakening of diminished or no significance if *tails*–which favors *heads*.

Of course, it should not be surprising to Monty Hall aficionados that the hidden role of protocol in halfing should be opaque to naive intuition, and we can use the problem to dramatize what's wrong with the Passepartoutian scheme. Suppose that a **big prize** is hidden behind one of three doors, each with equal objective chance. The hypothesis *Door i* corresponds to the state of affairs in which the **big prize** is behind Door $i$. If *Door 1*, then Beauty will have a single awakening, on Monday. If *Door 2*, Beauty will have a single awakening, on Tuesday. And, if *Door 3*, Beauty will have two awakenings, on Monday and Tuesday. Halfers of course assign each of the alternatives credence $\frac{1}{3}$ upon awakening.

Suppose now that a halfer learns what day it is, and is asked for her updated credence in *Door 3.* Note: if *Monday*, *Door 1* is eliminated. If *Tuesday*, *Door 2* is eliminated. *Door 3* cannot be eliminated. Recall that our halfer has prior credence $\frac{1}{3}$ in *Door i* for each $i$ and, if she accepts Elga's principle, *Monday* and *Tuesday* are equally likely conditioned on *Door 3.* Suppose our halfer learns *Monday.* Since the current protocol is isomorphic to that of the Monty Hall problem, her situation is precisely that of a Monty Hall contestant that has initially chosen *Door 3* and seen the hypothesis *Door 1* eliminated.

The Passepartoutian halfer, then, who updates credences by conditioning on *not Door 1*, is committing the well-known fallacy of those who answer $\frac{1}{2}$ in the Monty Hall problem, in defiance of the understood protocols. On the contrary, Beauty's credence in *Door 3* must remain $\frac{1}{3}$. Anything else violates the law of total probability, not to mention common sense; it just can't be that Beauty should update credence in *Door 3* from $\frac{1}{3}$ to $\frac{1}{2}$ upon learning what day it is *regardless of what day it is.* If that's all Beauty's got, she needs to revisit thirding.

Indeed, this *is* essentially Rosenthal's ([12]) argument for thirding; alter the original problem so that the single *heads* awakening occurs on either Monday or Tuesday (with equal probability). Rosenthal takes it as uncontentious (Passepartoutian halfers agree) that Beauty's credence in *heads* upon learning *Monday* is $\frac{1}{3}$. Since the same holds for *Tuesday,* absolute credence in *heads* must be $\frac{1}{3}$ by the law of total probability.

To recapitulate: the reason that halfer probabilities don't revert to objective chance upon elimination of collocated alternatives is that *halfer probabilities aren't absolute credences;* they're conditioned credences. They track, under a natural class of utility allocation protocols, Beauty's credences *conditioned on its mattering one way or the other.* They are, for all that, nevertheless probabilities, and must in particular be updated in the standard way, i.e. by conditionalization. Like Fogg's valet, whose obstinacy couldn't account for the International Date Line, and who would thereby have squandered his master's fortune but for an accident, halfers who don't conform will require, in the long run, many such accidents to preserve their own.

*Conclusion.* I've maintained that the one-third solution is best for the Sleeping Beauty problem as formulated, but for some arguably natural protocols governing the allocation of utility, behavior consistent with a one-half solution may be optimal. It's important for halfers to remember, however, which protocols support their scheme, lest their intuitions fail them as to how to update the conditioned credences they favor in light of *de se* evidence. The primary moral, for halfers and thirders alike: *credences are probabilities!*

## References

[1] Arntzenius, Frank. 2003. Some problems for conditionalization and reflection. *The Journal of Philosophy*. 100:356-370.

[2] Chalmers, David. 2002. The St. Petersburg Two-Envelope Paradox. *Analysis*. 62:155-57.

[3] Elga, Adam. 2000. Self-locating belief and the Sleeping Beauty problem. *Analysis* 60:143-147.

[4] Bostrom, Nick. Sleeping beauty and self location: A hybrid model. *Synthese*. 157:59-78.

[5] Gallager, Robert G. 2011. *Stochastic Process: Theory for Applications* (draft). Available at http://www.rle.mit.edu/rgallager/notes.htm

[6] Hawley, Patrick. 2012. Inertia, Optimism and Beauty. *Nous*. To appear. Available at http://philsci-archive.pitt.edu/5319/1/iob.pdf

[7] Horgan, Terry. 2004. Sleeping Beauty awakened: new odds at the dawn of the new day. *Analysis* 63: 10-21.

[8] Lewis, David. 1979. Attitudes *De Dicto* and *De Se*. *The Philosophical Review* 88: 513-543.

[9] Lewis, David. 2001. Sleeping Beauty: Reply to Elga. *Analysis* 61:171-176.

[10] Lewis, Peter J. 2007. Quantum Sleeping Beauty. *Analysis* 67: 59-65.

[11] Meacham, Christopher. 2008. Sleeping Beauty and the Dynamics of *De Se* Beliefs. *Philosophical Studies* 138: 24569.

[12] Rosenthal, J. S. 2009. A mathematical analysis of the Sleeping Beauty problem. *Mathematical Intelligencer* 31: 32-37.

[13] Ross, Jacob. 2010. Sleeping Beauty, countable additivity, and rational dilemmas. *The Philosophical Review* 119: 411-447.

[14] Shaw, James R. 2013. De se belief and rational choice. *Synthese* 190:491-508.

[15] Weintraub, Ruth. 2004. Sleeping Beauty: A Simple Solution. *Analysis* 64: 8-10.

[16] White, Roger. 2006. The generalized Sleeping Beauty problem: a challenge for thirders. *Analysis* 66: 114-119.

*rmcctchn@memphis.edu*