Sleeping Beauty in Flatland

preprint

Paul Franceschi
University of Corsica

revised January 2004

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

ABSTRACT. I describe in this paper a solution to the Sleeping Beauty problem. I begin with the consensual emerald case and I also discuss Bostrom's Incubator gedanken. I address then the Sleeping Beauty problem. I argue that the root cause of the flaw in the argument for 1/3 is an erroneous assimilation with a repeated experiment. Lastly, I present 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. The emerald case

Consider, to begin with, the following experiment:

Experiment 1: an urn contains p red balls and q green balls. You draw a ball at random from the urn. You try to 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. According to reasoning (I), you conclude then that $P(R) = p / (p + q)$ and $P(G) = q / (p + q)$.

Let us turn now to the situation corresponding to the *emerald case*, described by John Leslie (1996, p. 20):

Situation 1: the *emerald case*: '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'. What is then your credence that your emerald originates from the set of three humans?

Let us identify now the first set of three emeralds with red balls and the second set of five thousands emeralds with green balls. The situation is now equivalent to an urn that contains three red balls and five thousands green balls. At this stage, it should be clear that situation 1 is structurally analogous to experiment 1, with $p = 3$ and $q = 5000$. We get then accordingly: $P(R) = 3 / (3 + 5000) = 3/5003$ and $P(G) = 5000 / (3 + 5000) = 5000/5003$. The resulting probability that your emerald comes out from the set of three humans equals 3/5003; and the probability that it originates from the set of five thousand humans equals 5000/5003. The corresponding line of reasoning can be described more accurately as follows:

(1)	situation 1 (the *emerald case*) is analogous to experiment 1	analogy
(2)	reasoning (I) applies to experiment 1	premise
(3)	$\therefore$ reasoning (I) applies to situation 1 (the *emerald case*)	from (1),(2)

I take it that the above reasoning should be consensual. However at this step, agreement stops. Let us then proceed a bit further by reviewing some other experiments and situations.

2. The Incubator

Let us consider now the following experiment:

Experiment 2: The content of an urn depends on the flipping of a fair coin. If Heads, the urn contains one red ball; if Tails, it contains one red ball and one green ball. You try to evaluate the probability of drawing a red or a green ball. A fist line of reasoning (I) goes as follows. Consider, to begin with, the probability of drawing a red ball. If the coin has landed Heads then the probability of drawing a red ball is 1. Now if the coin has landed Tails then this latter probability equals 1/2. The probability of Heads and Tails being 1/2, we get accordingly: P(R) = 1 x 1/2 + 1/2 x 1/2 = 3/4. It is worth mentioning in passing that in the Tails case, the situation is in all respects analogous to experiment 1 with an urn that contains one red ball and one green ball, except that the probability of the corresponding situation is 1/2. Let us turn now to the probability of drawing a green ball. If the coin has landed Heads then the probability of drawing a green ball is 0. By contrast, if the coin has landed Tails then this latter probability is 1/2. Hence P(G) = 0 x 1/2 + 1/2 x 1/2 = 1/4. It is worth noting that the Tails case is analogous to experiment 1 with an urn that contains one red ball and one green ball, except that the probability of the corresponding situation is 1/2. To sum up, according to reasoning (I): P(R) = 3/4 and P(G) = 1/4.

However, an alternative reasoning (II) goes as follows. Let us term *iterated* e*xperiment 2*, experiment 2 repeated *n* times. If experiment 2 is repeated *n* times, say 1000 times, then there will be in total 1000 (1 x 1000 x 1/2 + 1 x 1000 x 1/2) red balls and 500 (0 x 1000 x 1/2 + 1 x 1000 x 1/2) 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.

I shall argue that reasoning (II) in experiment 2 is fallacious. 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. In what follows, my concern will be with showing that *iterated* e*xperiment 2* is not structurally analogous to experiment 1.

