Sleeping Beauty goes to the lab: The psychology of self-locating evidence

Matteo Colombo

Jun Lai

Vincenzo Crupi

Abstract. Analyses of the Sleeping Beauty Problem are polarised between those advocating the "1/2 view" ("halfers") and those endorsing the "1/3 view" ("thirders"). The disagreement concerns the evidential relevance of self-locating information. Unlike halfers, thirders regard self-locating information as evidentially relevant in the Sleeping Beauty Problem. In the present study, we systematically manipulate the kind of information available in different formulations of the Sleeping Beauty Problem. Our findings indicate that patterns of judgment on different formulations of the Sleeping Beauty Problem do not fit either the "1/2 view" or the "1/3 view." Human reasoners tend to acknowledge self-locating evidence as relevant, but discount its weight significantly. Accordingly, self-locating information may trigger more cautious judgments of confirmation than familiar kinds of statistical evidence. We also discuss how these results can advance the debate by providing a more nuanced and empirically grounded account or explication of the evidential impact of self-locating information.

Keywords: sleeping beauty problem; probability; reasoning; self-locating evidence.

Sleeping Beauty goes to the lab: The psychology of self-locating evidence

1. Introduction

The Sleeping Beauty Problem (SBP) is a challenging puzzle in probabilistic reasoning. It raises questions of unsuspected theoretical relevance for the foundations of probabilistic reasoning, belief update, decision-making, and beyond (Titelbaum, 2013).
In its standard formulation, the problem goes as follows:

> On a Sunday, 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. When you are awakened, to what degree ought you believe that the outcome of the coin toss is Heads? (Elga, 2000; see Piccione & Rubinstein, 1997, Example 5, for an earlier formulation)

In the SBP everyone agrees that, on Sunday, your degree of belief in Heads should be 1/2. Opinions are split, however, about what should happen to your belief when you are awakened. For so-called *halfers*, your credence in Heads should remain 1/2, whereas so-called *thirders* think that your credence in Heads should change to 1/3 (and, accordingly, 2/3 for Tails). Let us briefly lay down halfers' and thirders' arguments.

At the outset, you know all the details of the SBP experiment, including that the coin is fair and that you will lose your memory of an earlier awakening. When you wake up, all new information you have is contained in the statement 'I am awake now.' Statements of this type carry *self-locating* (or *centred*) information, which concerns only one's spatio-temporal location in the world or one's identity — such as the information conveyed by the statements 'Today is Monday,' or 'I am Jun.' Instead, non-self-locating (*uncentred*) information — such as the information conveyed by 'The coin landed Tails' or 'Jun was born in Nanjing' — concerns what the world is like. For the halfer, the piece of self-locating information contained in the statement 'I am awake now' bears no relevant connection to the outcome of the coin flip. Therefore, your initial degrees of credence should remain unchanged. And because before the experiment you know that the coin is fair, you should then retain a credence of $P(\text{Heads}) = P(\text{Tails}) = 1/2$ (Lewis, 2001; Arntzenius, 2002; Cozic, 2011; Hawley, 2013).

So-called *thirders* disagree and submit that your credence in Heads when you are awakened should be 1/3. Their argument goes as follows. Given the set-up of SBP, the pair {It is now Monday; It is now Tuesday} partitions the space of the possible situations in which you may find yourself when you wake up. When you wake up, thus, the statements 'I am awake and it is now Monday' and 'I am awake and it is now Tuesday' pick out jointly exhaustive and mutually exclusive states of affairs. Accordingly, when you consider the outcome of the coin toss, you will have to assign certain degrees of belief to the following distinct propositions:

H: It is now Monday and the fair coin landed Heads.

T_1: It is now Monday and the fair coin landed Tails.

T_2: It is now Tuesday and the fair coin landed Tails.

The thirder's solution of the SBP assumes an even distribution on the partition {*H*, T_1, T_2}. As a motivation, one can imagine repeating the experiment again and again: then, *H*, T_1, and T_2 would each tend to come out true an equal proportion of times. Given this even distribution, you would work out the degree of credence you ought to assign to Heads by the law of total probabilities as follows: *P*(Heads) = *P*(Heads | It is now Monday) x *P*(It is now Monday) + *P*(Heads | It is now Tuesday) x *P*(It is now Tuesday) = 1/2 x 2/3 + 0 x 1/3 = 1/3. Your degrees of belief should therefore be *P*(Heads) = 1/3 and *P*(Tails) = 2/3 on any particular awakening (Elga, 2000; Dorr, 2002; Weintraub, 2004; Titelbaum, 2008).

