ANALYSIS 61.3 JULY 2001

Sleeping Beauty: reply to Elga

DAVID LEWIS

1. *The problem*

Researchers at the Experimental Philosophy Laboratory have decided to carry out the following experiment. First they will tell Sleeping Beauty all that I am about to tell you in this paragraph, and they will see to it that she fully believes all she is told. Then on Sunday evening they will put her to sleep. On Monday they will awaken her briefly. At first they will not tell her what day it is, but later they will tell her that it is Monday. Then they will subject her to memory erasure. Perhaps they will again awaken her briefly on Tuesday. Whether they do will depend on the toss of a fair coin: if heads they will awaken her only on Monday, if tails they will awaken her on Tuesday as well. On Wednesday the experiment will be over and she will be allowed to wake up. The three possible brief awakenings during the experiment will be indistinguishable: she will have the same total evidence at her Monday awakening whatever the result of the coin toss may be, and if she is awakened on Tuesday the memory erasure on Monday will make sure that her total evidence at the Tuesday awakening is exactly the same as at the Monday awakening. However, she will be able, and she will be taught how, to distinguish her brief awakenings during the experiment from her Wednesday awakening after the experiment is over, and indeed from all other actual awakenings there have ever been, or ever will be.

Let's assume that Beauty is a paragon of probabilistic rationality, and always assigns the credences (subjective probabilities) she ought to. We shall need to consider her credence functions at three different times. Let P be her credence function just after she is awakened on Monday. Let P_+ be her credence function just after she's told that it's Monday. Let P_- be her credence function just before she's put to sleep on Sunday, but after she's been told how the experiment is to work.

At the beginning of her Monday awakening, what credence does Beauty assign to the hypothesis HEADS that the result of the coin toss is heads? What credence does she assign to the hypothesis TAILS that it's tails? Adam Elga (2000) argues that P(HEADS) = 1/3, P(TAILS) = 2/3. I disagree, and argue that P(HEADS) = P(TAILS) = 1/2.

ANALYSIS 61.3, July 2001, pp. 171–76. © David Lewis

I haven't said yet whether the coin was to be tossed before or after the Monday awakening. Elga's argument applies in the first instance to the case that it is tossed after; but he thinks, and I agree, that the answer to our question should be the same in both cases. My argument will apply equally to both cases.

2. *Common ground*

What gives our disagreement much of its interest is that we agree on so much else (including much that not everyone would agree with). Let me begin by running through the undisputed common ground.

We agree that there are two kinds of possibilities to which credences may be given. There are possibilities about what sort of possible world is actual; and there are possibilities about who one is and when one is and what sort of world one lives in. Following Quine (1969), we shall represent the latter possibilities as classes of *centred worlds*: possible worlds with designated individuals-at-times within them. Call the classes of centred worlds *centred possibilities*. (We could represent the former possibilities, the *uncentred possibilities*, as classes of *un*centred possible worlds; but we needn't bother, since we can subsume the uncentred possibilities under the centred ones.) It may happen that two centred worlds are situated within the same uncentred possible world: only their designated individuals-at-times differ. If so, I call them *collocated*.

When Beauty awakens during the experiment, three centred epistemic possibilities are compatible with her total evidence:

H_1: HEADS and it's Monday,
T_1: TAILS and it's Monday,
T_2: TAILS and it's Tuesday.

Elga writes, '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' (144). By 'proposition' he means an uncentred possibility. The reason the same propositions are true whether Beauty is in T_1 or T_2 is that the centred worlds that are members of T_1 are collocated with the corresponding members of T_2.

I accept Elga's 'highly restricted principle of indifference'.[1] So we have a further point of agreement:

[1] By ignoring the collocation of corresponding members of the two epistemic possibilities, we would get a less restricted principle of indifference, which would tell us that $P(H_1) = P(T_1) = P(T_2)$. That would afford a swift shortcut to Elga's conclusion – much too swift, and Elga is wise to have nothing to do with it. It has bizarre consequences: for instance, that it makes exactly no difference to the equality of $P(H_1)$ and $P(T_1)$ if,

(1) $P(T_1) = P(T_2)$.

