Self-locating belief and the Sleeping Beauty problem

Adam Elga

In addition to being uncertain about what the world is like, one can also be uncertain about one's own spatial or temporal location in the world. My aim is to pose a problem arising from the interaction between these two sorts of uncertainty, solve the problem, and draw two lessons from the solution.

1. *The Sleeping Beauty problem*[1]

Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you to back to sleep with a drug that makes you forget that waking.[2]

When you are first awakened, to what degree ought you believe that the outcome of the coin toss is Heads?

First answer: 1/2, of course! Initially you were certain that the coin was fair, and so initially your credence in the coin's landing Heads was 1/2. Upon being awakened, you receive no new information (you knew all along that you would be awakened). So your credence in the coin's landing Heads ought to remain 1/2.

Second answer: 1/3, of course! Imagine the experiment repeated many times. Then in the long run, about 1/3 of the wakings would be Heads-wakings – wakings that happen on trials in which the coin lands Heads. So

[1] So named by Robert Stalnaker (who first learned of examples of this kind in unpublished work by Arnold Zuboff). This problem appears as Example 5 of Piccione 1997, which motivates two distinct answers but suspends judgment as to which answer is correct (1997: 12–14). Aumann 1997 uses a fair lottery approach to analyse a similar problem. Adapted to the Sleeping Beauty problem, that analysis yields the same answer as the one I will defend in section 2. However, unlike the argument in Aumann 1997, my argument does not depend on betting considerations.

[2] The precise effect of the drug is to reset your belief-state to what it was just before you were put to sleep at the beginning of the experiment. If the existence of such a drug seems fanciful, note that it is possible to pose the problem without it – all that matters is that the person put to sleep believes that the setup is as I have described it.

Analysis 60.2, April 2000, pp. 143–47.

on any particular waking, you should have credence 1/3 that that waking is a Heads-waking, and hence have credence 1/3 in the coin's landing Heads on that trial. This consideration remains in force in the present circumstance, in which the experiment is performed just once.

I will argue that the correct answer is 1/3.

2. Suppose that the first waking happens on Monday, and that the second waking (if there is one) happens on Tuesday. Then when you wake up, you're certain that you're in one of three 'predicaments':

H_1 HEADS and it is Monday.
T_1 TAILS and it is Monday.
T_2 TAILS and it is Tuesday.

Notice that the difference between your being in T_1 and your being in T_2 is not a difference in which possible world is actual, but rather a difference in your temporal location within the world. (In a more technical treatment we might adopt a framework similar to the one suggested in Lewis 1983, according to which the elementary alternatives over which your credence is divided are not possible worlds, but rather centred possible worlds: possible worlds each of which is equipped with a designated individual and time. In such a framework, H_1, T_1, and T_2 would be represented by appropriate sets of centred worlds.)

Let P be the credence function you ought to have upon first awakening. Upon first awakening, you are certain of the following: you are in predicament H_1 if and only if the outcome of the coin toss is Heads. Therefore, calculating $P(H_1)$ is sufficient to solve the Sleeping Beauty problem. I will argue first that $P(T_1) = P(T_2)$, and then that $P(H_1) = P(T_1)$.

If (upon first awakening) you were to learn that the toss outcome is Tails, that would amount to your learning that you are in either T_1 or T_2 . Since being in T_1 is subjectively just like being in T_2, and since exactly the same propositions are true whether you are in T_1 or T_2, even a highly restricted principle of indifference yields that you ought then to have equal credence in each. But your credence that you are in T_1, after learning that the toss outcome is tails, ought to be the same as the conditional credence $P(T_1|T_1$ or $T_2)$, and likewise for T_2. So $P(T_1|T_1$ or $T_2) = P(T_2|T_1$ or $T_2)$, and hence $P(T_1) = P(T_2)$.

The researchers have the task of using a fair coin to determine whether to awaken you once or twice. They might accomplish their task by either

(1) *first* tossing the coin and then waking you up either once or twice depending on the outcome; or
(2) first waking you up once, *and* then tossing the coin to determine whether to wake you up a second time.

Your credence (upon awakening) in the coin's landing Heads ought to be the same regardless of whether the researchers use method (1) or (2). So without loss of generality suppose that they use – and you know that they use – method (2).

Now: if (upon awakening) you were to learn that it is Monday, that would amount to your learning that you are in either H_1 or T_1. Your credence that you are in H_1 would then be your credence that a fair coin, soon to be tossed, will land Heads. It is irrelevant that you will be awakened on the following day if and only if the coin lands Tails – in this circumstance, your credence that the coin will land Heads ought to be 1/2. But your credence that the coin will land Heads (after learning that it is Monday) ought to be the same as the conditional credence $P(H_1|H_1$ or $T_1)$. So $P(H_1|H_1$ or $T_1) = 1/2$, and hence $P(H_1) = P(T_1)$.