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 after the coin has landed Tails) balls. In this context, it should be clear that there only exists red-Heads, red-Tails and green-Tails balls in experiment 2.

The intuition underlying reasoning (II) in experiment 2 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 to compute frequencies. Let us consider first red-Heads balls. In the current context, red-Heads balls can be considered properly as single objects. Thus you are entitled to envisage drawing isolatedly red-Heads balls and these latter can acceptably be seen as single objects. By contrast, it appears that red-Tails balls are quite undissociatable 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 constituents and mere parts of one single object. In other words, red-Heads balls and, on the other hand, red-Tails and green-Tails balls, cannot be put on a par and considered as objects of the same type for frequency probability purposes. And this justifies the fact that one is not entitled to add red-Heads, red-Tails and green-Tails balls to compute probability 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. The correct intuition, however, is that red-Heads can be seen as single objects, while red-Tails and green-Tails balls ought to

be considered properly as mere parts of single objects which are on a par with red-Heads objects. Now the analogy with experiment 1 proves to be ungrounded, since in this latter case, red and green balls can legitimately be put on a par and considered as objects of the same type. This invalidates the analogy of iterated experiment 2 with experiment 1. It follows that reasoning (II) in experiment 2 is incorrect. This leaves us with reasoning (I). 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). The upshot, however, is that this latter interpretation of probabilities should not be adopted unrestricted and in particular, frequencies should not be computed by adding objects of intrinsically different types.

At this step, it is worth considering the situation corresponding to the *Incubator* (Bostrom 2002, p. 64):[1]

Situation 2: the *Incubator*, version 1: 'Stage (a) In an otherwise empty world, a machine called "the incubator" kicks into action. It starts by tossing a fair coin. If the coin falls [heads] then it creates one room and a man with a black beard inside it. If the coin falls [tails] then it creates two rooms, one with a black-bearded man and one with a white-bearded man. As the rooms are completely dark, nobody knows his beard color. Everybody who's been created is informed about all of the above. You find yourself inside one of the rooms.' Question: What should be your credence that you have a black or a white beard? [2]

Now the line of reasoning related to experiment 2 can be transposed straightforwardly to the *Incubator*. For let us identify a black-bearded man with a red ball and a white-bearded man with a green ball. The resulting situation is a machine that creates one red ball on Heads and one red ball and one green ball on Tails. This has the effect or rendering the situation corresponding to the *Incubator* analogous to experiment 2. Now given that reasoning (I) applies to experiment 2, it follows that reasoning (I) also applies to the *Incubator*. The corresponding reasoning by analogy can be stated thus as follows:

(4)	situation 2 (the *Incubator*) is analogous to experiment 2	analogy
(5)	reasoning (I) applies to experiment 2	premise
(6)	∴ reasoning (I) applies to situation 2 (the *Incubator*)	from (4),(5)

3. The Sleeping Beauty problem: a solution

Let us envisage now the following experiment:

Experiment 3: an urn contains one red ball and one green ball. If Heads then due to a filtering effect, you cannot see nor feel green balls and you can only see and feel one red ball. If Tails then there is no filter effect and you can see and feel one red ball and one green ball. Your task is to evaluate the probability of drawing a red or a green ball.

At this stage, it appears that experiment 3 is in all respects similar to experiment 2, except for what concerns the Heads case. In this latter case, in experiment 2 there is only one red ball in the urn, which is devoid of green balls. By contrast, in experiment 3, there is one red ball and one green ball in the urn, but you cannot see nor feel the green ball, due a to selection effect (Leslie 1989, Bostrom 2002). But the outcome is that this precludes your from drawing the green ball in the Heads case, thus rendering the situation equivalent -- from a probability standpoint -- to experiment 2. As a consequence, reasoning (I) also applies to experiment 3.

[1] The Heads and Tails cases are reverted here, with regard to Bostom's original description. The extensive version of the incubator also includes a later stage: "Stage (b): a little later, the lights are switched on, and you discover that you have a black beard. Question: What should your credence in Tails be now?".
[2] Bostrom's original question: " What should be your credence that the coin fell tails?".

