Sleeping Beauty: a critical survey (final version)

1. It's not Beauty: nor, less cutely, pretty

In case you've been systematically avoiding frustration, Sleeping Beauty (popularized by Elga 2000) is a rational agent taking part in an experiment. She is awakened Monday morning, asked her credence in *heads*, told what day it is and asked her credence again. If now the toss of a fair coin lands *heads*, she's put back to sleep until Wednesday morning. If *tails*, she's put back to sleep but, in her sleep, is given a drug that erases all memory of that morning's experiences, and then on Tuesday suffers an analogous interview. Beauty knows all of this in advance. The problem is what her personal credence in *heads* should be upon initial wakeup. A *halfer* says one-half. A *thirder* says one-third. For halfers, there is a second issue: Beauty's credence in *heads* upon learning *Monday*.

This is a critical survey (more *critical* than *survey*) of the Sleeping Beauty literature written by a mathematician. It claims to report on several cases of faulty probabilistic reasoning and a few instances of specious stewardship on the part of the philosophical community. In this brief first section, I'm going to say a bit about four examples of the latter.

1. **The uncharitable treatment of David Lewis** (who did not live to defend his position). Nowhere does it seem to be observed in the literature that Beauty's *tails* experience is a standard case of oversampling, and that halfer credences are to thirder credences as population proportions are to sample proportions in the wake of sample weight bias correction. This seems a natural first observation, and indeed looks to be implicit (on my reading) in the original paper of Lewis (2001). Lewis's provocative position is an admittedly strange fiction, but surely if the connection to statistical sampling had been made explicit its deeply resourceful inner logic would have been better appreciated.

2. **The rise of double halfing**. But Lewis died, almost everyone came to think of his counterintuitive position as crazy and a rival form of halfing ("double halfing") came to see a lot more ink. Double halfing tries to be less strange than Lewisian halfing but tests the patience of the probabilistically literate in trading in its coherence for a disturbingly cavalier repudiation of standard diachronic norms (i.e. conditionalization and reflection). Insofar as the aforementioned literate have voiced objection, they've been pushed unceremoniously to the fringe. Jenkins (2005) for example lobs a virtually unnoticed Monty Hall connection all the way from *Philosophia Mathematica* and a paper by Cian Dorr (2005) in which double halfing is completely vaporized along similar lines wasn't published at all.

3. **The celebration of a thought experiment** (and associated argument due to Jacob Ross purporting to show that there is a "deep tension" leading to "rational dilemmas" between thirder reasoning and the principle that subjective probabilities should be countably additive). One might think–either because it sounds too good to be true or because it requires (like the two envelope paradox, etc.) a physical quantity unbounded in expectation or even because the idea had been kicked around by graduate students studying stochastic analysis for decades–that philosophers would have seen nothing but red flags over this one.

Instead, it was consecrated in the *Philosopher's Annual* and the one published reply tries to construct a one-third argument that doesn't generalize to a one-fourth argument when Beauty has three awakenings if *tails*. (Because who needs good music if you've got chops?)

4. **The role of editors as self-appointed defenders of disciplinary boundaries.** There is little doubt in my mind that something (Stanley) Fishy is going on in analytic philosophy–just the place where you'd think everyone would both understand the danger and yet have the sense to steer clear. I'm not a social critic so I can't say it well, but thinking about philosophy has become an institutional activity. Philosophy is a profession. Graduate students are apprentices. Instructing them in the correct way to think about (but more especially, to write) philosophy and then purging non-conforming tokens keeps the money in the family. But at least some areas of philosophy are, by their nature, interdisciplinary. If philosophy wants to be the hand-maid of the sciences then it needs to have an at-least-adequate understanding of how science works. Responding to concerns from scientists that philosophy has lost its way with dismissive missives suggesting that philosophy will be prepared to listen when scientists have better learned the craft of philosophical writing encourages us to think of the Philosophy Department as something akin to a specialized branch of the English Department–dealing largely in fiction. That's not the view I want to have, but fear someone's been sleeping at the wheel. It's not pretty.

2. Polemics aside: this is the end, Beautiful friend

Beauty's days are numbered. But how? Well, that's the question.

