The Sleeping Beauty Problem

- The problem (Elga 2000):
  "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 back to sleep with a drug that makes you forget that waking."

- The researchers put you to sleep, and some time later you find yourself awake. Because of the possibility of memory erasure, you don't know whether it's your first awakening or second awakening. At this point, should your degree of belief that the coin came up heads be
  – 1/2?
  – less than 1/2?
  – greater than 1/2?

Two Plausible Principles

- The **Principal Principle** implies:
  – If you are certain that a particular outcome of an indeterministic process has a particular objective chance of occurring *and you have no inadmissible evidence*, your degree of belief that that outcome will occur should equal the objective chance.

- The **Relative Frequency Principle** (?):
  – If you are certain that in the long run many repetitions of an indeterministic process will converge on a particular relative frequency for a particular outcome, your degree of belief that that outcome will occur on one repetition of the process should equal the relative frequency.

Two Answers to the SBP

- "*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 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."

Updating by Conditionalization

- Suppose that between t_1 and t_2 you learn e.
- **Conditionalization**: For any x, $P_2(x) = P_1(x \mid e)$.
- Die-rolling story:
  – $P_1(\text{three}) = 1/6$
  – Between t_1 and t_2, you learn that the die came up odd.
  – $P_2(\text{three}) = P_1(\text{three} \mid \text{odd})$
  $$= P_1(\text{three \& odd})/P_1(\text{odd}) = 1/3$$

t_2 distribution
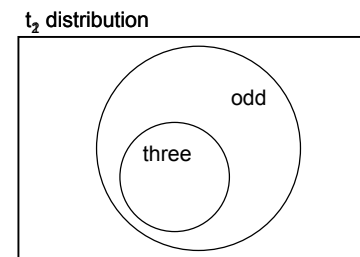
Elga's Argument

- Suppose the experiment begins Sunday night; the two (possible) awakenings are Monday morning and Tuesday morning.
- It doesn't make any difference whether the experimenters flip the coin after they put you to sleep on Sunday or after they put you to sleep on Monday. So let's suppose they flip it on Monday.
- Now suppose that you're awake all day Monday, and Monday night they tell you that it's Monday.
- "Your credence that [the coin will land Heads] would then be your credence that a fair coin, soon to be tossed, will land Heads."
  - By the Principal Principle, you should assign $P_2(h) = 1/2$.

Elga's Argument (Part II)

- $P_2(h) = 1/2$ [PP]
- $P_2(h) = P_1(h|mon)$ [Conditionalization]
- $P_1(h|mon) + P_1(t|mon) = 1$
- $P_1(h|mon) = P_1(t|mon) = 1/2$
- $P_1(h\&mon) = P_1(t\&mon)$
- $P_1(mon|t) = P_1(tues|t)$
        ["highly restricted principle of indifference"]
- $P_1(t\&mon) = P_1(t\&tues)$
- $P_1(h\&mon) + P_1(t\&mon) + P_1(t\&tues) = 1$
- $P_1(h\&mon) = 1/3$ [Conclusion]

- What's really going on (Law of Total Probability):
  $P_1(h) = P_1(h|mon) \cdot P_1(mon) + P_1(h|tues) \cdot P_1(tues)$
             <1   0

Lewis's Argument

- $P_0(h) = 1/2$ [PP]

- "Beauty gains no new uncentred evidence, relevant to Heads versus Tails, between the time when she has credence function P_0 and the time when she has credence function P_1. The only evidence she gains is the centred evidence that she is presently undergoing either the Monday awakening or the Tuesday awakening." (Lewis 2001)

- So $P_1(h) = 1/2$.

Lewis's Math

- Elga:

  $P_1(mon|t) = P_1(tues|t)$ [PI]
  $P_2(h) = P_1(h|mon)$ [Cond]

  $P_2(h) = 1/2$ [PP]

  …
  …
        $P_1(h) < P_2(h)$
  …
  …

  $\therefore P_1(h) = 1/3$

- Lewis:

  $P_1(mon|t) = P_1(tues|t)$ [PI]
  $P_2(h) = P_1(h|mon)$ [Cond]

  $P_1(h) = 1/2$ [last slide]

  …
  …
        $P_1(h) < P_2(h)$
  …
  …

  $\therefore P_2(h) = 2/3$

$P_2(h) = 2/3?!?$

- The Principal Principle "requires a proviso, which was satisfied when we used it to give us $P_1(h)=1/2$, but which is not satisfied when Elga uses it to give him $P_2(h)=1/2$. 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.