Despite extensive debate, the disagreement between halfers and thirders on the SBP persists. One of the reasons why disagreement continues is that it is unclear whether or not self-locating information is *evidentially relevant* to beliefs about non-self-locating (or uncentred) hypotheses. According to thirders, learning self-locating information in the SBP impacts rational credences about the outcome of the coin flip (e.g., Horgan 2004; Weintraub 2004; Titelbaum 2008; Draper 2013). Halfers believe that self-locating information is evidentially irrelevant to uncentred hypotheses; and so it has no evidential impact on rational credences about the outcome of the coin flip (e.g., Lewis 2001; White 2006; Bradley 2012; Hawley 2013).

In this work, we bring Sleeping Beauty to the lab for the first time. We report the first empirical study addressing how naïve reasoners' judgment compare with the predictions of standard competing analyses of SBP. The primary aim of our investigation is to devise a transparent version of SBP along with some relevant variations in order to provide an empirical assessment of the

halfer's and thirder's interpretations of the problem. The following sections will thus present the study design and results. In the subsequent discussion we will consider the implications and relate our findings to the traditional debate on SBP.

2. Overview of the experimental scenarios

We constructed four different scenarios which we employed in two studies. Across these scenarios, we experimentally manipulated the kind of evidence available to participants in order to test the predictions of the halfers' and thirders' theoretical analyses. The design was between-subject, ruling out carry-over effects across conditions. Each participant read one vignette only, and expressed a judgment of probability on a 7-point Likert scale. A Likert scale was employed for its simplicity in use and understanding, although responses are not obviously translated into numerical probabilities. Indeed, we expected an interval scale to be sufficient to test the relevant competing hypotheses with our materials (see below).

To begin with, we adapted the standard SBP to make it as transparent as possible to naïve participants, in the following version (labelled *Basic*).

BASIC version

On a Sunday, you will be administered one of two pills, depending on the toss of a fair coin (HEADS: regular pill; TAILS: strong pill). You will not be told the outcome of the coin toss, and the two pills look identical. However, you know the following.

If the coin landed HEADS:
– the pill you're given on Sunday (regular pill) will make you sleep for one day;
– then you will wake up (on Monday).

If the coin landed TAILS:
– the pill you're given on Sunday (strong pill) will make you sleep for one day;
– then you will wake up a first time (on Monday), and shortly afterwards fall back asleep for another day, forgetting that you just woke up;
– then you will finally wake up a second time (on Tuesday).

Imagine you've just woken up. You don't know which day it is, and you do not know whether or not you have already woken up before. You are now asked to express your belief about the outcome of the coin toss that was made on Sunday.

As anticipated, the halfer's and thirder's predictions diverge critically in this basic version. According the thirder, one should judge *P*(Heads) = 1/3 and *P*(Tails) = 2/3. According to the halfer, instead, the correct answer here is *P*(Heads) = *P*(Tails) = 1/2, just as in the following *No Evidence* version, which we employed as a control condition.

NO EVIDENCE version

[same introductory paragraph as above]

If the coin landed HEADS:
– the pill you're given on Sunday (regular pill) will make you sleep for one day;
– then you will wake up (on Monday).

If the coin landed TAILS:
– the pill you're given on Sunday (strong pill) will make you sleep for two days;
– then you will wake up (on Tuesday).

Imagine you've just woken up. You don't know which day it is. You are now asked to express your belief about the outcome of the coin toss that was made on Sunday.

As *P*(Heads) = *P*(Tails) = 1/2 is uncontroversially the correct response in this version, participants' judgments should differ between the *Basic* vs. *No Evidence* condition in case they reason as thirders and self-locating information has impact on their credences. In order to gain a more fine-grained understanding of the evidential impact of self-locating information, and further disentangle halfers' and thirders' predictions, we relied on yet another benchmark variant, where ordinary (non-self-locating) and evidentially relevant information was involved. This third version we called *Marble*.

MARBLE version

On a Sunday, five small, empty, and closed boxes are placed in front of you; and you will then be administered one of two pills, depending on the toss of a fair coin (HEADS: regular pill; TAILS: strong pill). You will not be told the outcome of the coin toss, and the two pills look identical. However, you know the following.