Several approaches to credences are possible, including frequencies, evidence and solutions to optimization problems. In the last case one can consider gamblers' stakes and maximize *capital* (utils), or invoke information theory and minimize *surprisal*. It's the policy governing accrual of the optimized quantity that matters. The naive view, which Ross (2010) calls[1] *Every Awakening Legitimacy* (*EAL*), is that relevant quantities accrue twice if *tails*. This supports thirding (Elga 2000; Horgan 2004; Rosenthal 2009 etc.). The competing view, *Single Awakening Legitimacy* (*SAL*), according to which relevant quantities accrue precisely once, supports halfing (Bostram 2007; Hawley 2012; Jenkins 2005; Halpern 2004; D. Lewis 2001; P. Lewis 2007; Meacham 2008; Pust 2012; White 2006 etc.) provided Beauty isn't "tipped off" as to the rate of accrual.

I assume (with apologies) familiarity with some concepts in information theory and stochastic analysis. I will however keep the arguments brief. Rather than connect dots, I invite (by navigation) readers to think about Sleeping Beauty's predicament as relentlessly as I have. Steps may be skipped, but there's enough direction to make everything rigorous.

Section 3 is an elaboration on Elga's (2000) argument for thirding, which I commend. In the current climate we must view *EAL* was an unformulated premise, but it would have

[1]More or less. I use the term somewhat more generally than Ross, and with different emphasis. At any rate use of "legitimate" here is innocent. *EAL* is just a rule governing accrual of certain decision-theoretic quantities; it's not a philosophical thesis.

been unrealistic to expect Elga to anticipate the somewhat bizarre rival *SAL*. I then answer objections by Pust (2012) to Elga's practice of conditionalization on intraday evidence.

In Section 4 I explain briefly (because it's obvious) how *SAL* leads to halfing. I then give some informal justifications (not so obvious, apparently) for the plausibility of *SAL*. This step is important: *SAL must* square with the *self indicating assumption*. This is necessary to avoid the "Doomsday" type arguments that defeat what one might call "overgeneralized Lewisian halfing".

Having escaped *Doomsday* there's no reason not to dispense with double halfing. I do this in the somewhat tedious Section 5, perfunctorily explaining along the way why the Monty Hallish example I use is more ruinous than other merely "embarassing" attacks (such as Titelbaum 2012; see however Dorr 2005 for the devastating treatment alluded to above).

In Section 6 I defend countable additivity of rational credences, both by an improved direct argument and by showing that one of the assumptions Ross needs to generate "rational dilemmas" from its perceived incompatibility with thirder reasoning spawns them *by itself*– i.e., independently of thirder reasoning. Reading between the lines, I may appear to press further, suggesting that Ross's scenarios can (or should) be ignored in virtue of their high rational cost and specious plausibility. I confess to the grounding temper. In fact I would characterize faith in the metaphysical possibility of Ross's scenarios as rank superstition. However, my argument has nothing to do with my temper.

Finally, in Section 7 I make a brief comparison with a related puzzle ("The Prisoner", from Arntzenius 2003) before wrapping things up in a whimsically indulgent final section.

3. Twice told tails: thirding and conditionalization on centered evidence

Subject to *EAL*, betting and frequency arguments capably vindicate thirding, as has been argued by many others. Another type of argument is from *surprisal*.[2] 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 $-\frac{1}{2}\log_2 p - 2 \cdot \frac{1}{2} \cdot \log_2(1-p)$, minimized at $p = \frac{1}{3}$.

Arguments from evidence have spawned the most convoluted debate. Elga notes that Beauty receives no new "information" upon waking. Lewis however caution us to "Beware" what Elga intends by "information" (which for Elga can only be uncentered), and

[2]Kierland and Monton (2005) minimize expected squared difference (Brier rule) rather than surprisal. It doesn't affect their conclusions, as the Brier rule is *proper*, but it's a poor choice: if $\epsilon = 10^{-50}$ and I assign the actual world credence ϵ while you assign it ϵ^2, my credences have outperformed yours *dramatically*; but if I assign the actual world credence $\frac{1}{2} + \epsilon$ and you assign it $\frac{1}{2} + \epsilon^2$, our performances aren't importantly different. The Brier rule doesn't treat the two differences as identical (as a linear rule would), which is good, but it does treat them as comparable (within an order of magnitude), and that's not.

makes a point of using "evidence" (which can be centered or uncentered) instead. He claims that Beauty doesn't receive any of that upon waking, either. Subsequent thirders (such as Horgan 2004), however, have argued that Beauty *does* receive new evidence upon waking. My own view (which I won't be explaining unfortunately, owing to the fact that it matters little and several reviewers have used appalling misinterpretations of it to dismiss everything else I say) is that whether or not Beauty does receive new evidence upon waking depends on which of *EAL* or *SAL* she adopts as a premise.

