# Rational credences are private

In this brief note I present a puzzle, or class of puzzles, that I ran across in my recent investigations into the Sleeping Beauty Problem. On the surface it looks like something of a throwaway, but the problems addressed here are actually, I believe, more central to philosophy than those that arise in the Sleeping Beauty problem proper. At the end of the paper I will make an advertisement for the solution I favor, though I think other solutions are defensible.

## *1. I can't believe Sleeping Beauty is Stupid*

According to Elga (2010), your subjective probabilities should be sharp. If $0 \leq q < p \leq 1$ then it cannot be that both $q$ and $p$ can play the "credence in proposition $P$" role for you in arbitrary betting scenarios. For if so, you will be indifferent (since $q$ can play the credence role) to a bet in which you win $1 - q$ if $P$ is true and lose $q$ if $P$ is false, as well as (since $p$ can play the credence role) a bet in which you lose $1 - p$ if $P$ is true and win $p$ if $P$ is false. It's not rational, argues Elga, to refuse both bets since by accepting both you could secure a sure gain of $p - q$. Call this the *arbitrage argument.* If it succeeds, all manner of "fuzzily" defined credences (such as interval-valued credences) are ruled out (says Elga). This would appear to include cases in which an agent "suspends judgment" as to a credence. I will refer to this claim as *sharpness.*

Meanwhile in the Sleeping Beauty thought experiment, it's generally taken as a truism that Beauty must have the same credences during each awakening.[1] So far as I know there are no objections from anyone on this point. Indeed, it's so taken for granted that it's not easy to find specific formulations of the claim, which I will refer to as *consistency.*[2] Finally, by *publicity* I will refer to the policy that Beauty reports her true credences.

Several solutions to the Sleeping Beauty problem have emerged, perhaps four of which might be considered "serious contenders." They are:

1. The one-third solution (Elga 2000). Here Beauty has credence $\frac{1}{3}$ in *heads* at wakeup, $\frac{2}{3}$ in *Monday* at wakeup and $\frac{1}{2}$ in *heads* after learning *Monday.* This is the majority view.

[1] In case you haven't heard, Sleeping Beauty is a subject in an odd experiment. She will be awakened on Monday morning, told what day it is, put back to sleep and then, if a toss of a fair coin lands *tails*, she will have all memory of Monday's awakening erased and will be awakened again on Tuesday, to again be told what day it is and put back to sleep. (If the coin lands heads, she sleeps through Tuesday.) In any event, she wakes up on Wednesday and the experiment is concluded. The problem is what Beauty's temporal credences ought to be at wakeup and after being told what day it is.

[2] I leave it open whether *consistency* follows from *sharpness.*

2. The one-half solution (Lewis 2001). Here Beauty has credence $\frac{1}{2}$ in *heads* at wakeup, $\frac{3}{4}$ in *Monday* at wakeup and $\frac{2}{3}$ in *heads* after learning *Monday.*

3. The double halfer position (Halpern 2004, White 2006; see Dorr 2005 for criticism). Here Beauty has credence $\frac{1}{2}$ in *heads* at wakeup, $\frac{3}{4}$ (or unspecified) in *Monday* at wakeup and $\frac{1}{2}$ in *heads* after learning *Monday.*

4. The optimist positions (Hawley 2012)[3]. Here Beauty has full credence in *Monday* at wakeup.

In this paper I argue that Beauty, in order to have coherent credences at all, must either (a) suspend judgment about the truth of a certain centered proposition, (b) allow herself the option of having distinct credences on the two days of the experiment, (c) keep her true credences private or (d) adopt an optimist position. The centered proposition referred to in (a) is harmless enough...in fact a version will be given in which it's simply the assertion that a certain coin currently lies heads.

**Main Claim.** If Beauty has coherent personal credences at wakeup and subscribes to all of *sharpness*, *consistency* and *publicity* then she must be an optimist.