If the coin landed HEADS:

– the pill you're given on Sunday (regular pill) will make you sleep for one day;
– meanwhile, one of the five boxes will be filled with a marble, then closed again (the other four boxes remain closed and empty);
– then you will wake up (on Monday), and open one of the five boxes at random.

If the coin landed TAILS:
– the pill you're given on Sunday (strong pill) will make you sleep for two days;
– meanwhile, all five boxes will be filled with five marbles (one each), then closed again;
– then you will wake up (on Tuesday), and open one of the five boxes at random.

Imagine you've just woken up. You don't know which day it is. You open one of the five boxes at random: you find a marble. You are now asked to express your belief about the outcome of the coin toss that was made on Sunday.

Like in the *No Evidence* version, the halfer's and the thirder's analyses must converge in this case. In fact, Bayes theorem implies that the probability of Tails given that one has found a marble is $P(\text{Tails}|\text{marble}) = P(\text{marble}|\text{Tails}) \times P(\text{Tails}) / [(P(\text{marble}|\text{Tails}) \times P(\text{Tails})) + (P(\text{marble}|\text{Heads}) \times P(\text{Heads}))] = (1 \times 1/2) / [(1 \times 1/2) + (1/5 \times 1/2)] = 5/6$. Moreover, and importantly, a significant analogy holds between the *Marble* version and SBP from a thirder's point of view. For the event of finding a marble could be one out of six, namely, the single one that could materialize after Heads, or each one of five which could materialize after Tails. Given that the box opening was random, there are thus five out six chances that the coin flip yielded Tails once one knows that the event of finding a marble actually occurred. In the thirder's analysis of the SBP, one can reason about awakenings in a similar way as we just did with marbles. In the halfer's analysis, on the contrary, the analogy is rejected entirely. As a consequence, the two analyses critically diverge as concerns the comparison of the *Marble* condition with the following counterpart version (labelled *Plus*).

PLUS version

On a Sunday, you will be administered one of two pills, depending on the toss of a fair coin (HEADS: regular pill; TAILS: strong pill). You will not be told the outcome of the coin toss, and the two pills look identical. However, you know the following.

If the coin landed HEADS:
– the pill you're given on Sunday (regular pill) will make you sleep for one day;
– then you will wake up (on Monday).

If the coin landed TAILS:

– the pill you're given on Sunday (strong pill) will make you sleep for one day;

– then you will wake up a first time (on Monday), and shortly afterwards fall back asleep for another day, forgetting that you just woke up;

– the same will happen on each of the following days, until you finally wake up a fifth time (on Friday).

Imagine you've just woken up. You don't know which day it is, and you do not know whether or not you have already woken up any time before. You are now asked to express your belief about the outcome of the coin toss that was made on Sunday.

The *Plus* version tries out the halfer's intuition further. According to the halfer, the switch from two to five awakenings (or ten, for that matter) would still leave the self-locating evidence irrelevant, so that judgments in the *Plus* condition are expected to differ from the *Marble* condition but not from the *Basic* (and *No Evidence*) condition. An opposite prediction arises from the thirder's analysis. Indeed, as anticipated above, responses in the *Plus* version should line up with those in the *Marble* version and differ from both the *Basic* and the *No Evidence* variants in case participants behave as thirders.

Table 1. Comparison between the halfers' and the thirders' judgments about the probability that the coin landed TAILS in different versions of the SBP.

	No Evidence	**Basic**	**Plus**	**Marble**
Halfers	1/2	1/2	1/2	5/6
Thirders	1/2	2/3	5/6	5/6

In summary, we experimentally manipulated the SBP's formulation as a function of the kind of evidence available to human reasoners (Table 1). We thereby addressed three related questions with our study: whether human reasoners acknowledge self-locating information as evidentially relevant in the SBP, whether the impact of self-locating information differs from the impact of objective statistical information like finding a marble in a randomly chosen box, and whether the standard theoretical accounts put forward by halfers and thirders are empirically adequate.

3. Experiment

3.1. Study 1

***Method*.** Two hundred and forty-three participants (Mean age, 38, SD = 11, male 137, female 106) were recruited using Amazon MTurk. We only allowed MTurk workers with an approval rate > 95% and with a number of HITs approved > 5000 to participate in our study. Instructions and material were presented in English on the Qualtrics Survey Software. Participants were randomly assigned to one of four experimental groups.