A more satsifying argument for thirding is that of Elga. I'll give my own, slightly different rendition of it. It starts with an assumption that all parties to the debate must accept.

> *Self Indicating Assumption (SIA).* Let $h_1, h_2, \ldots, h_k$ be mutually exclusive events having known objective chances. In the absence of further evidence, an observer's credence in h_i should be proportional to the product of h_i's objective chance and the expected number of unique observers, conditional on h_i.

For some halfers, this will not be taken as an auspicious beginning. Indeed...even thirders might see it as question-begging; many think the Sleeping Beauty problem just *is* the problem of self-indication. But the idea grounding *SIA* is merely that one should take oneself to have been sampled uniformly at random from the set of observers in a representative pool of uncentered worlds in which the worlds occur in numbers proportional to their objective chance. It falls innocuously out of, say, eternal recurrence plus the frequency perspective on credences–observers who take *SIA* as evidence that they inhabit a world with comparatively high numbers of observers are vindicated in proportion to those numbers. Denying *SIA*...either because one has misunderstood the nature of the one-half solution or one has overthought, and generalized it to cases that don't involve objective chances (see footnote 4 below), is the most serious error I am attempting to expose in these pages, and I see no other way to proceed but by stating this up front and asking that readers withhold judgment until they have considered carefully how all of the pieces fit together.

Aside from its seeming inevitability to seasoned intuition, denying *SIA*, say by taking your world to have been sampled by objective chance, then you by uniform random selection from its observers (what makes you so special that you get to be in all the underpopulated worlds I can't guess...not everyone can), renders one vulnerable to the so-called *Doomsday* argument: given near-one-half objective chance of near-term human extinction and your strange views, conditioning on self-locating evidence that one is an early human leads to a conviction that near-term extinction is practically certain. (Not an acceptable end.)

The problem with applying *SIA* to the Sleeping Beauty problem is that it's not clear whether non-communicating time-slices of the same individual qualify as distinct observers. Here of course "observers" is too suggestive. What we're really interested in are *novel observations*. But do Beauty's two *tales* awakenings count as distinct such? Applied to the accrual of novel observations as a decision-theoretic quantity, *EAL* says *yes*: two novel observations if *tails*, one if *heads*. *SIA* and Elga's indifference principle (which says that

Monday and *Tuesday* should be taken as equally likely conditional on *tails*) together yield the familiar Elgain centered credences:

	Monday	*Tuesday*
Heads	$\frac{1}{3}$	0
Tails	$\frac{1}{3}$	$\frac{1}{3}$

This ends the argument; the one-third solution follows from indifference, self-indication and *EAL*.

Elga advises that Beauty ought to respond to *intraday* evidence by using an extension of classical conditionalization to probabilities defined over centered events. Upon learning *Monday*, Elga updates credence in *heads* to $\frac{1}{2}$. As an expedient, he uses days rather than moments as what one might term *coarse* world centers, and advises that Beauty do the updating by classical conditioning of her centered credence function on the centered event *Monday*, that is, on the set of centered worlds coarsely centered at Monday. Lewis condones this expedient, but Pust (2012) objects to it, noting that, technically, world centers should be moments, not days, and that, since moments are so fleeting, the supports of one's centered credence functions at any two distinct times are disjoint–which is purported to imply that one cannot have evolved from the other via conditionalization.

Cisewski et. al. (2013) have capably defused Pust's worry, but I will give my own account, which has the virtue of brevity. Pust writes:

> There is a tendency in the literature to formulate the relevant hypotheses regarding Beauty's temporal location in terms of days (e.g. "It is Monday") rather than moments. This obscures the fact that when Beauty considers whether it is Monday, she is really considering whether it is now Monday, i.e. whether the present moment occurs on Monday. When she is informed that it is Monday, she is actually learning that the (then) present moment occurs on Monday, a proposition...previously inaccessible to her.