**Proof.** At wakeup, Beauty considers the following proposition:

**C**: Either today is Monday and the coin lands *heads* or today is Tuesday and my Monday credence in "**C**" was not more than my credence in *heads.*

By "**C**" Beauty intends (on Tuesday) the proposition that was identified yesterday (Monday) by the same characters that identify **C** today (Tuesday). Since therefore "**C**" has a referent which is fixed prior to the act of interpretation fixing **C**, there is no regress that would render **C** unintelligible.[4] Indeed, **C** can be made entirely concrete:

[3] The optimist position advanced by Hawley is of course the one in which Beauty has credence $\frac{1}{2}$ in *heads* at wakeup.

[4] In contrast, Caie (2013) seeks to sabotage coherence of rational credences with:

> (*) Yuko's credence that (*) is true isn't greater than or equal to 0.5.

But this "proposition" is self-referential, and will be excluded by any reasonable extension of type theory. Good riddance; like the liar's mantra, (*) can't be unpacked. One pass:

> (*) Yuko's credence that 'Yuko's credence that (*) is true isn't greater than or equal to 0.5' is true isn't greater than or equal to 0.5.

Which is no closer to ground, so there's a regress. Failing to parse, (*) is just nonsense.

> Harvey Dent (aka Two-Face) has captured Batman and plans to subject him to Monday and possibly Tuesday awakenings. Prior to the Monday awakening, Harvey has tossed a fair coin out of view. During it, he asks Batman for his credences in *heads* = *the original toss landed heads*, *Monday* and **heads** = *the coin currently lies heads.* If now the original toss landed tails, Batman's memory of the first awakening is erased and he has a second. Prior to the second awakening, Harvey places the coin in the heads position if and only if Batman's credence in **heads** was not greater than his credence in *heads* during the first awakening, then again asks Batman's credences. Batman knows the protocol. Harvey promises Batman that he'll die if he fails to respond with coherent credences.

Assuming *consistency*, Let $h$, $\mathbf{h}$ and $m$ be Batman's credences in *heads*, **heads** and *Monday*, respectively. Now **heads** holds if either this is Monday and *heads* or if this is Tuesday and $\mathbf{h} \leq h$. In other words, **heads** is co-extensive with **C**. We must establish that $m = 1$.

If $m = 0$ then $h = 0$, which means that (modulo null sets) **heads** holds if and only if $\mathbf{h} = Cred(\mathbf{heads}) = 0$, and Batman's credences are incoherent. So we may assume $m > 0$.

Case 1: $\mathbf{h} > h$. Here **heads** is coextensive with *heads*, as the coin lies tails on Tuesday. So $\mathbf{h} = Cred(\mathbf{heads}) = Cred(heads) = h$, a contradiction.

Case 2: $\mathbf{h} \leq h$. Here **heads** is true if it's Monday and *heads* or if it's Tuesday. So

$$\begin{aligned} h &\geq \mathbf{h} \\ &= Cred(\mathbf{heads}) \\ &= Cred(Monday) \cdot Cred(\mathbf{heads}|Monday) + Cred(Tuesday) \cdot Cred(\mathbf{heads}|Tuesday) \\ &= m \cdot \frac{h}{m} + (1-m) \cdot 1 = h + (1-m) \geq h. \end{aligned}$$

It follows that the inequalities are equalities, which implies that $m = 1$. □

*2. It was a Dark and unideal Knight*

So either *optimism* is necessary or *publicity, consistency,* and *sharpness* fail to cohere; Batman must either live with the former or get rid of one of the latter.

Jettisoning *publicity* seems an attractive option for Batman if he can pull it off, but it isn't clear that he can. Thought-reading devices that could detect his true credences, or at least detect lying, seem logically and perhaps even nomologically possible. Doing away with *consistency* looks problematic; indeed, I see no way of doing it except by employing some random process, in which case it would be difficult to defend the claim that the values generated represented credences at all. Finally to deny *sharpness*, say by suspending judgment in the Batman game, looks problematic (at least at first blush), as we could just require that Two-Face always place the coin in the heads position when Batman suspends judgment, so that suspension becomes as self-defeating and unstable a position as any sharp credence. (Though I don't favor it, I do nevertheless think that it's coherent to deny *sharpness.* See below.)

