To sleep, perchance to have indiscriminable collocated awakenings: ay, there's the rub

Beauty (see [3]) is a rational agent 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 her credence again and debriefed. If the coin landed heads, that's the end of the experiment. If *tails*, Beauty is given a drug that erases all memory of that morning's experiences and puts her back to sleep. Then on Tuesday she does it all again. Beauty knows all of this in advance. The problem is what her initial personal credence in *heads* should be on Monday. 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*.

In a famous early incarnation, credences were tools for gamblers interested in maximizing winnings. If credences are just solutions to optimization problems, the answer turns on what quantity is to be optimized, and what policies govern accrual of that quantity. What gets optimized is, it turns out, not so important. One can stick with stakes and maximize winnings (as proxy for *utility*), or invoke information theory and minimize *surprisal*. In the former case Beauty will strive to avoid vulnerability to a Dutch Book. In the latter, she'll try to maximize performance as measured by a (logarithmic) scoring rule. The policies governing accrual are what matter. The naive view, which Jacob Ross [10] calls *Every Awakening Legitimacy* (*EAL*), vindicates thirding (see [3], [5], [9], etc.) because whatever quantity one is optimizing accrues twice if *tails*.

There's a competing view, though–Ross calls it *Single Awakening Legitimacy* (*SAL*)– according to which "legitimate" consequences accrue only once, regardless of the outcome of the toss. That supports halfing (see [1], [4], [6], [7], [8], [13], etc.), modulo agreement that Beauty isn't "tipped off" as to the rate of accrual. I take it that Beauty's choice between halfing and thirding turns completely on whether she favors *SAL* or *EAL*, and that both choices support stable conventions. So I won't argue (seriously) for either here.

What I will argue is that Sleeping Beauty's predicament involves confrontations with neither the accepted laws of probability nor the conditionalization paradigm for updating credences in light of new evidence. This thesis leads me to question two recent claims. First is Ross's contention that there is a "deep tension" between the one-third solution and countable additivity of credences. Second is the claim by so-called "double halfers" ([1], [8] and [13]) that Beauty should update her credences in response to centered (*de se*) evidence by conditioning on the set of uncentered worlds consistent with that evidence.

Acknowledgement: An anonymous referee commented on a previous version of the paper.

1. I'm okay, you're okay: why thirders and Kolmogorov are okay

Subject to *EAL*, the betting arguments clearly vindicate thirding, as has been argued elsewhere. Another type of argument is from *surprisal*. If an agent has credence p in A, her surprisal (the number of bits of information acquired) upon learning A is $-\log_2 p$. Since to know more now is to be surprised by less later, agents seek to minimize surprisal.

According to *EAL*, Beauty is surprised twice if *tails*, so her expected surprisal during the experiment is $-\frac{1}{2}\log_2 p - 2 \cdot \frac{1}{2} \cdot \log_2(1-p)$, which is minimized at $p = \frac{1}{3}$.

Other arguments for Sleeping Beauty have focused on evidence. In order for there to be a change in credence, so the arguments go, there must be a change in Beauty's evidence. If that's right, thirders must argue either for new evidence or for lost evidence. I believe that both can be done persuasively. In the argument from new evidence, one uses so-called Jeffrey conditionalization on Monday's Debriefing (*MD*)–decisive but uncertain evidence for *tails*. The computation is straightforward: one has

$$P(heads) = P(MD)P(heads|MD) + P(\sim MD)P(heads|\sim MD) = \frac{1}{2}\big(1 - P(MD)\big),$$

while an indifference principle of Elga [3] says that $P(heads) + 2P(MD) = 1$, yielding $P(heads) = P(MD) = \frac{1}{3}$. In the argument from lost evidence, since Beauty has forgotten what day it is and since what day it is is relevant to *heads*, her credence in *heads* should revert to the expected value of yesterday's credence–namely one-half if today is Monday and one if today is Tuesday. The computation is essentially the same.

