Sleeping Beauty: in defence of Elga

Cian Dorr

The story of the Sleeping Beauty, in the version I will be concerned with, is this:

> Sleeping Beauty is a paradigm of rationality. On Sunday she learns for certain that she is to be the subject of an experiment. The experimenters will wake her up on Monday morning, and tell her some time later that it is Monday. When she goes back to sleep, they will toss a fair coin. If the outcome of the toss is Heads, they will do nothing. If the outcome is Tails, they will administer a drug whose effect is to destroy all memories from the previous day, so that when she wakes up on Tuesday, she will be unable to tell that it is not Monday.

The paradox of the Sleeping Beauty is this: Let P be her credence distribution immediately after waking up on Monday. Let P_+ be her credence distribution after having been told that it is Monday. Let HEADS be the proposition that the coin lands Heads. Let MONDAY be the 'centred proposition' that it is Monday. Then each of the following claims seems very plausible:

(1) $P(\text{HEADS}) = 1/2$
(2) $P_+(\text{HEADS}) = 1/2$
(3) $P_+(\text{HEADS}) = P(\text{HEADS}|\text{MONDAY})$
(4) $P(\text{HEADS}|\text{not-MONDAY}) = 0$
(5) $0 < P(\text{MONDAY}) < 1$

But these propositions are mutually inconsistent, given the probability calculus.

Proof
$P(\text{HEADS}) = P(\text{HEADS}|\text{MONDAY}) \cdot P(\text{MONDAY}) + P(\text{HEADS}|\text{not-MONDAY}) \cdot P(\text{not-MONDAY})$, by (5).
$P(\text{HEADS}) = P(\text{HEADS}|\text{MONDAY}) \cdot P(\text{MONDAY})$, by (4).

Analysis 62.4, October 2002, pp. 292–96.

$P(\text{HEADS}) = P_+(\text{HEADS}) \cdot P(\text{MONDAY})$, by (3).
$1/2 = 1/2 \cdot P(\text{MONDAY})$ by (1) and (2).
So $P(\text{MONDAY}) = 1$, contradicting (5).

Elga (2000) and Lewis (2001) both accept (3), (4) and (5), and I agree. Elga advocates giving up (1): he claims that P(HEADS) is 1/3, not 1/2. Lewis advocates giving up (2): he claims that P_+(HEADS) is 2/3, not 1/2.

Elga's case for (2) is based on the principle that when one is certain that a fair coin will be tossed in the future, one ought to believe to degree 1/2 that the outcome will be Heads. I agree that this principle gives us a strong prima facie reason to accept (2). (As Lewis points out, there are certain exotic possibilities, involving prophets, crystal balls and the like, in which this principle seems to fail; but any analogy between these strange cases and Beauty's situation is, to say the least, not obvious.) Nevertheless, I don't think that Elga has given us any reason to accept (2) *instead* of (1). One's beliefs about the temporal location of a coin-toss are relevant to the question what one should believe about its outcome only because (in the absence of crystal balls) we are limited in the sorts of evidence we can have about future coin-tosses. But often our evidence about coin-tosses we believe to be in the past is limited in the same way, and, when it is, our credence about the outcome of the toss is subject to the same constraint. For example, if I learn from a history-book that a certain fair coin was tossed, but am not told what the outcome was, or given any other information about matters causally connected to the outcome, I should believe to degree 1/2 that this coin landed Heads. So Elga's principle is plausible only because it is derived from a more general, if harder to state, principle to the effect that when one lacks evidence of a certain sort ('inadmissible' evidence, in the terminology of Lewis 1980) about the toss of a coin which one knows to be fair, one should believe to degree 1/2 that it lands Heads. Unfortunately, however we fill in the details of this more general principle, it is hard to see how it could support (2) any more strongly than it supports (1). When Beauty is told that it is Monday, it seems intuitively that she gains new evidence without losing any. So however plausible it is that she lacks inadmissible evidence about the coin-toss afterwards, it is at least as plausible that she also lacked inadmissible evidence about the coin-toss to begin with.

Thus, we should not expect our intuitive judgments about the constraints on rational belief about coin-tosses to decide between (1) and (2): it is these intuitive judgments that lead to the paradox in the first place. Nevertheless, I am persuaded that Elga's answer is the correct one. What persuades me is the following variation on the story:

> Again, Sleeping Beauty knows for certain on Sunday that she is to be the subject of an experiment. This time, the experimenters will defi-

> nitely wake her both on Monday and on Tuesday, administering an amnesia-inducing drug between the two awakenings. However, they have two amnesia-inducing drugs, and they will decide which one to administer by tossing a fair coin on Monday night. If the outcome of the toss is Tails, they will administer the amnesia-inducing drug that was used in the original version of the experiment. If the outcome is Heads, they will administer a much weaker amnesia-inducing drug, which merely *delays* the onset of memories from the previous day, rather than destroying them entirely. If Beauty receives this weaker drug, the first minute of her awakening on Tuesday will be just as it would have been if she had received the stronger drug, but after that the memories of Monday's awakening will come flooding back. She will then realize that it is Tuesday, and that the outcome of the toss must have been Heads.