As pointed out above, four conditions were sufficient to disentangle relevant predictions from standard halfer and thirder accounts, thus putting them to empirical test. Halfers and thirders agree on their predictions that *P*(Tails) = 1/2 in the *No Evidence* condition, and that *P*(Tails) = 5/6 in the *Marble* condition. For the *Basic* and *Plus* conditions, instead, halfers and thirders disagree (Cisewski *et al*., 2016; see also Ross 2010). Halfers predict that *P* (Tails) = 1/2 in both the *Basic* and the *Plus* condition, since they claim that self-locating information bears no relevance relation with the outcome of the coin toss. Instead, thirders predict that *P* (Tails) = 2/3 in the *Basic* condition, and that *P* (Tails) = 5/6 in the *Plus* condition as a function of the partition of self-locating possibilities.

Participants read only one version of the SBP, and were asked to express their belief about the outcome of the coin toss in the situation described. Participants' responses were collected on a 7-point scale ranging from 'After waking up, I would think the coin toss on Sunday is certain to have been Heads and not Tails' to '— certain to have been Tails and not Heads' (midpoint was labelled '— equally likely to have been Heads or Tails'). Finally, participants were asked to indicate their age, sex, and level of education, and to enter a Qualtrics-generated survey code for MTurk for validating their participation. Because two participants failed to enter the correct code, their answers were not considered for analysis, leaving us with a sample of two hundred and forty-one participants (Mean age, 35, SD = 10, male 137, female 104).

***Results*.** A Kruskal-Wallis test showed that each of our four manipulations had a significant effect on participants' judgment, χ^2 (3) = 28.76, p = 0.000. Across conditions, we also found significant differences concerning the degree of certainty that the outcome of the coin toss was Tails (ranging in 4-7, i.e. from "equally likely" to "certain"), χ^2 (3) = 60.16, p = 0.000. A Dunn's test was further performed to test all possible pairwise comparisons between the four different conditions. The

results showed that difference of scores between *Basic* & *Marble*, *Marble* & *No Evidence*, *Marble* & *Plus*, and *No Evidence* & *Plus* were significant ($p < 0.01$), while the other comparisons (*Basic* & *No Evidence*, *Basic* & *Plus*) were not. We found no effect of age, sex, or education.

3.2. Study 2

***Method*.** Study 1 revealed that participants' judgments of the SBP depended on the type of evidence available. In particular, its results are consistent with the idea that naïve reasoners acknowledge self-locating information as relevant in the SBP (*Plus* condition). Study 2 examined whether these results may have been affected by a focus on the coin mechanism in the question participants were asked.

A new sample of two hundred and forty participants (Mean age 36, SD = 10, male 139, female 101) was recruited from Amazon MTurk. As in Study 1, we only allowed MTurk workers with an approval rate > 95% and with a number of HITs approved > 5000 to participate in our study. Instructions and material were presented in English on the Qualtrics Survey Software. Participants were randomly assigned to one of four experimental groups, and none took part in more than one experiment.

Unlike in Study 1, participants did not express their belief about the outcome of the coin toss. Instead, participants expressed their belief about the pill they were administered in the situation they were asked to consider. Otherwise, the versions of the SBP used in this second study were identical to the versions we used in Study 1. Responses were again collected on a 7-point Likert scale and participants again provided their age, sex, level of education, and a Qualtrics-generated survey code for validating their participation. Because three participants failed to enter the correct code, their answers were not considered for analysis, leaving us with a sample of two hundred and thirty-seven participants (Mean age 34, SD = 10, male 137, female 100).

***Results*.** A Kruskal-Wallis test showed that all groups differed significantly in their answers, χ^2 (3) = 17.43, p = 0.001. All groups differed significantly in their certainty of a Tails outcome, χ^2 (3) = 29.18, p = 0.000. A Dunn's test was further performed to test all possible pairwise comparisons between the four different conditions. The results showed that difference of scores between *Basic* & *Marble*, and *Marble* & *No Evidence* were significant ($p < 0.01$), while the other comparisons were not. Like in Study 1, we found no effect of age, sex, and education.

A Mann-Whitney Test showed that there was no significant difference between the answers of the participants of this study (M = 4.31) and the answers of participants from Study 1 (M = 4.24), p = 0.83. Aggregating data from both studies, the difference between the *Basic* and the *No Evidence* condition did not reach significance, p = 0.42. However, the aggregate analysis did reveal a significant difference between the *Plus* and the *Basic* condition, p = 0.03, d = 0.29, and between the *Marble* and the *Plus* condition, p = 0.03, d = 0.25 (Table 2 and Figure 1).

