Beauty, Eh? In Defense of Thirding and (Lewisian) Halfing

So how's it goin', eh? Our topic for today is Sleeping Beauty.

It could be argued that there are already too many papers on this topic. But, there probably aren't many written by mathematicians, and fewer by mathematicians who have read at least half (well, okay, a third) of the literature. Being such, this one may offer a fresh perspective. Although I am primarily attempting to increase understanding of several neglected mathematical issues, however, I have not resisted the temptation to offer my own philosophical take on their significance for the problem. My overall conclusion will be that the one-third and the (Lewisian) one-half solutions are both fine, albeit for different readings of *credence*. Hopefully whatever philosophical naiveté is on display in these few paragraphs can be forgiven,[1] and will not prevent readers from seeing that the mathematical portions may in fact advance the overall discussion non-trivially.

1. Take off, eh? No roads lead to violations of countable additivity.

In [9] there is an argument purporting to show that the one-third solution to the Sleeping Beauty problem is at odds with countable additivity of probabilities. The argument requires a scenario in which a rational agent is subjected to an experiment whose expected duration, by her own lights, is infinite. In spite of this implausibility, the argument seems to have gained many adherents. Nevertheless, the scenario in question is of a kind familiar to mathematicians, who have routinely cautioned against the danger of adopting bad models of their type–the very confusion that [9] capitalizes on.[2] (Brutal, eh?)

In the version I will consider, Beauty is a rational agent participating in an experiment in which a coin is tossed on Sunday night. Beauty is awakened Monday morning, asked her credence in *heads* and told the outcome. If the coin landed heads, that's the end of the experiment. If *tails*, Beauty is given a drug that puts her to sleep for another 24 hours and erases all memory of her Monday awakening. Then on Tuesday morning she is again awakened, asked her credence in *heads*, told the outcome, and that's the end of the experiment. The problem is what Beauty's credence in *heads* should be on Monday morning. A *halfer* says one-half. A *thirder* says one-third. More generally, a Sleeping Beauty problem is defined to be "a problem in which a fully rational agent, Beauty, will undergo one or more mutually indistinguishable awakenings..." ([9]) where the number of

[1] Along with the summary offences of cluttering Sleeping Beauty bibliographies further and employing dubious cultural references (which temptation I have also failed to resist).

[2] Though I'll eschew technical details here, the original, repeated Sleeping Beauty problem can be modeled by a positive recurrent Markov chain, with the one-third solution corresponding to its stationary probability measure. The example of [9] (*Sleeping Beauty in St. Petersburg*; see below) is a null recurrent chain, and, as such, lacks stationary probability distributions; for the received philosophical interpretation of this state of affairs among mathematicians, see e.g. Chapter 6 of [3], in particular pages 286 and 294.

such awakenings is a function of a random variable into a set S of hypotheses.

One argument for the one-third solution is that, if the experiment were repeated many times, the expected asymptotic density of heads awakenings would be $\frac{1}{3}$. The argument in [9] begins with the claim that the thirder utilizes 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 other innocuous premises, leads to the following two principles (it does not seem that both are needed, and the former follows from the latter).

> *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.
>
> *Frequency.* In any sleeping Beauty problem, if Beauty knew that the experiment was to be repeated many times then her credence in any hypothesis $h \in S$ should be proportional to the expected long-run frequency of h-awakenings.

An example is given showing that *Frequency* (and/or *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.

> *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 $P(X = n) = 2^{-n}$, $n \in \mathbf{N}$.

As the expected duration of an *SBSP* experiment is infinite, the asymptotic frequency of $X = n$ awakenings is zero for all n. According to *Frequency*, then, Beauty should upon awakening have credence zero in each of the assertions $X = n$. But she has credence 1 in their disjunction. This violates *CA*. Alternatively, by *GTP* Beauty should, upon awakening, assign equal credences to the exhaustive and mutually exclusive assertions $X = n$. This too violates *CA*.

