Sleeping Beauty in Flatland

preprint
August 2003

Paul Franceschi
University of Corsica

p.franceschi@univ-corse.fr
http://www.univ-corse.fr/~franceschi

ABSTRACT. I present a solution to the Sleeping Beauty problem. I begin with the consensual *emerald case* and describe then a set of relevant urn analogies and situations. These latter experiments make it easier to diagnose the flaw in the thirder's line of reasoning. I discuss in detail the root cause of the flaw in the argument for 1/3 which is an erroneous assimilation with a repeated experiment. Lastly, I discuss an informative variant of the original Sleeping Beauty experiment that casts light on the diagnosis of the fallacy in the argument for 1/3.

1. Experiments and situations

Experiment 1: an urn contains 2 red balls and 1 green balls. You draw a ball at random from the urn. You evaluate the probability of drawing a red or a green ball. Let P(R) and P(G) denote respectively the probability of drawing a red or a green ball. Reasoning I: P(R) = 2/(2+1) = 2/3 and P(G) = 1/(2+1) = 1/3.

Situation 1: the *emerald case* (Leslie 1996, p. 20): 'At some point in time, three humans would each be given an emerald. Several centuries afterwards, when a completely different set of humans was alive, five thousands humans would again each be given an emerald in the experiment. You have no knowledge, however, of whether your century is the earlier century in which just three people were to be in this situation, or the later century in which five thousand were to be in it'. Let P(T) be the probability that your emerald comes from the set of three humans and P(F) the probability that your emerald originates from the set of five thousand humans. Reasoning I: P(T) = 3/(3+5000) = 3/5003 and P(F) = 5000/(3+5000) = 5000/5003.

Experiment 2: The content of an urn depends on the flipping of a fair coin. If Heads then the urn contains 1 red ball; if Tails then the urn contains 1 red ball and 1 green ball. You evaluate the probability of drawing a red or a green ball. Reasoning I: if the coin has landed Heads then the probability of drawing a red ball is 1; else if the coin has landed Tails then the probability of drawing a red ball is 1/2. In this latter case, we face a situation which is in all respects analogous to experiment 1 with an urn that contains 1 red ball and 1 green ball, except that the probability of Tails is 1/2, thus yielding a probability of drawing a red ball that equals 1/2x1/(1+1) = 1/2x1/2. On the other hand, if the coin has landed Heads then the probability of drawing a green ball is 0; else if the coin has landed Tails then the probability of drawing a green ball is 1/2. This latter case is analogous to experiment 1 with an urn that contains 1 red ball and 1 green ball, except that the probability of Tails is 1/2, thus yielding a probability of drawing a green ball that equals 1/2x1/(1+1) = 1/2x1/2. Hence P(R) = 1x1/2 + 1/2x1/2 = 3/4 ; P(G) = 0x1/2 + 1/2x1/2 = 1/4. Reasoning II: if the experiment is repeated *n* times, say 1000 times then there will be in total 1000 (1x1000x1/2+1x1000x1/2) red balls and 500 (1x1000x1/2) green balls. According to reasoning II this experiment is equivalent, in the long run, to a type 1 experiment with an urn that contains 1500 balls from whose 1000 red balls and 500 green balls. Hence P(R) = 1000/1500 = 2/3; and P(G) = 500/1500 = 1/3.

Experiment 3: experiment 2 repeated 1000 times. A type 2 experiment is repeated *n* times, from T_1 to T_n. Let *n* = 1000. Reasoning I: on each drawing, the odds are the same as in experiment 2. Hence P(R) = 1x1/2 + 1/2x1/2 = 3/4 ; P(G) = 0x1/2 + 1/2x1/2 = 1/4. Reasoning II: in this case, there will be in total circa 1000 (2x1000x1/2) red balls and 500 (1x1000x1/2) green balls. Thus one finds oneself in a situation which is equivalent to a type 1 experiment. Consequently: P(R) = 1000/1500 = 2/3; and P(G) = 500/1500 = 1/3.

Experiment 4: an urn contains 1 red ball and 1 green ball. If Heads then due to a filtering effect, you cannot see nor feel green balls and you can only see and feel 1 red ball. If Tails then there is no filter effect and you can see and feel 1 red ball and 1 green ball. Your task is to evaluate the probability of drawing a red or a green ball. Reasoning I: just as in experiment 2, if the coin has landed Heads then the probability of drawing a red ball is 1; else if the coin has landed Tails then the probability of drawing a red ball is 1/2. In addition, if the coin has landed Heads then the probability of drawing a green ball is 0; else if the coin has landed Tails then the probability of drawing a green ball is 1/2. Hence P(R) = 1x1/2 + 1/2x1/2 = 3/4 ; P(G) = 0x1/2 + 1/2x1/2 = 1/4. Reasoning II: if the experiment is repeated *n* times, say 1000 times then I will see in total 1000 (1x1000x1/2+1x1000x1/2) red balls and 500 (1x1000x1/2) green balls. According to reasoning II, in the long run, this experiment is equivalent to a type 1 experiment, with an urn that contains 1500 balls from whose 1000 red balls and 500 green balls. Hence P(R) = 1000/1500 = 2/3; and P(G) = 500/1500 = 1/3.