  (cont'd)

$P_2(h) = 2/3?!?$ (Part II)

- "Now when Beauty is told during her Monday awakening that it's not Monday… 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…. Elga agrees…. Therefore the proviso applies, and we cannot rely on it that $P_2(h)=Ch(h)$ and $P_2(t)=Ch(t)$. 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."

- Say *what?!?*

Why Isn't This Easy?

- Everyone agrees on exactly what Beauty's Sunday night (t_0) degrees of belief should be.

- Why can't we look at the evidence she gains between t_0 and t_1, then update by conditionalizing to determine her $P_1(h)$ value?
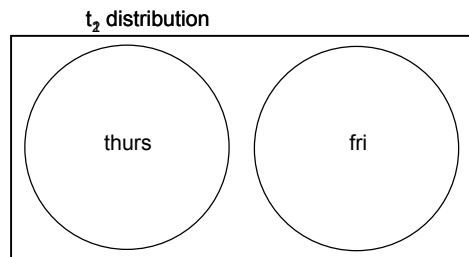
Centred vs. Uncentred Evidence
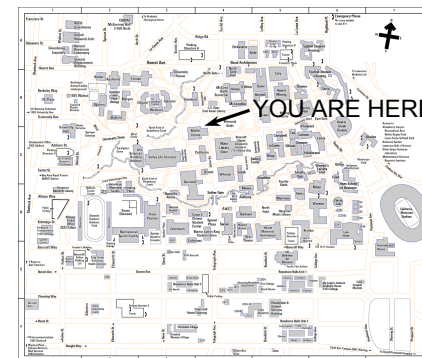
- "Uncentred" evidence says something about what the world is like. "Centred" evidence says something about where you're located in it. (Sometimes called "self-locating" evidence.)

Lewis's Argument (again)

- "Beauty gains no new uncentred evidence, relevant to Heads versus Tails, between the time when she has credence function P_0 and the time when she has credence function P_1. The only evidence she gains is the centred evidence that she is presently undergoing either the Monday awakening or the Tuesday awakening."

- **Relevance-Limiting Thesis**:
  It is never rational for an agent who gains only centred evidence to respond by altering an uncentred degree of belief.

- Isn't this something we can test using conditionalization?

Conditionalization Retains Certainties

- With a bit of math, we can show that if I assign $P_1(x) = 1$ and update by conditionalizing, I will assign $P_2(x) = 1$.
- This is because conditionalization strictly eliminates options.

t_2 distribution

- Now suppose I assign P_1(Today is Thursday) = 1. Must I assign P_2(Today is Thursday) = 1?
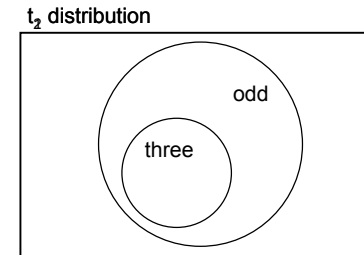
Conditionalization and Centred Evidence

- In general, conditionalization can't be used to update centred beliefs. This is because centred evidence *shifts* possibilities instead of *eliminating* them.

t_2 distribution

- In the Sleeping Beauty Problem, Beauty assigns P_0(sun) = 1. She can't update using conditionalization when she comes to believe "mon $\vee$ tues". But Lewis's entire argument is about the proper reaction to that piece of evidence. So we can't use conditionalization to evaluate Lewis's argument!

Solving the Problem

- To solve the Sleeping Beauty Problem, we need to know how agents should update their degrees of belief in response to centred (self-locating) evidence.
- This will also allow us to evaluate the Relevance-Limiting Thesis.

- So how should they?
  - One suggestion can be found in
    "The Relevance of Self-Locating Beliefs" by M. Titelbaum
    (at http://webfiles.berkeley.edu/titelbaum)

- By the way, the answer to the Sleeping Beauty Problem is 1/3.