The problem with this argument is that no rational thirder would literally adopt *FSBI*, for the simple reason that, with non-zero probability, the experiment will end between any first and second tails awakenings. That our original thirder ignores such remote possibilities is innocent–it matters little, and it would be difficult to communicate without such sanctions. The innocence dries up quickly in the *SBSP* example, however. Indeed, consideration of a single possibility as fantastic as that while Beauty is asleep she is, through a series of quantum mishaps, tranformed into a golden toque (or a Golden Molson, eh?) suffices to

make finite Beauty's expectation of the experiment's duration, which effectively eliminates any perceived tension involving countable additivity.[3]

Here are some details. First, if we are being literal, *FSBI* needs to be replaced by:

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

Here $Ch(\cdot)$ is objective chance. Observe now that *FSBD*, 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 kth 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

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

It is convenient now to define the *fidelity* of an implementation of a Sleeping Beauty experiment as $Ch(N = M)$, 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 have seen that under suitable hypotheses Beauty's credence in h is $P(h) = \frac{Ch(h) \cdot E(M|h)}{E(M)}$. It follows that if $E(N) < \infty$ then as fidelity approaches 1 Beauty's credences will 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 will approach zero for every h and, consequently, Beauty's credence functions will diverge in variation. It is natural to say that in the former case, the distribution Q constitutes a solution to the problem, while in the latter case there is no solution.

[3]It is tempting to object that this amounts to changing the problem, but that is not so. The change came earlier, when Beauty decided that *Monday* and *Tuesday tails* awakenings were equally likely. This seemingly insignficiant change was justified because it simplified her model. But modeling assumptions are always subject to the condition that outputs of interest (e.g. conformity with *CA*) be continuous relative to inputs. Models that fail this condition (null recurrent Markov chains such as *SBSP*, in particular) are bad models–they fail to cast light on the behavior of systems with approximating inputs.

This is consistent with *CA*. It also recovers the one-third solution. Unless, of course, one objects that now Beauty is assigning credence .3333343209... to heads and is therefore no longer a real thirder; perhaps real thirders still run afoul of *CA* per [9]. I'm not sure about that but won't argue. One can say what one likes about thirders in that sense because in that sense there aren't any thirders to take offense.

2. Let's make a deal, eh? Monty Hall shuts the door on non-Lewisian halfers.

Let's analyze the problem using the logarithmic scoring rule,[4] where the agent with estimate p for $P(heads)$ scores the negative of surprisal, here $\log p$ if *heads* and $\log(1-p)$ if *tails*. Assuming a fair coin, minimum expected surprisal occurs at $p = \frac{1}{2}$ (a calculus exercise). But Beauty is *surprised twice* if the coin lands tails; some might contend that her surprisal is actually $2\log(1-p)$ in this event, with minimum expectation at $p = \frac{1}{3}$. Obviously the thirder is committed to the latter mode of accounting, the halfer to the former. I believe that both modes of accounting make sense; the thirder regards *de se* information as admissible, whereas for the halfer, only propositional information is admissible.[5]

Here's another argument that both solutions make sense: consider what would happen in the original repeated experiment if instead of erasing Beauty's memory we just didn't tell her the times or results of the tosses. So Beauty is just living life as usual and in the next room someone is tossing a coin some evenings, skipping a evening after each *tails*. What credence should Beauty assign to the previous toss having landed *heads*? I hope the question sounds ambiguous. Say the previous toss was the seventeenth toss. There's a *de dicto* reading, on which Beauty should assign credence $\frac{1}{3}$ to the previous toss, whichever one that was, having landed *heads*, and a *de re* reading, on which Beauty should assign credence $\frac{1}{2}$ to the actual previous toss, i.e. the seventeenth toss, having landed heads. Beauty herself doesn't suffer this ambiguity, as the *de re* reading is not available to her. But in the original problem, where her memory is erased, it is. The $\frac{1}{3}$ reading is still available too, though one should probably call it a *de se* reading now.

If this is right, then there should be, even among thirders, a strong inclination to charitability regarding efforts to hammer out the details of a one-half solution.[6] The halfer

[4]Only the logarithmic rule is linear with respect to surprisal, which is, it would seem, the quantity rational agents trafficking in information seek to minimize.

[5]To elaborate: in order for both to count, the Monday and Tuesday *tails* surprisals must derive from non-identical (independent, in fact, if both are to count fully) revelations. As propositions, the Monday and Tuesday *tails* surprisals do derive from identical revelations, namely *the coin landed tails*. Beauty may, for all we know, have been (propositionally) omniscient but for knowledge of that toss (even though she doesn't know what day it is; cf. the omniscient gods in [6]), and she can't be scored minus two bits if she came in only one bit shy of a complete state description. In contrast, Monday's and Tuesday's *de se* surprisals derive from the doxastically independent assertions *this is a tails awakening*.