Combining results, we have that $P(H_1) = P(T_1) = P(T_2)$. Since these credences sum to 1, $P(H_1) = 1/3$.

3. Let H be the proposition that the outcome of the coin toss is Heads. Before being put to sleep, your credence in H was 1/2. I've just argued that when you are awakened on Monday, that credence ought to change to 1/3. This belief change is unusual. It is not the result of your receiving new information – you were already certain that you would be awakened on Monday.[3] (We may even suppose that you knew at the start of the experiment exactly what sensory experiences you would have upon being awakened on Monday.) Neither is this belief change the result of your suffering any cognitive mishaps during the intervening time – recall that the forgetting drug isn't administered until well after you are first awakened. So what justifies it?

The answer is that you have gone from a situation in which you count your own temporal location as irrelevant to the truth of H, to one in which you count your own temporal location as relevant to the truth of H.[4] Suppose, for example, that at the start of the experiment, you weren't sure whether it was 1:01 or 1:02. At that time, you counted your temporal location as irrelevant to the truth of H: your credence in H, conditional on its

[3] To say that an agent receives new information (as I shall use that expression) is to say that the agent receives evidence that rules out possible worlds not already ruled out by her previous evidence. Put another way, an agent receives new information when she learns the truth of a proposition expressible by an eternal sentence (Quine 1960: 191) of some appropriately rich language.

[4] To say that an agent counts her temporal location as relevant to the truth of a certain proposition is to say that there is a time t such that the agent's beliefs are compatible with her being located at t, and her credence in the proposition, conditional on her being located at t, differs from her unconditional credence in the proposition.

being 1:01, was 1/2, and your credence in *H*, conditional on its being 1:02, was also 1/2.

In contrast (assuming that you update your beliefs rationally), when you are awakened on Monday you count your current temporal location as relevant to the truth of *H*: your credence in *H*, conditional on its being Monday, is 1/2, but your credence in *H*, conditional on its being Tuesday, is 0. On Monday, your unconditional credence in *H* differs from 1/2 because it is a weighted average of these two conditional credences – that is, a weighted average of 1/2 and 0.

It is no surprise that the manner in which an agent counts her own temporal location as relevant to the truth of some proposition can change over time. What is surprising – and this is the first lesson – is that this sort of change can happen to a perfectly rational agent during a period in which that agent neither receives new information nor suffers a cognitive mishap.

At the start of the experiment, you had credence 1/2 in *H*. But you were also certain that upon being awakened on Monday you would have credence 1/3 in *H* – even though you were certain that you would receive no new information and suffer no cognitive mishaps during the intervening time. Thus the Sleeping Beauty example provides a new variety of counterexample to Bas Van Fraassen's 'Reflection Principle' (1984: 244, 1995: 19), even an extremely qualified version of which entails the following:

> Any agent who is certain that she will tomorrow have credence *x* in proposition *R* (though she will neither receive new information nor suffer any cognitive mishaps in the intervening time) ought *now* to have credence *x* in *R*.[5]

David Lewis once asked 'what happens to decision theory if we [replace the space of possible worlds by the space of centred possible worlds]?' and answered 'Not much'. (Lewis 1983: 149) A second lesson of the Sleeping Beauty problem is that something does happen: a new question arises about how a rational agent ought to update her beliefs over time.[6]

Massachusetts Institute of Technology
Cambridge, MA 02139-4307, USA
adam@mit.edu

[5] I am indebted to Ned Hall for pointing out that an answer of 1/3 conflicts with the Reflection Principle.

[6] Many thanks to Jamie Dreier, Gary Gates, Ned Hall, Vann McGee, Robert Stalnaker, Roger White, Sarah Wright, the participants in a 1999 conference at Brown University (at which an earlier version of this paper was presented), and an anonymous referee.

References

Aumann, R. J., S. Hart, and M. Perry. 1997. The forgetful passenger. *Games and Economic Behavior* 20: 117–20.

Lewis, D. 1983. Attitudes *de dicto* and *de se*. In his *Philosophical Papers, Volume I*, 133–159. New York: Oxford University Press.

Piccione, M. and A. Rubenstein. 1997. On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior* 20: 3–24.

Quine, W. V. 1960. *Word and Object*. Cambridge, Mass.: The MIT Press.

van Fraassen, B. C. 1984. Belief and the will. *Journal of Philosophy* 81: 235–56.

van Fraassen, B. C. 1995. Belief and the problem of Ulysses and the sirens. *Philosophical Studies* 77: 7–37.

Analysis 60.2, April 2000, p. 147. © V. H. Dudman