Many arguments for thirding are possible when one starts by accepting *EAL*. In the third section I will discuss halfer responses to the arguments. In each case, the response will be characterizable as simple substitution of *SAL* for *EAL*.

Next I give a diachronic Dutch Book argument for:

> *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.

Note: the argument requires only a bounded number of stakes.

Let X be a random variable on the naturals and consider a credence function P such that $\sum_{n=1}^{\infty} P(X = n) = 1 - \epsilon < 1$. For a large M, let $(X_i)_{i=1}^{M}$ be independent random variables distributed as X is. An agent subscribing to P has X_i revealed to her in turn. After $X_1, \ldots, X_{i-1}$ are revealed, she may bet a dollar that $X_i > \max\{X_j | 1 \leq j < i\}$. If she wins, she gets $\frac{2}{\epsilon}$ dollars. For any k, $P(X_i > k) \geq \epsilon$, so she'll take the bets.

Next, imagine that we have $M!$ agents, all subscribing to P. Each is assigned a different permutation π of $\{1, 2, \ldots, M\}$ and is offered a series of bets like that of the previous paragraph, but with the X_i's revealed in the order $X_{\pi(1)}, \ldots, X_{\pi(M)}$ (the agent wins the ith bet if $X_{\pi(i)} > \max\{X_{\pi(j)} | 1 \leq j < i\}$). They all bet from the same account. To break even, the proportion of bets they win must be at least $\frac{\epsilon}{2}$. But if X_i is the kth largest out of $X_1, \ldots, X_M$ (ties broken arbitrarily), the probability of a randomly selected agent winning when X_i is revealed is at most $\frac{1}{k}$, meaning that the proportion of winning bets is at most $\frac{1}{M}(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{M}) \approx \frac{\log M}{M}$, which tends to zero as M increases. So for large M, the P-subscribers collectively suffer a sure loss.

The moral of the story is that, subject to *EAL*, both *CA* and the one-third solution are okay. In fact, they're non-negotiable.

2. To sleep No more: why no accessible roads lead to violations of countable additivity

According to Ross [10], however, there are situations in which we are unable to subscribe to thirder reasoning while simultaneously satisfying *CA*. The situation he describes is a Sleeping Beauty problem ("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 taking values in a set S of "hypotheses") in which the expected number of awakenings is infinite. His claims about what thirders are committed to starts with 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 premises (details omitted), leads to a:

> *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.

A pathological example is introduced, purporting to show that *GTP* conflicts with *CA*:

> *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*.

As mentioned, in *SBSP* the expected number of awakenings, $\sum_{h \in H} Ch(h)N(h)$, is infinite. Here $Ch(\cdot)$ denotes objective chance and $N(h)$ is the number of awakenings associated with h. It follows of course that *SBSP* can't be faithfully implemented at our world, nor at any nomologically accessible world, nor for that matter at any world subject to a reasonably time stationary threat of mortality. So it's unclear how to interpret Ross's reports of a "deep tension" between *GTP* and *CA*.

I see two ways. Either "tension" is to be predicated of a set of principles when there exist worlds, no matter how unfamiliar, at which they conflict, or it's a two place relation between sets of principles and worlds...with tension between *CA* and *GTP* occurring only at very special, remote worlds. The former reading seems inevitable, given that Ross doesn't qualify "tension" in any way, and plainly means for it apply *here*. However, even if such a weak modal notion should concern actual agents, pitfalls remain.

Chiefly: *GTP* appears to be a red herring. Ross argues from conflict between *GTP* and *CA* to "rational dilemmas", i.e. "contexts in which full rationality is impossible". It would threaten his thesis if the only worlds at which the conflict can arise are so crazy that even *halfers* find it impossible to adopt "fully rational" credences there. Ross takes this possibility seriously, for he briefly considers the following premise:

> *Sleeping Beauty Indifference (SBI).* In any Sleeping Beauty problem, for any hypothesis h in S, upon first awakening, Beauty should have equal credence in each of the awakening possibilities associated with h.