[6]The surprisal-minimization argument gives some inkling as to how this might be done.

must make two critical choices. First: whether or not to accept Elga's "restricted principle of indifference" [1], namely whether to accept that $P(\textit{Monday tails}) = P(\textit{Tuesday tails})$. I will assume that halfers who do not explicitly deny the principle accept it. (Cf. [2]; I don't consider here the view, which may have merit, that halfing Beauty should leave *de se* events unmeasurable.) Second: how to update propositional credences in the light of *de se* evidence.

Because it's convenient (and amusing), I'll analyze the second question using a more-than-vaguely-familiar Sleeping Beauty problem in which there are three hypotheses of equal objective chance. The setup: there are three doors. A roll of a fair die determines which of the doors will have a new car placed behind it. The other doors will have chickens placed behind them. The hypothesis *Door i* corresponds to the state of affairs in which the new car 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 are committed to assigning each of the three doors credence $\frac{1}{3}$ upon awakening. If this isn't obvious, consider that if Beauty asks whether *Door 1* obtains and is told *no*, she is in the position of the original problem with respect to the remaining possibilities, and so must assign them equal credences. On the other hand, she updates on the proposition *not Door 1* by traditional conditioning, hence her prior credences in *Door 2* and *Door 3* are equal. *Mutatis mutandis*, her prior credences in *Door 2* and *Door 3* are equal.

Now the halfer asks what day it is, and after hearing the answer is asked for her updated credence in *Door 3*. Some observations: if the answer is *Monday*, this rules out *Door 1*. If the answer is *Tuesday*, this rules out *Door 2*. *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 restricted principle of indifference, *Monday* and *Tuesday* are equally likely conditioned on *Door 3*. Such a halfer's predicament is therefore isomorphic in protocol to that of the Monty Hall problem when she has initially chosen *Door 3* and seen the hypothesis *Door 1* eliminated. Accordingly, any halfer who updates credences by simple conditioning on *not Door 1* is committing the well-known fallacy of those who answer $\frac{1}{2}$ in the Monty Hall problem. Namely, conditioning on the proposition learned instead of conditioning on the fact that it, among other candidate propositions, was learned. No, Beauty's credence in *Door 3* must remain $\frac{1}{3}$ upon learning what day it is. How do extant halfing schemes fare?

Roger White in [11] argues for the one-half solution based upon a variant of the original problem in which, at each possibly current centered world, Beauty wakes up with prob-

---

Briefly...if only one tails surprisal is to count, a "one vote" policy must be instituted, according to which Monday and Tuesday tails versions of Beauty, between them, get a single information-theoretic vote, responsibility for which might be apportioned by stipulation, lots or weighted average. I take it that the details of the one-third solution, meanwhile, are uncontroversial.

ability $q < 1$. White advocates updating credences at each experimental awakening by conditioning on the propositional event *there is at least one awakening*. The general policy I take this to sanction is that one update credences in response to *de se* evidence by conditioning on the largest propositional event consistent with that evidence. But in our Monty Hallish scenario, where Beauty finds out that it's Monday, this sanctions the fallacy of conditioning on *not Door 1*, arriving at $\frac{1}{2}$. (Failure.)

Christopher Meacham meanwhile in [9] introduces an intricate scheme, termed *compartmentalized conditionalization*, which is designed to yield values $P(A|B)$ when A, B or both are *de se* events. To gloss over details somewhat, in the current scenario, when Beauty learns *Tuesday*, her $\frac{1}{6}$ credence in *Monday Door 3* is first transferred entirely to its surviving collocated alternative, *Tuesday Door 3*, then her $\frac{1}{3}$ credence in the *Door 1* hypothesis, which has no surviving collocated alternative, is distributed among the alternatives of the remaining hypotheses in proportion to their existing credences. (Failure again.[7])

Finally, we consider the halfing scheme of David Lewis [3]. Like Meacham, Lewis accepts Elga's restricted principle of indifference, and so assigns both *Monday Door 3* and *Tuesday Door 3* credences of $\frac{1}{6}$. He also uses standard conditionalization upon elimination of *de se* alternatives, so his updated credence in *Door 3* is $\frac{1}{3}$, which is, well...correct. (Beauty eh?)