At this step, it is worth considering the *Sleeping Beauty problem* (Elga 2000, p. 143):

Situation 3a: the *Sleeping Beauty problem*: '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'. Once awakened, what should Sleeping Beauty's credence be that (i) it is a Monday waking; and (ii) the coin has landed Heads?[3]

'*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'.

I shall argue here that the situation corresponding to the *Sleeping Beauty problem* is structurally analogous to experiment 3. In the Sleeping Beauty experiment, the time variable includes two temporal locations: Monday and Tuesday. Moreover, it appears that Beauty faces a selection effect in the case where the coin lands Heads, for in this latter 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 short, in the Heads case, Sleeping Beauty perceives the first time location (Monday) but is unable to perceive the second temporal location (Tuesday). However, in the Tails case, she is able to perceive both time locations (Monday and Tuesday). Let us identify now Monday with a red ball and Tuesday with a green ball. Now a Monday-waking can be assimilated with drawing a red ball and a Tuesday-waking can be identified with drawing a green ball. Furthermore, being not awaken on Tuesday in the Heads case can be assimilated with being incapable of seeing nor feeling the green ball due to a filtering effect. At this stage, it should be clear that the situation corresponding to the Sleeping Beauty problem is analogous to experiment 3. It follows that reasoning (I) also applies to the situation corresponding to the Sleeping Beauty problem. This can be stated step-by-step as follows:

(7)	situation 3a (*Sleeping Beauty problem*) is analogous to experiment 3	analogy
(8)	reasoning (I) applies to experiment 3	premise
(9)	$\therefore$ reasoning (I) applies to situation 3a (*Sleeping Beauty problem*)	from (7),(8)

At this stage, we are in a position to diagnose precisely the flaw in the argument for 1/3 in the Sleeping Beauty problem. For this purpose, it is worth stating more accurately the argument for 1/3. It begins informally with the transposition of reasoning (II) in experiment 3. The argument for 1/3 rests crucially on the fact that if the Sleeping Beauty experiment is repeated, it can be assimilated to a type 1 experiment. The corresponding line of reasoning runs as follows: if the Sleeping Beauty experiment is repeated *n* times, say 1000 times, then there will be in total 1000 (1 x 1000 x 1/2 + 1 x 1000 x 1/2) Monday-wakings and 500 (0 x 1000 x 1/2 + 1 x 1000 x 1/2) Tuesday-wakings. This experiment is equivalent in the long run, the argument goes, to a type 1 experiment with an urn that contains 1500 balls from whose 1000 red and 500 green balls. The respective probabilities of a Monday-waking and of a Tuesday-waking are computed accordingly: P(Monday-waking) = 1000/1500 = 2/3 and P(Tuesday-waking) = 500/1500 = 1/3.

At this point, it is worth mentioning some additional steps which are specific to the argument for 1/3 and which are targeted at computing P(Heads) and P(Tails). Given that the total number of Heads-wakings on Monday and of Tails-waking on Monday amounts respectively to 1 x 1000 x 1/2 = 500, it follows that the probability of a Heads-waking on Monday and of a Tails-waking on Monday equals 500/1500 = 1/3. Now given that the probability of Heads equals the probability of a Heads-waking on

[3] 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.

Monday, it follows that P(Heads) = P(Heads-waking on Monday) = 1/3. In parallel, the probability of Tails equals the probability of a Tails-waking on Monday plus the probability of a Tails-waking on Tuesday. Hence, P(Tails) = P(Tails-waking on Monday) + P(Tuesday-waking) = 1/3 + 1/3 = 2/3.