Figure 1. Mean scores and standard deviations (SD) for each group condition for Study 1, Study 2, and Studies 1 and 2 combined. Scores range from 1 = 'Certain that it was Heads and not Tails' to 7 = 'Certain that it was Tails and not Heads. Group conditions are on the horizontal axis; mean scores on the vertical axis. Aggregating data from both studies, the difference between the *Basic* and the *No Evidence* condition did not reach significance, p = .42. A statistically significant difference was found between the *Plus* and the *Basic* condition, p = .03, d = .29, and between the *Marble* and the *Plus* condition, p = .03, d = .25.

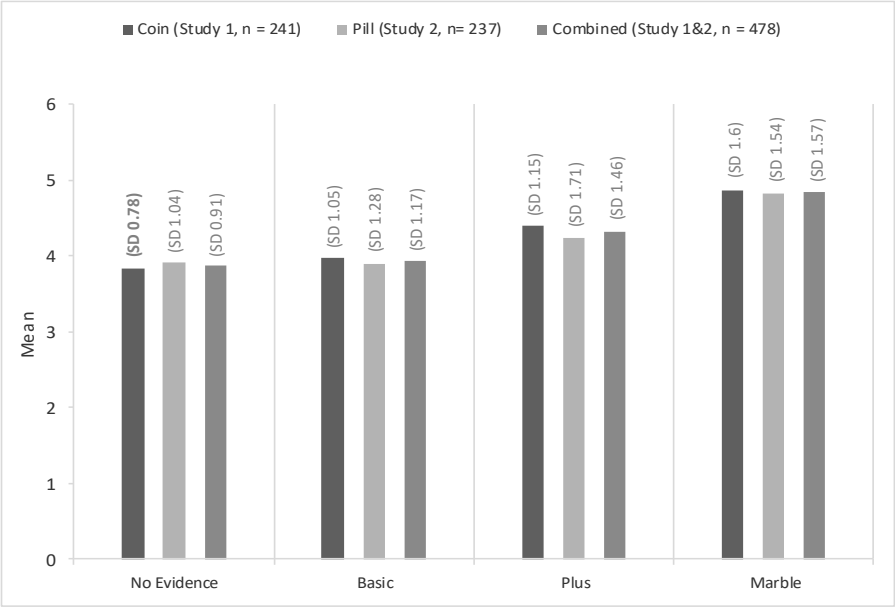

Table 2. Mean scores and standard deviations (SD) for each group condition. Responses ranged from 1 = "certain that it was Heads and not Tails" to 7 = "certain that it was Tails and not Heads".

Conditions	**No Evidence**		**Basic**		**Plus**		**Marble**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Coin (Study 1, n = 241)	3.83	.78	3.98	1.05	4.40	1.15	4.87	1.60
Pill (Study 2, n = 237)	3.92	1.04	3.90	1.28	4.23	1.71	4.82	1.54
Combined (1&2, n = 478)	3.87	.91	3.94	1.17	4.32	1.46	4.85	1.57

4. Discussion

Our results show that experimentally observed reasoning in the SBP did not simply fit either the halfer's or the thirder's analyses. The halfer's analysis is consistent with the lack of a significant difference between the *Basic* and *No Evidence* conditions, but is at odds with the finding that the *Plus* and *Basic* conditions reliably differed (recall that for the halfer one should have *P*(Tails) = ½ in both cases). The thirder's analysis, on the other hand, is consistent with the latter result, but is inconsistent with the finding that the probability of Tails is reliably judged to be higher in the *Marble* than in the *Plus* condition (for the thirder, one should have *P*(Tails) = 5/6 in both cases).

In order to evaluate the implications of our findings for the SBP, it is useful to distinguish two attitudes that one may have concerning a crucial issue, namely, the relationship between the normative and the descriptive aspects of the study of reasoning. The first attitude we will label *dividist*, following the Humean idea of a sharp logical divide between *is* and *ought*.