So what exactly was it about Lewis's scheme that seemed implausible enough that several halfers saw a need to amend it? Compartmentalized conditionalization looks, at first blush, like a plausible plan to rescue halfing from the attacks of thirders ([5] and [10], among others) whose arguments may have appeared to suggest that Lewis's conditioning scheme wasn't consistent with his original choice to assign *heads* a credence of $\frac{1}{2}$ in Beauty's original predicament. Indeed, if we consider a situation in which *heads* and *tails* are to each result in two awakenings (on Monday and Tuesday), and we imagine that, during an awakening, Beauty finds out that *Tuesday heads* does not obtain, it might seem that Lewis is under conflicting obligations to update credence in *heads* to: (a) $\frac{1}{3}$, because Lewis had previously assigned credence $\frac{1}{4}$ to each possible alternative and uses standard conditioning to update credences, and (b) $\frac{1}{2}$, since the situation is now just like that of the original problem. If Lewis is in trouble, compartmentalized conditionalization might generate some sympathy.

However, Lewis is not in trouble. (Not, at least, on consistency grounds.) In his scheme, the elimination of an alternative does not restore credences to what they would have been had the alternative not existed in the first place. As he sees it, to get credences as they would be if an alternative hadn't existed in the first place, you distribute its credence among its collocation partners. If the alternative gets eliminated, on the other hand, its credence gets distributed among all surviving alternatives, collocated or not. Lewis is not explicit on the point, but we can infer it from the fact that Lewis updates Beauty's credence in *heads* to $\frac{2}{3}$ when *Tuesday* is eliminated, whereas if the Tuesday awakening

[7]The updating schemes of White and Meacham appear to be identical when updating credence on propositions by elimination of alternatives, so their fates are linked here.

hadn't existed in the first place, Beauty's credence in heads would just be $\frac{1}{2}$.[8]

So much for halfers who accept Elga's restricted principle of indifference. What of those who deny it? The Monty Hall argument doesn't apply to them. Are there such halfers? In fact there are. Patrick Hawley [4] argues that in the original problem, Beauty should assign credence 1 to *Monday* (and thus in particular to *Monday* conditioned on *tails*). His argument has two premises. (1) $P(heads) = \frac{1}{2}$, and (2) $P(heads|Monday) = \frac{1}{2}$. An implicit third premise, then, is the multiplication rule (3) $P(A \wedge B) = P(A)P(B|A)$. These three premises are not simultaneously satisfied in any of the systems we've looked at. Halfers accept (1), thirders, White and Meacham accept (2), thirders and Lewis accept (3). The status of this argument is therefore unclear.

The key point, perhaps, is merely that the premises are independently plausible and consistent, so that one may simply adopt them as axioms. This view is coherent. If Lewis can ask (and he has, in effect) that Beauty's Monday and Tuesday *tails* versions each be weighted by a factor of one-half, or perhaps that a coin toss will determine which is to receive full weight, why not just always count Monday's version? On the other hand, the view is quite arbitrary, for whatever one can say in favor of assigning full weight to Monday Beauty, one can as easily say in favor of assigning full weight to Tuesday Beauty.[9] Hawley writes: "the best compromise...might well be to believe to degree 1 that it is Monday whenever she awakens. She will be surely be right on Monday, and possibly never be wrong about the day during the experiment." Consider however so-called *Mondrue*, which is like Monday whenever the coin lies heads and like Tuesday whenever the coin lies tails. In favor of this freshly minted day one might counter: "the best compromise...might well be to believe to degree 1 that it is *Mondrue* whenever she awakens. She will be surely be right on *Mondrue*, and possibly never be wrong about the day during the experiment."

Well, that's our topic for today. To recapitulate my findings: thirding coheres with accepted principles of probability. Arguments presented by others on previous occasions show it to be also reasoned and natural. Lewisian halfing, which has been kicked about rather unfairly, turns out to be coherent and reasoned as well. On the other hand it's probably natural only to the most eccentric of rational agents. Hawley-style halfing is coherent too, but arbitrary. Other forms of halfing are incoherent.

So good day, eh?

---

[8] Lewis's policies are an example of "one vote accounting". Suppose Beauty sends herself a postcard during each awakening, and on Wednesday morning (her memory having been wiped again), must read her credence in *heads* off of a single postcard (selected, in case of tails, at random). She will then record Lewisian probabilities.

[9] If that's not arbitrary enough, one could presumably assign Monday weight q and Tuesday weight $1 - q$ for any $q \in [0, 1]$. Only $q = \frac{1}{2}$ (i.e. Lewisian halfing) appears to have any respect for symmetry.

References

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

[2] Elga, Adam. 2010. "Subjective probabilities should be sharp." *Philosophers' Imprint* 10(5).

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

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

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

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

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

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

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

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

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

rmcctchn@memphis.edu