Now the flaw in the thirder's line of reasoning can be accurately diagnosed. The erroneous step is the *analogy stage*, namely the consideration that if the experiment is repeated *n* times, it will be 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 3. What is at the origin of the problem is the misleading intuition that each waking is intuitively considered as one single event. And the apparent plausibility of the argument for 1/3 emerges 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 draw first a distinction 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 considered properly 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. Sleeping Beauty in Flatland

At this step, one could wonder whether the above analysis would be confirmed. Such a confirmation ought to satisfy the following requirements. What is needed is a situation structurally analogous to experiment 3, but where the real objects which represent the balls cannot be intuitively added. This can be done by considering the *Sleeping Beauty in Flatland* variant.

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. Furthermore, *Flatland* can also be regarded as an introductory text to 4-dimensional objects and to higher-dimensional polytopes.

Let us describe then the situation corresponding to the Sleeping Beauty in Flatland experiment:

Situation 3b: *Sleeping Beauty in Flatland*: Some researchers give you a cube and then put you to sleep. During your sleep they will place you with the cube, depending on the toss of a fair coin, either in Flatland (Heads) or in Spaceland (Tails). After that, they will 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 then (i) the probability of seeing a square and (ii) the probability that the coin has landed Heads.

The Sleeping Beauty in Flatland variant has a remarkable feature. 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. But our pre-theoretical intuition also tells us that there is something which impedes reasoning (II) to get moving in the Sleeping Beauty in Flatland variation. Hence, reasoning (II) seems intuitively less appealing in the Sleeping Beauty in Flatland variant than in the original Sleeping Beauty experiment. This intriguing phenomenon stands in need of explanation. But let us pause for a while before suggesting the corresponding explanation.

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 third 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.

At this step, it is worth showing that the Sleeping Beauty in Flatland variant is structurally analogous to experiment 3. 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) into the Sleeping Beauty in Flatland variation. For reasoning (II) can be now revived as follows. Imagine the experiment repeated many times. In the long run, Beauty will see 500 square-Heads, 500 square-Tails and 500 quasi-cube-Tails, giving now apparently strong grounds to the conclusion, now in line with reasoning (II), that the probability of drawing a square equals 2/3 and that P(Heads) = 1/3.

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. At this stage, the same diagnosis as above of the flaw in reasoning (II) in experiment 3 applies straightforwardly. Now it can be pointed out 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). It follows then accordingly:

(10)	situation 3b (*Sleeping Beauty in Flatland*) is analogous to experiment 3	analogy
(11)	reasoning (I) applies to experiment 3	premise
(12)	∴ reasoning (I) applies to situation 3b (*Sleeping Beauty in Flatland*)	from (10),(11)

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 inclined 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. 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. We should bear in mind that according to an unformalised axiom of probability theory, a given situation is standardly modelled with the help of urns, dices, balls, etc. But the rules that allow for these simplifications lack an explicit formulation. However in certain situations, in order to reason properly, it is also necessary to take into account somewhat unfamiliar objects whose constituents are pairs of indissociable balls or of mutually unseparable wakings, or 3-dimensional complements of 2-dimensional objects such as quasi-cubes, etc. This lesson was anticipated by Nelson Goodman, who pointed out in *Ways of Worldmaking* that some objects which are prima facie completely different from our familiar objects also deserve some consideration: 'we do not welcome molecules or concreta as elements of our everyday world, or combine tomatoes and triangles and typewriters and tyrants and tornadoes into a single kind'.[4] As we have seen, we cannot add unrestrictedly objects of the Heads-world with objects of the Tails-world. For some objects of the Heads-world are mere parts of objects in the Tails-world and reciprocally, some mere constituents of objects of the Tails-world are genuine and independent objects in the Heads-world. Now the status of our paradigm probabilistic object, namely a ball, appears world-relative, since it can be a whole in the Heads-world and a part in the Tails-world. Once this goodmanian step accomplished, we should be less vulnerable to certain subtle cognitive traps in probabilistic reasoning.[5]

[4] Goodman (1978, p. 21).

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

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, 143-7

Goodman, N. (1978) *Ways of Worldmaking*. Indianapolis: Hackett Publishing Company

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