Experiment 5: experiment 4 repeated 1000 times. A type 4 experiment is repeated *n* times, from T_1 to T_n. Let *n* = 1000. In this case, I will draw in total circa 1000 (2x1000x1/2) red balls and 500 (1x1000x1/2) green balls. Reasoning I: on each drawing, the odds are the same as in experiment 4. Hence P(R) = 1x1/2 + 1/2x1/2 = 3/4 ; P(G) = 0x1/2 + 1/2x1/2 = 1/4. Reasoning II: one finds oneself in a situation which is equivalent to a type 1 experiment. Consequently: P(R) = 1000/1500 = 2/3; and P(G) = 500/1500 = 1/3.

Situation 4a: *Sleeping Beauty problem* (Elga 2000, p. 143): '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'. Once awakened, to what degree should Sleeping Beauty believe that (i) it is a Heads-waking and (ii) the coin has landed Heads?[1] '*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 (...). 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 (...). So on any particular waking, you should have credence 1/3 that that waking is Heads-waking, and hence have credence 1/3 in the coin's landing Heads on that trial'.

Situation 4b: *Sleeping Beauty in Flatland*: Some researchers give you first a cube and then put you to sleep. During your sleep they place you with the cube, depending on the toss of a fair coin, either in Flatland (Heads) or in Spaceland (Tails). After that, they wake you up once. Once awakened in Flatland (Heads), you will see a 2-dimensional object, a square. Conversely, if awakened in Spaceland, you will see a 3-dimensional object, a cube. You evaluate (i) the probability of being awakened in Flatland or in Spaceland and (ii) the probability that the coin has landed Heads.

2. The reasoning

[1] Adapted from Elga (2000). Elga's original text: 'When you are *first* [my emphasis] awakened, to what degree ought you believe that the outcome of the coin toss is Heads?'. Considering here *any* waking (Heads-waking on Monday, Tails-waking on Monday or Tails-waking on Tuesday) is more general and equally allowed by the formulation of the problem, since all wakings are indistinguishable.

From the foregoing experiments and situations, the following reasoning can be straightforwardly derived (assuming that steps (6) and (11) are true):

(1)	situation 1 (*emerald case*) is analogous to experiment 1	analogy
(2)	reasoning I applies to experiment 1	premise
(3)	∴ reasoning I applies to situation 1 (*emerald case*)	from (1),(2)
(4)	either reasoning I or reasoning II applies to experiment 2	dichotomy
(5)	according to reasoning II from experiment 2, experiment 3 (iterated experiment 2) is structurally identical to experiment 1	premise
(6)	experiment 3 (iterated experiment 2) is not structurally identical to experiment 1	premise
(7)	∴ reasoning II does not apply to experiment 2	from (5),(6)
(8)	∴ reasoning I applies to experiment 2	from (4),(7)
(9)	either reasoning I or reasoning II applies to experiment 4	dichotomy
(10)	according to reasoning II in experiment 4, experiment 5 (iterated experiment 4) is structurally identical to experiment 1	premise
(11)	experiment 5 (iterated experiment 4) is not structurally identical to experiment 1	premise
(12)	∴ reasoning II does not apply to experiment 4	from (10),(11)
(13)	∴ reasoning I applies to experiment 4	from (9),(12)
(14)	situation 4a (*Sleeping Beauty problem*) is analogous to experiment 4	analogy
(15)	reasoning I applies to experiment 4	from (13)
(16)	∴ reasoning I applies to situation 4a (*Sleeping Beauty problem*)	from (14),(15)
(17)	situation 4b (*Sleeping Beauty in Flatland*) is analogous to experiment 4	analogy
(18)	reasoning I applies to experiment 4	from (13)
(19)	∴ reasoning I applies to situation 4b (*Sleeping Beauty in Flatland*)	from (17),(18)

Assuming the hypothesis that steps (6) and (11) are true, I take it that the whole reasoning (1)-(19) should be consensual. However, the current state of the philosophical debate is that step (6) should be regarded as more or less consensual while, on the other hand, step (11) remains at this stage fully controversial. For this reason, I shall now concentrate on the rationale that makes these two specific steps true.