For the dividist, the *is* and *ought* of reasoning are essentially independent and should be kept so. According to this approach, the status of a normative analysis remains untouched by whatever descriptive finding. A dividist *halfer* would then see our results as indicating that naïve reasoners can consistently be misguided in problems involving self-locating evidence such as the SBP. People's inclination to provide responses higher than ½ in variants of the SBP (as illustrated, in particular, in our *Plus* scenario) would then amount to a bias of judgment akin to, say, the gambler's fallacy (e.g., Clotfelter & Cook, 1993; Terrell, 1998), and an illustration that humans' probabilistic reasoning may systematically fail to comply with rational standards. In this view, an explanation of the bias would then simply be left to further descriptive (psychological) research.

A dividist *thirder* could easily take a similar stance. Our participants departed from the correct solution of the SBP (now meant to be different from ½), and that's all there is to say. Interestingly, though, our results are likely to invite a more articulated reaction in this case, and one which deserves more careful discussion.[1] As it happens, a thirder might be tempted to contend or mitigate the conclusion that people are not reasoning as thirders. The thirder would initially point out (in fact, correctly) that our comparison of the *No Evidence* and *Basic* versions only yielded a negative result: no significant difference was detected with our procedure, which does not imply

[1] Here we have to acknowledge very helpful remarks from two reviewers.

that the two problems are generally taken to have exactly the same solution. But how should one then interpret the rest of our findings? Could they depend on a general tendency to *conservatism* in probability updating, according to which people's belief updating is generally conservative relative to the predictions of Bayesian conditioning (Phillips & Edwards 1966; Edwards 1968; Fischoff & Beyth-Marom, 1983; Slovic & Lichtenstein, 1972), so that the provision of new evidence would have less impact on people's beliefs than what Bayesian conditioning predicts? As interesting as it is, the idea of a general tendency to conservativism in probability updating still does not explain the difference we found between judgments in the *Marble* and *Plus* conditions, which a thirder is bound to consider structurally analogous from a probabilistic point of view.

A more subtle way to see our participants as quasi-thirders would go as follows. Sometimes appropriate facilitating conditions are in order for people's reasoning competence to emerge (see, e.g., Hoffrage *et al*., 2000; Pighin, Tentori, and Girotto, 2017). Consider an analogy with another famous probabilistic puzzle, the Monty Hall problem, which is known to invite "½" as a largely dominant (but mistaken) response because of the representational and computational difficulty of the task for the unaided human mind (Krauss & Wang, 2003). One idea is that our *Plus* condition fostered more correct responses in the SBP (from a thirder's perspective) much as it happens when we adapt the Monty Hall problem by multiplying the doors (from three in the standard version to a larger number; see Granberg, 1996). According to this reading, in a version of SBP with even more awakenings, even more people would become able to adequately appreciate the evidential relevance of the information provided, and the difference that we found between the *Plus* and the *Marble* versions would itself tend to vanish.[2]

The latter interpretation is surely clever and appealing, but ultimately unconvincing, we submit. To recap, it implies that very few of our participants were able to depart from response "1/2" and behave as thirders in the *Basic* condition, while a larger proportion should have done so in the *Plus* condition, with the normatively correct assessment (here, 5/6) being facilitated by the scenario

[2] Once properly specified, this hypothesis might find support from further investigation within our experimental paradigm. One could, in particular, consider how probability judgments differ in our *Basic* vs. *Plus* version (and/or some modification thereof) and make a quantitative comparison with the variation in the *Plus* vs. *Marble* version. Notice, however, that responses would have to lie on an interval scale for making this comparison. Given our elicitation procedure (with a 7-point scale), we deemed appropriate not to rely on this assumption in our analyses of data. (We thank the editor for raising this point.)

with multiple awakenings. Crucially, this pattern of behavior would generate a bimodal distribution in our *Plus* condition. In turn, a bimodal distribution in the *Plus* condition would show up in a systematically higher variance in comparison with the *Marble* condition (where no bimodal distribution is expected, according to the interpretation we're discussing). And yet, this is not what we found: the variance of participants' responses did not systematically increase from the *Marble* to the *Plus* condition (see *Table 2*). More generally, and for the same reasons, the idea that our participants were simply split into a group of halfers and a group of thirders does not account for our results.

We conclude that a dividist, of either the halfer or the thirder strand, would have to understand our results as showing that people's intuitive judgments of the evidential relevance of self-locating evidence are systematically biased in one way or another. However, dividism is not the only possible approach. A major alternative view can be labelled *explicationism*.