He then writes "One might claim...that...almost everyone is committed to SBI. And since CA conflicts with SBI, all parties to the debate should reject CA, regardless of whether they accept the Generalized Thirder Principle." This could undermine his claims about the significance of the conflict between *GTP* and *CA*, and Ross is quick to deflate it, in two ways. First by claiming that, for halfers "the conflict between *SBI* and *CA* arises only when (*CA*) is understood to range not only over ordinary propositions but also over centered propositions...." And, second, by substituting *FSBI* for *SBI*. If the deflationary project is successful, Ross will have shown that at some worlds, *CA* by itself is okay, while *CA* and *GTP* are in conflict.

Given Ross's ambitions, though, conformity with *CA* for halfers at the crazy worlds isn't enough. Halfing Beauty must find it possible to be fully rational there. This seems unlikely–infinite expected lifespan arguably gives rise to an unbounded utility function, which elevates the status of known paradoxes relating to runaway utility (in particular the St. Petersburg two envelope paradox; see [2]) from logical curiosities to genuine rational quandaries. It's open to Ross to argue against unbounded utility at worlds supporting faithful implementation of *SBSP*, but Ross gives no such argument, and it can't be the default view.[1] The upshot of this is that credences are rationally constrained by:

> *Finite Expectation (FE).* In any Sleeping Beauty problem, if $N(h)$ denotes the number of awakenings associated with h, then Beauty's credences $\{P(h) : h \in S\}$ should satisfy $\sum_{h \in S} P(h)N(h) < \infty$.

For consider a situation in which Beauty has been sentenced to an *SBSP*-style incarceration involving a mild form of torture...say, being isolated in a room with only a copy of *Word and Object* from which all but the middle chapters have been removed. Beauty is free to choose between two rival groups of scientists (groups *Dull* and *Duller*) to carry out the sentence. Each has already computed the number of awakenings she would experience at their hands. She's chosen group *Dull*, but this choice is arbitrary. Now she will be offered a sequence of two trades that she'll have to accept if she fails to subscribe to *FE*, but which will leave her worse off. First, the judge (who doesn't know the values N) offers

[1] It would be somewhat shocking if utility were bounded for an agent having infinite expected lifespan, as this would imply virtual indifference, asymptotically, between (even vast eons of) pleasure and pain.

her a halving of her sentence to switch groups. By indifference, she accepts, and switches to group *Duller*. Next, the judge asks group *Dull* to reveal their number and offers to let her switch back. At a price–the quadrupling of her previously halved sentence. This is twice as much time as she was originally going to serve if she hadn't taken the first trade. Nevertheless she accepts the trade, as $E(N_{Duller}) = \infty$.

Full rationality lives and dies with bounded utility. Which, in our context, is reflected in the *FE* constraint. What Ross shows is that when (non-rational by default) agents determine that rational conformity with *FE* is impossible, they must renounce either *GTP* or *CA* as well. Ontologically, this discovery is almost surely vacuous but, even if it isn't, the conflict between *GTP* and *CA* does no work in the spawning of rational dilemmas. At best, it recycles them.

3. A day late and a dollar short: why halfers are okay

There exist natural betting protocols whereby Beauty at least *acts* like a halfer. Nick Bostrom [1] proposes a thought experiment (*Beauty the high roller*) where bets are offered to Beauty on Mondays only. Given such a protocol (perhaps we agree that phony bets are offered on Tuesdays so she won't be tipped off), Beauty should behave as a halfer in the style of Hawley [4], who assigns *Monday* probability 1 conditioned on *tails*.[2] If the one real bet were to be offered on *Monday* or *Tuesday* with equal likelihood conditioned on *tails*, Beauty will identify strongly with Peter J. Lewis [7]'s *quantum Sleeping Beauty* interpretation. Shaw [11] introduces (in essence) bets that Beauty can make only (and only *once*) by agreeing to them during each awakening of the experiment. This evokes Lewis [6]'s answer (familiar to any statistician) to tails world oversampling: sample weight dilution of the *tails* awakenings.