Pust's dense prose demands time, sobriety and something like aspirin. But the point is hardly valid anyway. Though she doesn't know the global time, Beauty can at wakeup nevertheless use local coordinates (identifying surrounding times according to their position relative to $z :=$*now*) to form credences about what might occur at nearby future times. Then, when she observes what does occur, she can condition on those priors.[3] For example, when Beauty is told that it's Monday, the content of her evidence can be characterized, according to Pust, as *the present moment occurs on Monday.* Pust's claim is that this proposition was previously inaccessible to her, but it wasn't. Five minutes ago, when she awoke, she would have expressed it as *the moment five minutes in the future occurs on*

[3]If she likes, she can then recenter her credences in the present using something like what Meacham (2008) calls a *continuity principle*, but it's not really necessary.

Monday and assigned it credence $\frac{2}{3}$ (if she's a thirder). Conditioning her wakeup credence function on this event, *heads* updates to credence $\frac{1}{2}$, recovering Elga's result.

Pust appears to anticipate this counter-argument and makes a further objection. For our example, its content is that *the moment five minutes in the future occurs on Monday* entertained at wakeup, does not, even if it represents the same proposition, have the same *cognitive significance* as *the present moment occurs on Monday*. This objection makes even less sense than the first. Issues of cognitive significance only come into play when an agent entertains the same proposition in two different guises without realizing they are the same proposition. (Cf. *Hesperus is Hesperus* and *Hesperus is Phosphorus*.) That doesn't happen here, so cognitive significance is irrelevant.

4. Asked and answered: how (and why) Lewisian halfers are diluting themselves

Any argument that Beauty gains evidence on waking uses *EAL* as an implied premise (perhaps by being parasitic on the completed one-third solution). *SAL*, on the other hand, is motivated by and coheres with Lewis's claim that Beauty has no new relevant evidence.

In favor of *SAL* are several senses in which *EAL* violates the spirit of prior decision theoretic practice. If one's decision theory is grounded in utility maximization, for example, it is natural to assume that a rational agent will have access to her "utility balance" relative to some fixed point in time; in particular, it will typically supervene on her complete cognitive state. The idea here is that if the agent can't, upon ideal reflection, figure out that she's been punished or rewarded, she hasn't been. That violates *EAL* because no *Monday tails* utility exchange will survive her cognitive state reset. Another property of prior decision theoretic practice is that accrued surprisal measures total information acquired since initiation of scoring, a property clearly violated by *EAL*. Indeed, in cases where answers are withheld logarithmic scoring requires not merely that we not repeat questions; it requires questions that are independent in the information-theoretic sense. In particular, questions cannot be scored that have been previously asked and answered.

The previous paragraph is more than loose talk, as the justification Lewisians give for *SAL* determines whether they will reject *SIA*. To do so is disastrous, as it leaves one vulnerable to the Doomsday argument. (See also the *Sadistic Scientist* argument in Meacham 2008.) I believe it's evident that the justifications I have given square with *SIA*. In particular, they point to a credence in *heads* of one-third in the doppelganger version of the Sleeping Beauty problem. (That a question was asked and answered by my doppelganger has nothing whatsoever to do with me.[4]) To attack Lewis on the basis of his presumed rejection of *SIA* is simply to attack a straw man; such arguments miss their mark.

[4]Kierland and Monton (2005) think that the choice between halfing and thirding depends on one's *epistemic goals*. However, they favor the one-half solution in the doppelganger version of the problem. They are concerned that the one-third solution could fall prey to a "replicating world" thought experiment; I may be fairly certain that the actual world isn't replicating itself at regular intervals now, but if I apply thirder reasoning to the doppelganger version of the problem, I may soon have near-1 credence in that proposition.

It's straightforward that *SAL* entails the one-half solution. The information-theoretic argument is representative. Total surprisal becomes $-\frac{1}{2}\log_2 p - \frac{1}{2}\log_2(1-p)$, which is minimized at $p = \frac{1}{2}$. All that remains is to distribute the *tails* credence over Beauty's two awakenings. A plausible "indifference principle" attributed to Elga requires that this credence be distributed uniformly.

Hawley (2012) disagrees, using a principle of "inertia" to peddle wholesale disenfranchisement of Tuesday's awakening, assigning credence 1 to *Monday*. Inertia is intriguing. Indeed, it closely resembles the solution I myself favored when at first I grasped the depth of Beauty's troubles (though I preferred that Beauty assign *Muesday*–Monday if *heads*, *Tuesday* if *tails*–probability one). Like Lewis's view, these are useful fictions. Ultimately, however, they cripple an agent's capacity to respond appropriately to evidence indicating that they are inhabiting the disenfranchised location in a manner too dramatic to live down. Consider for example the case of Guildenstern, who is subjected to a two-awakening experiment with memory erasure. During the first awakening, he is given a fair coin. During the second, a seriously biased coin (it always lands *heads*). It doesn't take you or me long to figure out, to a high degree of certainty, what day it is. Guildenstern however subscribes to *inertia*, so no matter how many times he sees the coin land *heads*, he never suspects that it might be biased.