And yet *optimism* seems an affront to the intuition that, for all Batman knows, it might well be Tuesday. Moreover, adopting *optimism* doesn't appear to get one out of the woods anyway, as the problem can be made to appear outside of Sleeping Beauty experiments. Consider the following.

> The Great Predictor has placed a coin under a cup and then asked you to give your credence in *heads.* However, the Great Predictor assures you that his prediction skills are very great indeed, and that he has placed the coin in the heads position if, and only if, your credence $h$ in *heads* will be not more than $\frac{1}{2}$.

Like the Batman case, The Great Predictor example purports to subject a rational agent (you) to a bout of "believed-of-oneself anti-expertise": any credence in *heads* you might adopt appears to be self-defeating, despite the fact that what your credences are can't seem to affect the status of the coin, which was placed before your credences were determined.

Many philosophers view "Great Predictors" with skepticism. (As David Lewis once said of the original Newcomb thought experiment, "Some philosophers have refused to learn anything from such a tall tale.") At any rate, what I think the Predictor scenario shows is that it's either *publicity* or *sharpness* that must be parted with in the Batman case, as denial of *consistency* or *optimism* won't carry over to the Predictor case anyway.

The issue even appears to arise in the following "low budget" experiment: I (who have no memory-erasing drug or prediction skills) ask your credence that I will presently place a coin in the heads position after promising that I will place it in the heads position if and only if your reported credence is not greater than $\frac{1}{2}$. Of course I can't read your mind, so the temptation to drop *publicity* here is strong. Otherwise, *sharpness* is in trouble. Of course, one could try to selectively create exclusions to *sharpness*...not require yourself to have credence in any event lying in the causal cone of your would-be credence in it.

But that's problematic for decision theory. Can I not have a credence in the proposition that I will die in a traffic accident, because such credence will affect how I decide to drive? Probably in that case there's some equilibrium credence that makes sense. But do equilibrium credences always make sense? If a madman asks your credence $r$ that he will kill you in five minutes after promising to do so with probability $1 - r^2$, must you necessarily adopt $r = \frac{-1+\sqrt{5}}{2}$? You would much rather answer with something like $r = 1$.

But enough. If Elga's argument is all *sharpness* has going for it then trashing it looks plausible, as the arbitrage argument doesn't seem to work in the problem cases, try as one might to make it so. (Details, which are murky indeed, left to the reader.) Nevertheless, I prefer to bite the bullet and get rid of *publicity* instead. My intuition is that I should always have credences. *Privacy* allows me to have stable ones. It's also something of a godsend for the evidential version of decision theory, which is almost surely a good thing. These considerations look weightier to me than does the threat of a mind-reading device that no one, not even an ideal rational agent, can fool.

Of course some (including David Lewis) have claimed that one should adopt a decision theory appropriate to a *non-ideal* rational agent. I disagree. In order for a theory to persist in the face of a departure from idealness, there has to be some systematicity in the departure. Real gasses, for example, may have behavior that is not explained precisely by the ideal gas law, but, owing to the systematicity of the departure of real gasses from idealness, scientists may be able to gather empirical data and build a more predictive (if more complicated) theory for real gasses. But what goes for gasses doesn't go for agents striving to behave rationally, because to whatever extent their irrationality is systematic, it can be corrected, for, as Hegel put it, "the very fact that something is determined as a limitation implies that the limitation is already transcended."

Yes, I quoted Hegel. No, that doesn't make for much of an argument. But then this isn't much of a paper.

---

Caie, Michael. 2013. Rational Probabilistic Incoherence. *Philosophical Review.* 122:527-575.

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.

Elga, Adam. 2010. Subjective probabilities should be sharp. *Philosophers' Imprint.* 10(05).

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.

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

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

*rmcctchn@memphis.edu*