The accord between Beauty's betting behavior under such protocols and *SAL* is obvious. Whether there is an accord between Beauty's betting behavior under such protocols and her *personal credences*, however, is less so. Indeed, there doesn't seem to be any good reason for the distribution of bets to constrain Beauty's personal credences at all.

The situation with surprisal, however, is potentially more friendly to halfing. Yes, Beauty is surprised twice if *tails*, but from one viewpoint what accrued surprisal is supposed to measure is information acquired since initiation of scoring, which will hardly be the case if the same piece of information gets counted twice. From the informational perspective, there seems to be little difference between gaining, losing and regaining information and simply gaining it once. In some sense, the initial gain is canceled by the loss. Back on the wagering front, any contradictory intuition that a lost bet can't be canceled just in virtue

[2]Hawley's position might be described as *First Awakening Legitimacy*. It and its cousins (*Last Awakening Legitimacy*, etc.) surrender Beauty (confident that she'll only be scored when she's right) to the indignity of assigning zero credence to events that turn out to obtain.

of one participant having forgotten its outcome is perhaps an artifact of a fairness norm, or that we imagine money having irreversibly changed hands. But the betting game is really just a thought experiment Beauty plays with herself. When a running tab between wagering rivals is kept "in the head", and both parties forget an outcome, they likely forget to score it as well.[3] Perhaps information loss should be scored as a "negative surprisal', or, when one expects multiple surprisals, be weighted in inverse proportion to the number expected.

I don't think this ragtag of aphoristic support for halfing does much to derail thirding, but it does suggest that perhaps the concept *credence* as previously applied was not fine-grained enough to handle Sleeping Beauty cases, and that there are two viable options for its explication. For the sake of exploring halfing further, I'll grant as much. If you don't agree, the rest of the paper could still potentially be interesting to you as a commentary on a particular type of optimization problem; it just won't be about *credences*.

When one incorporates *SAL* into arguments for thirding, they become arguments for halfing. It's obvious why one-wager protocols (betting paradigm versions of *SAL*) lead to halfing, and straightforward enough how *SAL* changes the information-theoretic argument: under *SAL*, only one tails surprisal should be scored, which means that the quantity to be minimized is $-\frac{1}{2}\log_2 p - \frac{1}{2}\log_2(1-p)$ (this occurs at $p = \frac{1}{2}$). In the arguments from evidence, meanwhile, *SAL* manifests itself in the form of either *disenfranchisement*, a policy whereby one of the tails awakenings is essentially silenced, or *dilution*, in which the tails awakening get only "half votes" each. This policy vindicates Lewis's view that Beauty acquires no evidence relevant to *heads*. Such evidence took the form of uncertain evidence for *tails* acquired Monday but reflected upon Tuesday and for disenfranchising halfers (such as Hawley), Tuesday's gone. (With the wind, perhaps...how else?)

So for all thirders have said to the contrary, halfing seems to be at least coherent, if counterintuitive. When Beauty learns *Monday*, though, things get strange.

4.'Deal' breaker: a more calamitous embarrassment for double halfers

According to Lewis [6], when Beauty learns *Monday*, her credence in *heads* jumps to $\frac{2}{3}$, a move made even more counterintuitive when one grants (as Lewis does) that the coin can be flipped later that Monday.

Thirders think this is an embarrassment for Lewis, and some subsequent halfers agree. A "double halfer" is a halfer who continues, contra Lewis, to assign *heads* credence one-half

[3]A point that seems never to be discussed in the literature is that as *Tuesday* Beauty contemplates betting on heads she might simply notice that she's a *dollar short* and, knowing what she does about her predicament, infer from this that she's a *day late*, and simply not bet. (Or better still, bet on *tails*.) If the terms of the experiment disallow knowledge that she's a dollar short and stakes are in fact a proxy for utility, it could be argued that thirders owe an account of how what Beauty doesn't know can matter to her.

upon elimination of a *tails* scenario. That is, upon learning either *Monday* or *Muesday* (the latter being like Monday if *heads* and like Tuesday if *tails*). Double halfing is halfing combined with a scheme whereby Beauty updates propositional credences in response to *de se* evidence by conditioning on the proposition corresponding to the set of worlds consistent with the evidence. Such updating is advocated in [1], [8] and [13].