When, part-way through her Monday awakening, Beauty is told that it's Monday, her credence function changes from P to P_+. Elga and I agree that this change takes place by conditioning on her new evidence, which can be expressed as not-T_2, or as $(H_1 \vee T_1)$. So

(2) $P_+(\text{HEADS}) = P(\text{HEADS} \mid H_1 \vee T_1)$,
$P_+(\text{TAILS}) = P(\text{TAILS} \mid H_1 \vee T_1)$.

Her total evidence at the start of her Monday awakening tells her that HEADS is true iff H_1 is true of her at that time; and likewise for TAILS and $(T_1 \vee T_2)$. It is routine to restate a conditional credence as a quotient of unconditional credences. So we agree that

(3) $P(\text{HEADS} \mid H_1 \vee T_1) = P(H_1 \mid H_1 \vee T_1) = P(H_1)/[P(H_1) + P(T_1)]$,
$P(\text{TAILS} \mid H_1 \vee T_1) = P([T_1 \vee T_2] \mid H_1 \vee T_1) = P(T_1)/[P(H_1) + P(T_1)]$,

and further that

(4) $P(\text{HEADS}) = P(H_1)$,
$P(\text{TAILS}) = P(T_1 \vee T_2) = P(T_1) + P(T_2)$.

We further agree that

(5) $P\_(\text{HEADS}) = 1/2 = P\_(\text{TAILS})$,

(6) Beauty gains no new uncentred evidence, relevant to HEADS versus TAILS, between the time when she has credence function P_- and the time when she has credence function P. The only evidence she gains is the centred evidence[2] that she is presently undergoing either the Monday awakening or the Tuesday awakening: that is, $(H_1 \vee T_1 \vee T_2)$.

Here agreement ends. Now I shall look at our respective analyses of the problem, including our different conclusions about P(HEADS).

3. *Elga's argument*

Elga, considering only the case that the coin toss comes after the Monday awakening, argues from later to earlier.

(E1) $P_+(\text{HEADS}) = 1/2$ (Premiss).
(E2) $\therefore P(H_1)/[P(H_1) + P(T_1)] = 1/2$ (E1, 2, 3).
(E3) $\therefore P(H_1) = P(T_1)$ (E2).

instead of being told that the coin is fair, Beauty is instead told that it is biased 999 to 1 in favour of heads!

[2] Beware: Elga speaks of 'new information'. But in his terminology 'centred information' doesn't count as 'information' at all (145, n. 3). That needn't stop him from calling it 'evidence'.

(E4) $\therefore$ $P(H_1) = P(T_1) = P(T_2) = 1/3$ (E3, 1).
(E5) $\therefore$ $P(\text{HEADS}) = 1/3$, $P(\text{TAILS}) = 2/3$ (E4, 4).

Quod erat demonstrandum; but we note two further consequences.

(E6) $\therefore$ $P_+(\text{HEADS}) = P(\text{HEADS}) + 1/6$ (E1, E5).
(E7) $\therefore$ A change in credence from P_- to P was not produced by new relevant uncentred evidence; either it was not produced by relevant evidence at all, or else the centred evidence $(H_1 \vee T_1 \vee T_2)$ was relevant to HEADS versus TAILS (E5, 5, 6).

4. *My argument*

I, considering both cases, argue from earlier to later.

(L1) Only new relevant evidence, centred or uncentred, produces a change in credence; and the evidence $(H_1 \vee H_2 \vee H_3)$ is not relevant to HEADS versus TAILS (Premiss).
(L2) $\therefore$ $P(\text{HEADS}) = 1/2 = P(\text{TAILS})$ (L1, 5, 6).

Quod erat demonstrandum; but we note five further consequences.

(L3) $\therefore$ $P(H_1) = 1/2 = P(T_1) + P(T_2)$ (L2, 4).
(L4) $\therefore$ $P(T_1) = 1/4 = P(T_2)$ (L3, 1).
(L5) $\therefore$ $P(\text{HEADS} \mid H_1 \vee T_1) = 2/3$,
$P(\text{TAILS} \mid H_1 \vee T_1) = 1/3$ (L3, L4, 3).
(L6) $\therefore$ $P_+(\text{HEADS}) = 2/3$,
$P_+(\text{TAILS}) = 1/3$ (L5, 2).
(L7) $\therefore$ $P_+(\text{HEADS}) = P(\text{HEADS}) + 1/6$ (L2, L6).