Granting Elga's indifference principle, then, *SAL* cashes out as Lewisian centered credences:

	Monday	*Tuesday*
Heads	$\frac{1}{2}$	0
Tails	$\frac{1}{4}$	$\frac{1}{4}$

Here *tails* awakenings are *diluted* relative to *heads* awakenings. As I have pointed out repeatedly (because it is so important), Lewisians, like thirders, are committed to *SIA*. However, for them *tails* observations count as "half observations". This is the standard response to oversampling; to correct sampling bias, apply lesser weights to oversampled outcomes (in inverse proportion to sample rate). On this view, halfer credences are analogous to recovered population proportions. The question, for Lewis, is always "what are the proportions in the population from which these samples are drawn?"

(See also the *Many Brains* argument in Meacham 2008 and the *Presumptuous Philosopher* argument in Bostrom 2007.) So they choose Charybdis (i.e. Doomsday) where I choose Scylla. (Double halfers throw sand.) Why don't I fear replicating worlds? All of these thought experiments betray a confusion over how *SIA* works. Worlds are sampled by objective chance, not by subjective probability. One's handling of "objective chance in contingency" cases does not commit one to a similar handling of "subjective probability in necessity" cases. In the former one believes that the contingent event occurs with asymptotic frequency $\frac{1}{2}$ (say), where in the latter one is indifferent to asymptotic frequencies of 0 and 1. The consequence is that the cases require very different treatments. Kierland and Monton, Bostrom and Meacham are all guilty of an elementary overgeneralization.

The above matrix also serves as Lewis's Sunday prior probabilties matrix, which explains Lewis's claim that Beauty learns nothing on waking. On the view of his decision theory, half of Beauty's self awakens on Monday and the other half awakens on Tuesday; no part of her legitimately experiences *both* awakenings. It also explains why *I am awakened Monday* isn't tautologous for Lewisians.[5]

Did Lewis think all of this of his own theory? Maybe and maybe not. But certainly he would have defended his approach against later attacks in precisely the way I have.

5. 'Deal' breaker: double halfers and the flouting of protocol

The most infamous artifact of dilution is that when Lewis conditions on *Monday*, his credence in *heads* jumps to $\frac{2}{3}$. Self-indication cuts both ways, and Monday's wakeup counts as just half an observation if *tails*; put another way, Monday half-awakenings confirm *heads* to the degree that they might be second half-awakenings. For thirders and some would-be halfers, this is too much. A "double halfer" is a halfer who continues, contra Lewis, to assign *heads* credence one-half upon elimination of a *tails* scenario.[6] Double halfing is halfing combined with a scheme whereby Beauty updates propositional credences in response to centered evidence by conditioning on the proposition corresponding to the set of worlds consistent with the evidence. Such updating is advocated by (Bostrom 2007; Halpern 2004; Meacham 2008; Pust 2012; White 2006).

However, it's a scheme that fails viability by virtue of its neglect of *protocol*, i.e. those infamous rules governing what evidence comes to us in various situations. The locus classicus for protocol's role in updating is Monty Hall, and in fact one defender of Lewisian halfing (Jenkins 2005) promises that a "fruitful comparison" can be made between Monty Hall and the problem of how halfers should update upon learning *Monday*. It's possible to deliver on this promise in an extremely direct way.

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*, 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.

[5]Imagine a 10% salt solution in a beaker marked "Monday". If *tails*, the experimenters add an equal portion of water and put half of the (now 5%) solution in a beaker marked "Tuesday". If we now learn, of a particular sodium ion, that it's in the beaker marked "Monday", we've learned something non-tautologous, and our credence in *heads* will increase to $\frac{2}{3}$. Such is the dilution metaphor, anyway.

[6]Or at least tries to. As shown in Dorr (2005), if there are n equally likely, mutually discriminable ways that Beauty's awakening could go down, double (but not Lewisian) halfer credence in *heads* upon observing one of them is actually $\frac{n}{3n-1}$, so that in practice double halfers are effectively thirders.