3. A solution to the Sleeping Beauty problem

Let us begin with step (6). Step (6) states thus against reasoning II that experiment 3 (iterated experiment 2) is not structurally identical to experiment 1. Reasoning II rests on the fact that experiment 2 can be repeated and the corresponding situation is then analogous to a type 1 experiment with 1500 balls from whose 1000 red and 500 green balls.

I shall argue that step (6) is true and that reasoning II in experiment 2 is fallacious. For the sake of clarity, let us draw first a distinction between red-HEADS (red balls created after the coin has landed Heads), red-TAILS (red balls created in the Tails case) and green-TAILS (green balls created in the Tails case) balls. In this context, it should be clear that there only exists red-HEADS, red-TAILS and green-TAILS balls in experiments 2 and 3.

The intuition underlying reasoning II in experiments 2 and 3 is that one is entitled to add red and green balls to compute frequencies. However, I shall argue that this intuition is misleading. With our terminology, it means that one feels intuitively entitled to add red-HEADS, red-TAILS and green-TAILS balls. Let us begin with red-HEADS balls. In the current context, red-HEADS balls can be considered properly as single objects. Thus you are entitled to envisage drawing isolately red-HEADS balls and these latter can acceptably be seen as single objects. By contrast, it appears that red-TAILS balls are quite undissociable from green-TAILS balls. For you cannot draw a red-TAILS ball without drawing the associated green-TAILS ball. And conversely, you cannot pick a green-TAILS ball without picking the associated red-TAILS ball. From this viewpoint, it is mistaken to consider red-

TAILS and green-TAILS balls as separate objects. The correct intuition is that the association of a red-TAILS and a green-TAILS ball constitute one single object, in the same sense as red-HEADS balls constitute single objects. And red-TAILS and green-TAILS balls are best seen intuitively as parts of one single object, whose constituents are one red-TAILS and one green-TAILS ball. In other words, red-HEADS balls and, on the other hand, red-TAILS and green-TAILS balls, cannot be considered as objects of the same type for frequency probability purposes. And this situation motivates the fact that one is not entitled to add red-HEADS, red-TAILS and green-TAILS balls to compute frequencies. If experiment 2 is repeated, one is not entitled to add (i) red-HEADS and red-TAILS balls and (ii) red-HEADS and green-TAILS balls to compute the corresponding frequencies. For in both cases, you add objects of intrinsically different types, i.e. you add one single object with the mere part of another single object. It follows that reasoning II in experiment 2 is erroneous. Hence, reasoning I is correct. As we have seen, the whole idea of reasoning as if experiment 2 were repeated is related to the frequentist interpretation of probabilities (Hájek 2002) and the repeatability of thought experiments. The upshot, however, is that this latter interpretation of probabilities should not be adopted unrestricted.

Let us consider, second, step (11). Step (11) claims contra reasoning II that experiment 5 (iterated experiment 4) is not structurally identical to experiment 1. In this context, reasoning II is based on the fact that experiment 4 can be repeated and the corresponding situation is then analogous to a type 1 experiment with 1500 balls from whose 1000 red and 500 green balls. The targeted situation is the Sleeping Beauty Problem.

On closer scrutiny, it appears that experiment 4 is the same as experiment 2, except that a selection effect (Leslie 1989, Bostrom 2002) is present in the former case. In effect, if the coin has landed Heads then a selection effect precludes you from feeling and seeing the green ball. In this context, from the observer's standpoint, the situation is identical to experiment 2. However, if the coin has landed Tails, there is no selection effect and you can feel and see properly both red and green balls.

Now the diagnosis of the fallacy in reasoning II is the same as in experiment 2. What is at the origin of the problem is that each red ball is intuitively considered as a single object. But this intuition proves to be mistaken. And what creates the illusion is that one seems pre-theoretically entitled to add red balls and green ones to compute frequencies. But what the above analysis reveals, is that one must distinguish between red-HEADS, red-TAILS and green-TAILS balls. Once this step accomplished, it is patent that the correct intuition is that red-HEADS can be seen as single objects, while red-TAILS and green-TAILS balls must be considered properly as mere parts of single objects which are on a par with red-HEADS objects. In sum, red-HEADS balls and, on the other hand, red-TAILS and green-TAILS balls, cannot be considered as objects as the same type. The upshot is that one is no longer entitled to add (i) red-HEADS and red-TAILS balls and (ii) red-HEADS and green-TAILS balls, to compute frequencies. Consequently, reasoning II appears now fallacious, making it clear that experiment 5 is not analogous to experiment 1. What remains thus in force is reasoning I.