Epistemologists and philosophers of science have developed several probabilistic explications of the concept of *evidential relevance* (Fitelson 1999; Crupi, Chater, & Tentori, 2013; Crupi & Tentori, 2014; Brössel, 2013; Festa & Cevolani, 2016). One way of evaluating the adequacy of these alternative explications is to examine their degree of similarity to ordinary usage and judgment (Carnap, 1950; Kemeny & Oppenheim, 1952). To the extent that an explication will only illuminate a concept if it fits central cases of usage and judgment, an explicationist will allow that empirical results from the psychology of reasoning possibly bear on philosophical issues (see Schupbach 2017; Colombo 2017). In this perspective, psychological results will provide philosophers with data helpful to discover and assess cases in which a pluralistic approach to explication is in order; they can help philosophers to identify the explicandum's central features and their relation with other concepts; and they can point to sources of bias affecting philosophers' judgments themselves, including instances of plain normative uncertainty such as the SBP (Shepherd & Justus, 2015).[3]

[3] In fact, our notion of explicationism allows for significant nuances. Carnap (1950, p. 3) characterizes the task of explication as that of "transforming a given more or less inexact concept into an exact one, or, rather, in replacing the first by the second." According to Carnap (1950, p. 7), an adequate explicatum should be similar to the explicandum in respecting prior usage — though "close similarity is not required" and "considerable differences are permitted." It should be more exact than the explicandum. It should be fruitful in the sense of being "useful for the formulation of [...] empirical laws [or] logical theorems." And last, the explicatum should be simple. Given its emphasis on the requirement of fruitfulness, Carnapian explication can be aptly described as aiming at "concept engineering" (Kitcher 2008). Kemeny & Oppenheim (1952, p. 308), on the other hand, distinguished their project from Carnap's in these

Here, one might want to object that the SBP is decidedly not an ordinary scenario for the assessment of evidential relevance.[4] This is true, but inconsequential for an explicationist's project concerned with the SBP. For the SBP surely is a central case for intuitive judgments of evidential relevance *when self-locating information is involved*. To an explicationist, then, our results might suggest that an adequate account of the evidential relevance of self-locating information allow that self-locating information may trigger more cautious judgments of confirmation than familiar kinds of statistical evidence.

While illustrating potential implications of our findings for diverse perspectives such as dividism or explicationism is a relevant concern of the present work, adjudicating between those approaches would require itself a whole (and different) paper. Accordingly, our summary and conclusion is that our results indicate a pattern of judgment qualitatively different from either the halfer's or thirder's analyses of the SBP, where self-locating evidence is acknowledged as relevant but its quantitative impact is discounted significantly as compared to more standard statistical evidence. Other factors were previously shown to have such diluting effects on reasoning with evidence, such as second-order uncertainty about the values of a relevant statistical distribution (Tentori, Crupi, & Osherson, 2007, 2010). Interestingly, although "mixed" models of the SBP exist (Bostrom, 2007; Meacham 2008), they do not take into account this specific diluting, conservative effect involved in reasoning with self-locating information.

Acknowledgments. Work on this project was supported by the Deutsche Forschungsgemeinshaft (DFG) as part of the priority program *New Frameworks of Rationality* (SPP 1516).

---

terms: "The commonest procedure of explication is to apply a trial and error method till one arrives at an ingenious guess, and then try to find intuitive reasons to justify the proposed *explicatum.* This procedure is clearly very dangerous: The intuition of the most honest and well-trained philosopher is likely at times to become a tool for grinding an axe. […] We feel that we must first put down clearly all that our intuition tells us about the *explicandum,* and then find the precise definitions that satisfy our intuitive requirements." Given a stronger emphasis on the requirement of similarity, the goal of Oppenheimian explication is more one of concept clarification instead of concept engineering.

[4] We thank an anonymous reviewer for prompting this clarification.

References

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

Bostrom N. (2007). Sleeping Beauty and self-location: A hybrid model. *Synthese*, *157* (1), 59-78.

Bradley D. (2012). Four problems about self-locating belief. *Philosophical Review* 121: 149–177.

Brössel P. (2013). The problem of measure sensitivity redux. *Philosophy of Science* 80: 378-397.

Carnap R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Cisewski J., Kadane J.B., Schervish M.J., Seidenfeld T., & Stern R. (2016). Sleeping Beauty's credences. *Philosophy of Science*, 83: 324-347.

Clotfelter C. & Cook P.J. (1993). The 'gambler's fallacy' in lottery play. *Management Science*, 39: 1521-1525.