Bostrom refers to his such brand of halfing as a "hybrid model". Indeed, double halfers seem to suffer from multiple personality disorder. Like Lewis, they start out ostensibly respecting *SAL*, but when a *tails* scenario is eliminated, double halfers follow thirders in assuming that the remaining scenario carries full weight. In this respect double halfing seems to be a reflex response to the most counterintuitive feature of Lewis's scheme.

The road to double halfing is paved with the best of intentions. After all, if Beauty learns *Monday* (or *Muesday*), there are no longer any indiscriminable collocated centered worlds to be found. Why would credences not revert, then, to objective chance? It sounds like a reasonable question. Arguably, though, it betrays a blindness to how peculiar halfing is in the first place. To be a halfer is to buy into *SAL*, and if you buy into *SAL*, you have to split the legitimacy of the *tails* awakenings. So if you eliminate one, you lose some legitimacy. On this view the alternative is thirding–not a novel form of halfing.

The problem with double halfing...more generally, with updating propositional credences in response to *de se* evidence by conditioning on the proposition corresponding to the set of worlds consistent with the evidence...is that it ignores the role of protocol. The locus classicus for protocol's role in updating is Monty Hall. So (with apologies)...here we go.

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.

Accordingly, halfers who update credences by conditioning on *not Door 1* are 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 $\frac{1}{3}$.[4]

[4]What else? It can't be that Beauty should update credence in *Door 3* from $\frac{1}{3}$ to $\frac{1}{2}$ upon

This "embarrassment" for double halfers differs from that of Titelbaum [12]'s in an important respect. The main consequence of the observations in [12] is that if Beauty subscribes to Elga's indifference principle and performs the fateful flip herself (hence a corresponding meaningless flip on Tuesday as well) then in order to maintain credence $\frac{1}{2}$ in Monday's flip landing *heads* she has to assign credence $\frac{5}{8}$ to the centered proposition *today's flip will land heads*. As this applies to Lewis as well, Titelbaum clearly intends for his indictment to extend to other halfers, and only singles out double halfers because Lewis has already embraced similar counterintuitive consequences in print.[5] The mishandling of Monty Hall, however, isn't merely an embarrassment...it's a deal breaker. And, it's entirely on double halfers. Lewis responds correctly to the given protocol, and, given that most double halfers probably have independent reasons for favoring halfing, I take the bulk of the suasory force of the argument to flow not from double halfing to thirding, but rather from double to Lewisian halfing.

References

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

[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] Hawley, Patrick. 2012. Inertia, Optimism and Beauty. *Nous*. 47:85-103.

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

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

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

[8] Meacham, Christopher. 2008. Sleeping Beauty and the Dynamics of *De Se* Beliefs.

learning what day it is *regardless of what day it is*. In fact that's Rosenthal [9]'s argument for thirding; alter the original problem so that the single *heads* awakening occurs on either Monday or Tuesday (with equal probabilities). Rosenthal takes it as uncontentious (double 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 $\frac{1}{3}$ as well. (Note that Rosenthal's argument doesn't work against Lewis, who takes it that Beauty's credence in *heads* upon learning *Monday* is $\frac{1}{2}$.)

[5]Not everything counterintuitive is embarrassing, and it's not clear to me why Lewis's $\frac{2}{3}$ and Titelbaum's $\frac{5}{8}$, counterintuitive as they are, should be more embarrassing than the halfer's original choice of $\frac{1}{2}$. The relevant intuitions are surely a product of *EAL*'s default status, and the original $\frac{1}{2}$ is equally bad at conforming to rational expectations under *EAL*.

Philosophical Studies 138: 24569.

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

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

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

[12] Titelbaum, Michael. 2012. An embarrassment for double halfers. *Thought* 1:146-151.

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

rmcctchn@memphis.edu