The situation of the original Sleeping Beauty problem parallels the urn analogy from experiment 4. In the Sleeping Beauty problem, it appears that Beauty faces a selection effect in the case where the coin lands Heads. For in this last case, Beauty is not awakened on Tuesday. By contrast, Beauty faces no selection effect in the Tails case, since she is awakened on both Monday and Tuesday. In the Sleeping Beauty experiment, the time variable includes two temporal locations: Monday and Tuesday. In the Heads case, Sleeping Beauty perceives the first time location (Monday) but is unable to perceive the second location (Tuesday). However, in the Tails case, she is able to perceive both time locations (Monday and Tuesday).

Let now P(M) be the probability of being awakened on Monday and P(T) be the probability of being awakened on Tuesday. Given the above set of experiments and situations, we are now in a position to state the argument for 1/3 in the Sleeping Beauty problem more accurately:

(20) if the Sleeping Beauty experiment is repeated *n* times hypothesis
(21) then there will be 1/3 Heads-wakings on Monday, 1/3 Tails-wakings on Monday and 1/3 Tails-wakings on Tuesday
(22) then the experiment will be structurally identical to experiment 1

(23) in experiment 1, P(R) = 2/3 and P(G) = 1/3

(24) in the Sleeping Beauty experiment, P(M) = 2/3 and P(T) = 1/3 from (22),(23)

(25) ∴ P(TAILS) = P(M) = 2/3 and P(HEADS) = P(T) = 1/3 from (24)

From the above, it should be clear that the argument for 1/3 in the Sleeping Beauty problem identifies itself with the above-mentioned reasoning II in experiments 2 and 4. The argument for 1/3 rests crucially on the fact that if experiment 4 is repeated, it can be assimilated to a type 1 experiment.

Now the erroneous step in the thirder's line of reasoning can be accurately identified. The erroneous step is the inference from (21) to (22), namely the consideration that if the experiment is repeated *n* times, it is equivalent to a type 1 experiment. And the diagnosis of the fallacy in the argument for 1/3 now parallels the flaw in reasoning II in experiments 2 and 4. What creates the problem is the misleading intuition that each waking is intuitively considered as a single event. And the apparent plausibility of the argument for 1/3 comes from the fact that one feels pre-theoretically entitled to add Monday wakings and Tuesday wakings to compute frequencies. However, as underlined above, one must distinguish first between Monday-HEADS, Monday-TAILS and Tuesday-TAILS wakings. It follows then that Monday-HEADS wakings and, on the other hand, Monday-TAILS and Tuesday-TAILS wakings cannot be properly considered as objects as the same type. For Monday-TAILS wakings are undissociable from Tuesday-TAILS wakings. And this finally prohibits adding (i) Monday-HEADS and Monday-TAILS wakings and (ii) Monday-HEADS and Tuesday-TAILS wakings to compute frequencies. This renders reasoning II fallacious and finally does justice to reasoning I.

4. Lessons of Sleeping Beauty in Flatland

Lastly, the Sleeping Beauty in Flatland variant of the original problem is worth taking into account. *Flatland* is a small book written by Edwin E. Abbott and published in 1884, which has undergone an increasing popularity, until our present day. Although it is also a social satire on the rigidities of Victorian England, the main concern of *Flatland* is with introducing the geometry of higher dimensions. The protagonist of the book, A Square, is an inhabitant of a 2-dimensional world, which only contains flat individuals. Abbott investigates what it would mean for such inhabitants to interact with beings and objects from of a 3-dimensional world. The underlying analogy, which also applies to our current situation, is that the inhabitants of a *n*-dimensional world would face a situation of the same nature when interacting with objets of a *n*+1 dimensional world. In this context, *Flatland* can also be regarded as an introductory text to 4-dimensional objects and higher-dimensional polytopes.

In the Sleeping Beauty in Flatland variant, it appears that when Beauty is thrown in Flatland in the Heads case, she faces an observation selection effect that precludes her from perceiving the 3rd spatial dimension of the cube and causes her seeing a mere square. A 3-dimensional object which is transferred in a 2-dimensional world such as Flatland is seen there as a 2-dimensional object. Its spatial 3rd dimension is hidden to all observers. By contrast, when the same object is plugged into a 3-dimensional world, it appears in its entirety as a 3-dimensional object. And when Beauty is plugged accordingly into the 3-dimensional world of Spaceland in the Tails case, she is enabled to perceive the cube with all its three dimensions.