Colombo, M. (2017). Experimental philosophy of explanation rising: The case for a plurality of concepts of explanation. *Cognitive science*, *41*(2), 503-517.

Cozic M. (2011). Imaging and Sleeping Beauty: A case for double-halfers. *International Journal of Approximate Reasoning*. 52: 137–143.

Crupi V. & Tentori K. (2014). Measuring information and confirmation. *Studies in the History and Philosophy of Science* 47: 81-90.

Crupi V., Chater N., & Tentori K. (2013). New axioms for probability and likelihood ratio measures. *British Journal for the Philosophy of Science* 64: 189-204.

Dorr C. (2002). Sleeping Beauty: In defense of Elga. *Analysis* 62: 292-296

Draper K. (2013). The evidential relevance of self-locating information. *Philosophical studies*, *166*(1), 185-202.

Edwards W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment* (pp. 17–52). New York, NY: Wiley.

Elga A. (2000). Self-locating Belief and the Sleeping Beauty problem. *Analysis, 60* (2): 143-147.

Festa R. & Cevolani G. (2017). Unfolding the grammar of Bayesian confirmation: Likelihood and anti-likelihood principles. *Philosophy of Science* 84: 56-81.

Fischoff B. & Beyth-Marom R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review* 90 (3) 239-260.

Fitelson B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, 66: S362–S378.

Granberg D. (1996). The Monty Hall problem: To switch or not to switch. In M. vos Savant, *The Power of Logical Thinking* (pp. 169-196). St. Martin's Press, NY.

Hawley, P. (2013) Inertia, optimisim, and Beauty. *Noûs* 47:1 85-103.

Hoffrage U., Lindsey S., Hertwig R., & Gigerenzer G. (2000). Communicating statistical information. *Science* 290 (issue 5500): 2261-2262.

Horgan T. (2004). Sleeping Beauty awakened: New odds at the dawn of the new day. *Analysis*. 64: 10–21.

Kemeny J.G. & Oppenheim P. (1952). Degree of factual support. Philosophy of Science, 19: 307–324.

Kitcher P. (2008). Carnap and the caterpillar. *Philosophical Topics* 36: 111-127.

Krauss S., & Wang X.T. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General*, *132* (1), 3-22.

Lewis D. (2001). Sleeping beauty: Reply to Elga. *Analysis*, *61* (3): 171-176.

Meacham C. (2008). Sleeping Beauty and the dynamics of *de se* belief. *Philosophical Studies* 138:2, 245-269.

Phillips L.D. & Edwards W. (1966). Conservatism in a simple probability inference task. Journal of Experimental Psychology, 72, 346–354.

Piccione M. & Rubinstein A. (1997). On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior*, *20* (1), 3-24.

Pighin S., Tentori K., & Girotto V. (2017). Another chance for good reasoning. *Psychonomic Bulletin & Review*.

Ross, J. (2010). Sleeping Beauty, countable additivity, and rational dilemmas. *Philosophical Review*, *119*(4), 411-447.

Schupbach J.N. (2017). Experimental explication. *Philosophy and Phenomenological Research* 94: 672-710.

Shepherd J. & Justus J. (2015). X-phi and Carnapian explication. *Erkenntnis* 80: 381-402.

Slovic P. & Lichtenstein S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behavior & Human Processes* 6, 649-744.

Tentori K., Crupi V., & Osherson D. (2007). Determinants of confirmation. *Psychonomic Bulletin & Review*, 14: 877-883.

Tentori, K., V. Crupi, & Osherson, D. (2010). Second-order probability affects hypothesis confirmation. *Psychonomic Bulletin & Review*, *17*, 129–34.

Terrell D. (1994). A test of the gambler's fallacy: Evidence from pari-mutuel games. *Journal of Risk and Uncertainty* 8: 309-317.

Terrell D. (1998). Biases in assessments of probabilities: New evidence from greyhound races. *Journal of Risk and Uncertainty* 17: 151-166.

Titelbaum, M.G. (2008). The relevance of self-locating beliefs. *Philsophical Review* 117: 555–605.

Titelbaum, M.G. (2013). Ten reasons to care about the Sleeping Beauty problem. *Philosophy Compass*, *8* (11), 1003-1017.

Weintraub, R. (2004). Sleeping Beauty: A simple solution. *Analysis* 64(1): 8-10.

White, R. (2006). The generalized Sleeping Beauty problem: A challenge for thirders. *Analysis*, *66*(290), 114-119.