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 very error 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 $\frac{1}{3}$.[7]

This "embarrassment" for double halfers differs from that of Titelbaum's (2012) in an important respect. The main consequence of his observations is that if Beauty subscribes to Elga's indifference principle and performs the fateful toss herself (and a corresponding meaningless toss on Tuesday) then in order to maintain credence $\frac{1}{2}$ in Monday's toss landing *heads* she has to assign credence $\frac{5}{8}$ to the centered proposition *today's toss 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 similarly counterintuitive consequences in print.[8] The mishandling of Monty Hall, however, isn't merely an embarrassment...it's a deal breaker. And as Lewis responds correctly to the given protocol, it's entirely on double halfers.

6. Sleeping Methuselah: on self-indication and countable additivity

Rational credences are generally taken to be constrained by:

> *Countable Additivity (CA).* For any countable, pairwise incompatible set of propositions, the sum of one's credences in the propositions in the set must equal one's credence in their disjunction.

Not by everyone. Some question the legitimacy of the several extant Dutch Book arguments in support of *CA*. An argument with finitely many stakes should answer these questions.

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

[7] It violates reflection for Beauty to 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. All rational violations of reflection are of the same type: failure to satisfy the hypotheses of the bounded martingale stopping theorem (see Schervish et. al. 2004). That's not the case here, so double halfing is irrational.

[8] Not everything counterintuitive is embarrassing, and I see little reason why Lewis's $\frac{2}{3}$ should be more embarrassing than his original choice of $\frac{1}{2}$, which is equally (and as intentionally) bad at conforming to rational expectations under the more natural premise *EAL*–surely the source of any intuitive insult. Ditto Titelbaum's $\frac{5}{8}$ (for Lewis, anyway).

$X_1, \ldots, X_{i-1}$ 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. For large M, the P-subscribers collectively suffer a sure loss, so it's irrational to subscribe to P.

Ross (2010) doesn't reject *CA*, but he does claim that there are situations in which one is unable to subscribe to thirder reasoning while simultaneously satisfying *CA*. The one 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.

If there's any such probability function, one would imagine. But I will skip over this point. At any rate *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* is in conflict 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. So *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 (which, arguably, includes all metaphysically possible worlds). As the

example requires infinite expectation in order to do its work, it isn't clear, therefore, how to interpret Ross's reports of a deep tension between *GTP* and *CA*.

More seriously, in the context of Ross's ambitions *GTP* is a red herring. Ross argues from conflict between *GTP* and *CA* to rational dilemmas, i.e. "contexts in which full rationality is impossible". But what if the only worlds at which the conflict can arise are so crazy that *everyone* finds it impossible to achieve full rationality 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.

Ross notes that if everyone is committed to *SBI*, then everyone should reject *CA* (hence full rationality is impossible for everyone), regardless of whether they accept *GTP*. This would undermine his thesis, and he's quick to deflate it, in particular by substituting *FSBI* for *SBI*, which he hopes will pull halfers back in line with *CA*. But *SIA* is mandatory (even for halfers), so everyone taking *SBSP* at face value should *still* reject *CA*. For any world supporting faithful implementation of *SBSP* will also support a version with unique subjects, and self-indication in such contexts still runs afoul of *CA*.[9]

It's implausible anyway, but halfers can't even get out of this by rejecting *SIA*, as face-value interpretation of *SBSP* wrecks rational decision theory entirely on its own. For consider a situation in which Beauty has been sentenced to an *SBSP*-style incarceration (without memory erasure) involving mild torture. She's free to choose between two rival detention facilities (A and B) to carry out the sentence. Each has already computed the number of days they would confine her. She's chosen A, but this choice is arbitrary. Now she will be offered a sequence of two trades that she'll have to accept if she believes that her expected time of incarceration is infinite, but which will leave her worse off. First, the judge (who doesn't know the values N) offers her a halving of her sentence to switch facilities. By indifference, she accepts, and switches to B. Next, the judge asks representatives of A 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. Nevertheless she accepts, as $E(N_B) = \infty$.

The upshot is that there are finite expectation constraints on rationality. For Beauty:

> *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 $\sum_{h \in S} P(h)N(h) < \infty$.