5. *The shape of the disagreement*

Elga rejects my premiss: his (E7) contradicts my (L1). I reject Elga's premiss: my (L6) contradicts his (E1). That's the entire source of our disagreement. We must reject one premiss or the other. Given both (and the agreed-upon common ground) we end up with an antinomy. What more can be said?

Elga's reason for rejecting my premiss is that he is following where argument leads. He regards it as a surprising discovery that '[a change in credence] can happen to a perfectly rational agent during a period in which that agent neither receives new information' – that is, no new uncentred evidence – 'nor suffers a cognitive mishap' (146). In my view it would be equally surprising, and equally suspect, to discover that the centred evidence $(H_1 \vee T_1 \vee T_2)$ was relevant to HEADS versus TAILS; or to discover that Beauty had suffered a cognitive mishap. (Memory erasure is indeed a cognitive mishap; but that happens later, so it is irrelevant to the change at issue between P_- and P.) Fair enough; but it would be nicer if one's rejection of the other's premiss were independently motivated.

I, on the other hand, claim that my reason for rejecting Elga's premiss *is* independently motivated. Where did Elga get it? At the time of P_+, in the case Elga is considering in which the coin toss happens after that time, Beauty knows that there will be a future toss of a fair coin. There is a well-known principle which says that credences about future chance events should equal the known chances. (See Mellor 1971; Lewis 1980.) It is just this principle that gave us the agreed-upon (5), applying both to the case that the coin toss will happen before the Monday awakening and to the case that it will happen after. The same principle would seem to say also that $P_+(\text{HEADS}) = \text{chance}(\text{HEADS}) = 1/2$ and that $P_+(\text{TAILS}) = \text{chance}(\text{TAILS}) = 1/2$, thereby providing Elga's premiss (E1).

I reply that the principle requires a proviso, which was satisfied when we used it to give us (5), but which is not satisfied when Elga uses it to give him his premiss (E1). Imagine that there is a prophet whose extraordinary record of success forces us to take seriously the hypothesis that he is getting news from the future by means of some sort of backward causation. Seldom does the prophet tell us outright what will happen, but often he advises us what our credences about the outcome should be, and sometimes his advice disagrees with what we would get by setting our credences equal to the known chances. What should we do? If the prophet's success record is good enough, I say we should take the prophet's advice and disregard the known chances.

Now when Beauty is told during her Monday awakening that it's Monday, or equivalently not-T_2, she is getting evidence – centred evidence – about the future: namely that she is not now in it. That's new evidence: before she was told that it was Monday, she did not yet have it. To be sure, she is not getting this new evidence from a prophet or by way of backward causation, but neither is she getting it just by setting her credences equal to the known chances. The news is relevant to HEADS, since it raises her credence in it by 1/6; see my (L7). Elga agrees; see his (E6). Therefore the proviso applies, and we cannot rely on it that $P_+(\text{HEADS}) = \text{chance}(\text{HEADS})$ and $P_+(\text{TAILS}) = \text{chance}(\text{TAILS})$. I admit that this is a novel and surprising application of the proviso, and I am most grateful to Elga for bringing it to my attention. Nevertheless I find it fairly convincing, independently of wishing to follow where my argument leads.[3]

Princeton University
Princeton, NJ 08544, USA

[3] I thank Adam Elga and the Editor of *Analysis* for valuable comments.

References

Elga, A. 2000. Self-locating belief and the Sleeping Beauty problem. *Analysis* 60: 143–47.

Lewis, D. 1980. A subjectivist's guide to objective chance. In *Studies in Inductive Logic and Probability*, ed. R. C. Jeffrey, Vol. 2, 263–93. Berkeley: University of California Press. Reprinted in D. Lewis, *Philosophical Papers*, Vol. 2. Oxford: Oxford University Press.

Mellor, D. H. 1971. *The Matter of Chance*. Cambridge: Cambridge University Press.

Quine, W. V. 1969. Propositional objects. In his *Ontological Relativity & Other Essays*. New York: Columbia University Press.

Analysis 61.3, July 2001, pp. 176–87. © Jc Beall