There is something puzzling with the Sleeping Beauty in Flatland variant. Our prima facie reasoning is that if the experiment is repeated, say 1000 times, Beauty will see circa 500 squares and 500 cubes, a line of reasoning quite in line with reasoning I and the conclusion that P(HEADS) = P(TAILS) = 1/2. By contrast, our pre-theoretical intuition is that there is something which impedes to get reasoning II moving in the Sleeping Beauty in Flatland variant. Hence, reasoning II seems intuitively less natural in the Sleeping Beauty in Flatland variant than in the original Sleeping Beauty problem. This intriguing phenomenon stands in need of an explanation.

Let us begin with showing how the Sleeping Beauty in Flatland variant is structurally analogous to experiment 4. Let us first term a quasi-cube an object that is a cube one face of which is missing. Now it is patent that when Sleeping Beauty sees a cube in Spaceland, she also sees an object whose constituents are a square and a quasi-cube. Let us forget about cubes for a while and concentrate on squares and quasi-cubes. Let us also draw a distinction between square-HEADS, square-TAILS and

quasi-cube-TAILS. With the relevant machinery at hand, we are now in a position to reframe reasoning II in the Sleeping Beauty in Flatland variation. For we can now imagine the experiment repeated many times. It appears then that in the long run, Beauty will see 500 square-HEADS, 500 square-TAILS and 500 quasi-cube-TAILS, giving now apparently strong grounds for the conclusion, in line with reasoning II, that P(HEADS) = 1/3. At this stage, the same diagnosis as above of the flaw in reasoning II in experiment 4 applies straightforwardly. Now it appears that square-HEADS and, on the other hand, square-TAILS and quasi-cube-TAILS, cannot be considered as objects of the same type. The upshot is that one is no longer entitled to add (i) square-HEADS and square-TAILS and (ii) square-HEADS and quasi-cube-TAILS to compute frequencies. This undercuts reasoning II and finally vindicates reasoning I.

Now it should be clear that the Sleeping Beauty in Flatland variant is the same as in the original Sleeping Beauty problem, to the difference that the variable is spatial in the Sleeping Beauty in Flatland variant while it is temporal in the original Sleeping Beauty experiment. In this context, the Sleeping Beauty in Flatland variant appears now informative. In effect, in this latter case, one does not feel pre-theoretically entitled to add square-HEADS, square-TAILS and quasi-cube-TAILS. Rather, one feels intuitively entitled to add squares and cubes, for the reason that we are only familiar with these latter objects. In particular, quasi-cubes are unfamiliar to us, being uncommon objects and concepts. Now adding squares and quasi-cubes is for us unnatural. This explains why, although structurally identical to the original Sleeping Beauty experiment, the Sleeping Beauty in Flatland variant is not equally suited for reasoning II.

Finally, the lesson of the Sleeping Beauty Problem is that our current and familiar objects or concepts such as balls, wakings, etc. should not be considered as the sole relevant classes of objects for probability purposes. For in certain situations, in order to reason properly, it is also necessary to take into account somewhat unfamiliar objects such as pairs of indissociable balls, pairs of mutually unseparable wakings, 3-dimensional complements of 2-dimensional objects such as quasi-cubes, etc. Once this goodmanian step accomplished, we should be less vulnerable to certain subtle cognitive traps in probabilistic reasoning.[2]

References

Abbott, E. A. (1884) *Flatland: A Romance of Many Dimensions*, http://www.geom.umn.edu/~banchoff/Flatland

Arntzenius, F. (2002) Reflections on Sleeping Beauty, *Analysis*, 62-1, 53-62

Bostrom, N. (2002) *Anthropic Bias: Observation Selection Effects in Science and Philosophy,* New York, Routledge

Bradley, D. (2003) Sleeping Beauty: a note on Dorr's argument for 1/3, *Analysis*, 63, 266-8

Delahaye, J.-P., (2003) La Belle au bois dormant, la fin du monde et les extraterrestres, *Pour la Science*, 309, 98-103

Dorr, C. (2002) Sleeping Beauty: in Defence of Elga, *Analysis*, 62, 292-6

Elga, A. (2000) Self-locating Belief and the Sleeping Beauty Problem, *Analysis*, 60, 143-147

Hájek, A. (2002) Interpretations of Probability, *The Stanford Encyclopedia of Philosophy* (Winter 2002 Edition), E. N. Zalta (ed.), http://plato.stanford.edu/archives/win2002/entries/probability-interpret

Leslie, J. (1989) *Universes*, London, Routledge

Leslie, J. (1996) *The End of the World: the science and ethics of human extinction*, London, Routledge

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

Monton, B. (2002) Sleeping Beauty and the Forgetful Bayesian, *Analysis*, 62, 47-53

[2] I thank Jean-Paul Delahaye and Claude Panaccio for useful discussion.