[9] Cisewski et. al. (2013) take *SBSP* at face value but adopt the halfer position. That's beside the point; Ross's claim was that *thirders* run afoul of *CA*. Also they don't say whether they're advocates for *SIA* in distinct-subject experiments. If they are, Ross's worry will end up getting them anyway, and if they aren't, they'll face Doomsday.

Is adoption of *FE* tantamount to changing the subject? That's an irrational pessimism. If Beauty takes *SBSP* at face value, she's abandoned effective decision (and so embraced nihilism about rationality) irrespective of her views on self-indication. That self-indication can make subsequent trouble for *CA* seems quite beside the point. No...far better for Beauty to take *FE* at face value. Probably she can't go wrong, but if she does, at least she'll know that there was nothing she could have done to make it any better.

The question of the steep rational toll of eschewing finite expectation constraints vs. the power-to-model or expressiveness costs of adhering to them is interesting, but for another time. Suffice it to say it's an old topic (see, e.g., Arntzenius and McCarthy 1997 or Gallager 2014 (Chapter 6, esp. Summary[10]) quite separate from Ross's ostensive concerns.

7. The univocality of ostensive indexicals and The Prisoner

A comparison of the one-third solution to Sleeping Beauty with Arntzenius's (2003) Prisoner is enlightening. The Prisoner is waiting in his cell, where there is no clock, hoping for a stay of his scheduled execution. Right now, his credence in *I am executed* is one-half. If he is to be executed, it will happen at precisely midnight, so swiftly that he never sees it coming. At 11:59, The Prisoner will surely be alive but won't be sure whether or not it's past midnight, and will take his suspicion that it might be as partial evidence in favor of his stay having been granted. Let's assume that his internal clock (apart from his being alive) assigns *after midnight* probability one-half. Then, like Beauty's credence in *heads*, The Prisoner's credence in *executed* will have dropped to one-third.

Let's be clear: The Prisoner's credences evolve by classical conditionalization on his evidence, and his evidence is uncentered. About this there can be no debate. Suppose he sees a clock at 6 P.M. By propogation of this evidence through his time slices he learns, for any future "internal time" x, the probability $c(x)$ that the actual time is past midnight, and it's conditionalization on the further fact that he's alive at the internal time x he experiences at 11:59 (an uncentered proposition) that causes his credence in *executed* to have fallen. There is no need, in particular, for The Prisoner to appeal to self-indication.

On the other hand, it's not technically wrong for The Prisoner to forego conditionalization and take the more circuitous *SIA* route that (thirder) Beauty was forced to travel. The expected number of observations of the type characterizing his current evidence is $\frac{1}{2}$ if *executed* and 1 if *stay*, which in light of *SIA* evokes the one-third solution.[11]

[10]In particular, the paradox Ross constructs isn't new. Any null recurrent Markov chain generates seeming violations of *CA* for a finite-state mind moving in one of its orbits.

[11]Halfers too can reach for a parallel here: Sleeping Beauty's analog to The Prisoner's *I'm alive when my internal state is so-and-so* is, since Beauty's internal state is constant across wakeups, is the null assertion *I'm awakened when I'm awakened.* Thirders won't admit that as Beauty's total relevant centered evidence; *EAL* implies that *when I'm awakened* is equivocal if tails. Beauty's total relevant centered evidence, they will say, is the univocal *I am awakened now*, which may, for all Beauty knows, parse as *I am awakened Tuesday.*

8. Odds and end(game)s: why thirding is Beautiful

We've looked at four possible solutions. Thirding, Lewisian halfing, double halfing, inertia (Hawley's and the similar one assigning probability 1 to *Muesday*). What's the score?

The first property any approach to Sleeping Beauty must satisfy is conformity with the laws of probability. Everyone passes here. Next: conformity with diachronic norms (reflection and conditionalization). That eliminates double halfing, which is ugly. Reflection violations in the others are for well-documented reasons and are no cause for concern.

After that, we want an approach that stands up to our best efforts to embarrass it. Keirland and Monton, Bostrom and Meacham have thought experiments designed to embarrass thirding, but they are guilty of overgeneralization. The Doomsday arguments are often taken to embarrass Lewisian halfing, but those too fail once we realize that Lewis is embracing self-indication but tempering it with dilution. Inertia, however, is badly embarrassed by the Guildenstern thought experiment. Inertia's cute, but cute doesn't cut it.