Let Q_- be Beauty's credence function immediately after being woken on Monday in the variant case. Clearly Q_- should give positive credence to each of the following four hypotheses:

H_1 The coin lands Heads and it's Monday
H_2 The coin lands Heads and it's Tuesday
T_1 The coin lands Tails and it's Monday
T_2 The coin lands Tails and it's Tuesday

How should this credence be distributed? It seems to me that only one answer to this question is remotely credible: namely, $Q_-(H_1) = Q_-(H_2) = Q_-(T_1) = Q_-(T_2) = 1/4$. For during the first minute after she has woken up, the difference between the strong and the weak amnesia drugs is completely irrelevant from Beauty's point of view. Her credences should be just as they would have been if the experimenters had resolved to administer the strong drug in any case, so that nothing at all depended on the outcome of the coin-toss. And surely we must agree that in *that* case, the four hypotheses should get equal credence.

Let Q be Beauty's credence function after a minute has passed on Monday. On the assumption that she has sufficient introspective powers to be absolutely certain that she has not experienced the flooding-back of memories she would have experienced had H_2 been true, $Q(H_2) = 0$. But nothing in her experience during the first minute does anything to discriminate between the other three hypotheses. So the ratio of her credences in H_1, T_1 and T_2 will remain unchanged: that is, her credences in these hypotheses will be updated by conditionalizing on the negation of H_2. Hence, $Q(H_1) = Q(T_1) = Q(T_2) = 1/3$; so $Q(\text{HEADS}) = Q(H_1) = 1/3$.

This strikes me as a decisive reason to side with Elga's view about the rational distribution of credence in the original case. Of course, the evi-

dence Beauty has after a minute has gone by in the variant case is not *exactly* the same as the evidence she has immediately after she has woken up in the original case. In the variant case, her evidence includes memories of the minute that has passed since waking up, as well as memories of having been told that the variant experiment, rather than the original one, was to be performed. But could these differences really be relevant to the question what Beauty's credence in HEADS ought to be? I can't see how they could: once a minute has passed, the question whether it is the variant experiment or the original one that is being performed seems utterly immaterial from Beauty's point of view. To support this judgment, we can imagine a series of cases, in which the weaker amnesia drug is made successively weaker and weaker, shortening the time period required for the return of the memories from the previous day; until eventually it is reduced to zero, so that the drug does nothing at all. It is hard to see how any of these steps other than the last one could be decisive in abruptly changing the final distribution of credence from 1/3–2/3 to 1/2–1/2. But it would be crazy to maintain that the difference between an awakening in which one's memories from the previous day are delayed by, say, a tenth of a second, and one in which they are present from the beginning, can have such a pronounced effect on what it is rational for one to believe once one has fully woken up. Moreover, any attempt to draw a significant distinction between this case and the original Sleeping Beauty case seems deeply unfaithful to the phenomenology of waking up. Waking up, even in the absence of amnesia drugs, very often is quite a gradual process; it can take time for one to summon up all the memories relevant to one's current situation (to convince yourself of this, see Proust 1913: 1–9). A credible theory of rationality in belief should not make the facts about the credences one ought to have at the end of this process depend on the precise facts about how the process works.

Lewis (2001) argues for (1), on the grounds that on Sunday night Beauty's credence in HEADS was 1/2, and she gains no new evidence relevant to HEADS between Sunday night and Monday morning: in other words, the difference between these two states of evidence is irrelevant to the question to what degree one ought to believe HEADS. By considering the variant case, we can see what is wrong with Lewis's premiss. In the variant case, Beauty's continued failure to remember anything from Monday after the first minute *is* evidence relevant to HEADS: it decreases her credence in HEADS from 1/2 to 1/3. In Lewis's terms, it is 'evidence about the future' – namely, that if the outcome of the toss is Heads, she is not now in it. If something as unspectacular as a lack of memories of a certain sort can constitute evidence relevant to HEADS in the variant case, presumably it can also do so in the original case. The only difference is that

in the original case, there is no intervening period when she is unsure about her temporal location while still having degree of belief 1/2 in HEADS. But why should that matter?[1]

New York University
New York, NY 10003-6688, USA
cian.dorr@nyu.edu

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. Repr. in D. Lewis, *Philosophical Papers*, vol. 2, Oxford: Oxford University Press, 1986.

Lewis, D. 2001. Sleeping Beauty: reply to Elga. *Analysis* 61: 171–76.

Proust, M. 1913. *Swann's Way*, vol. 1 of *In Search of Lost Time*. Tr. C. K. S. Moncrieff and T. Kilmartin. Rev. ed. by D. J. Enright. New York: The Modern Library, 1992.

[1] Thanks to Adam Elga for several helpful discussions. My greatest debt of gratitude is to David Lewis.