What's the next criterion? If continuity with past decision theoretic practice, Lewis looks good. If friendliness to intuition, thirding. Halfing is my favorite. Not just out of a desire to be charitable to the philosopher I most admire, but because of its intricacy. It's only slightly crazy. A short time ago I introduced the problem to my colleague Steve Kalikow, whose Beautiful Mind harbors far better probabilistic instincts than my immanently practical one. In the course of perhaps seven minutes I'd pay good money to be able to put on YouTube (just picture a more curious, less streetwise Gabe Kaplan with astonishing brilliance and a learning disorder that makes thinking physically painful), he labored to reconstruct Lewis's solution. When I asked him what he'd do if he now learned *Monday*, he was delighted to discover that he'd have to update to $\frac{2}{3}$. Halfing really is a pretty theory.

But pretty doesn't pass, either. The deciding criterion is *endgame*: what happens to Beauty's credence in *heads* when the experiment's over? Of course we can't tell her the outcome, nor what day it is during her awakenings. But, after she gives her credence in *heads* on the last day of the experiment, we inform her that the experiment is over and that she will now go to sleep and wake up Wednesday, with no memory erasure. What's her credence in *heads* on Wednesday? On Thursday? The rest of her life? Well...it needs to be $\frac{1}{2}$. At least, that's what my intuitions tell me. But for Lewis it's $\frac{2}{3}$...forever.

Maybe it's just me. $\frac{2}{3}$ was fun for a day. Two, even. It shouldn't make a difference how long. But, like a tattoo that was cool when you were young, that has to get old. Forty years down the line, recounting that Sleeping Beauty experiment you were in: "they never told me the outcome of the toss, but I'm diluted if *tails*, so probably it was *heads*". Not that there's any grave consequence...unlike Hawley, she'll learn well enough if she's told the coin landed tails. And there *is* the bit about population proportions. Still, it's just sort of quirky. Quirky is fun and can even be pretty. But not Beautiful. Beauty requires symmetry. And to be Beautiful while you're Sleeping, well...you have to be perfect; skin the color of bread and eyes like green almonds I think it was–I'll never look into those eyes again. So Beauty, eh, that's our topic for today...hey I guess I'm a thirder in the end.

References

Arntzenius, Frank. 2003. Some problems for conditionalization and reflection. *Journal of Philosophy*. 7: 356-370.

Arntzenius, Frank and McCarthy, David. 1997. The two envelope paradox and infinite expectations. *Analysis*. 57:42-50.

Bostrom, Nick. 2007. Sleeping beauty and self location: a hybrid model. *Synthese*. 157:59-78.

Cisewski, Jessi and Kadane, Joseph B. and Schervish, Mark and Seidenfeld, Teddy and Stern, Rafael. 2013. The rest of Sleeping Beauty. Online. Accessed Sept. 4, 2014.

Dorr, Cian. 2005. A challenge for halfers. Online. Accessed Sept. 4, 2014.

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

Gallager, Robert G. 2011. *Stochastic Process: Theory for Applications*. Cambridge University Press. 2014.

Halpern, Joseph. 2004. Sleeping Beauty reconsidered: conditioning and reflection in asynchronous systems. *Oxford Studies in Epistemology*. Oxford University Press.

Hawley, Patrick. 2012. Inertia, optimism and Beauty. *Nous*. 47:85-103.

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

Jenkins, Carrie Ichikawa. 2005. Sleeping Beauty: a wake up call. *Philosophia Mathematica* 13:194-201.

Kierland, Brian and Monton, Bradley. 2005. Minimizing inaccuracy for self-locating beliefs. *Philosophy and Phenomenological Research*. 70:384-395.

Lewis, David. 2001. Sleeping Beauty: reply to Elga. *Analysis* 61:171-176.

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

McGee, Vann. 1999. An airtight Dutch book. *Analysis* 59:257-265.

Meacham, Christopher. 2008. Sleeping Beauty and the dynamics of *de se* beliefs. *Philosophical Studies* 138:24569.

Pust, Joel. 2008. Horgon on Sleeping Beauty. *Synthese* 160:97-101.

Pust, Joel. 2012. Conditionalization and essentially indexical credence. *Journal of Philosophy* 109:295-315.

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

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

Schervish, M.J., Seidenfeld, T. and Kadane, J.B. 2004. Stopping to reflect. *Journal of Philosophy*. 6:315-322.

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

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

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

rmcctchn@memphis.edu