DEMYSTIFYING THE MYSTERY ROOM

Sylvia Wenmackers
KU Leuven, Institute of Philosophy, Centre for Logic and Philosophy of Science (CLPS)
Kardinaal Mercierplein 2 - bus 3200, 3000 Leuven, BELGIUM
sylvia.wenmackers@kuleuven.be

Forthcoming in *Thought* (accepted in 2019)

Abstract

The Mystery Room problem is a close variant of the Mystery Bag scenario (due to Titelbaum). It is argued here that dealing with this problem requires no revision of the Bayesian formalism, since there exists a solution for it in which indexicals or demonstratives play no essential role. The solution does require labels, which are internal to the probabilistic model. While there needs to be a connection between at least one label and one indexical or demonstrative, that connection is external to the probabilistic model that is used to determine the relevant conditional probability; hence, it does not complicate the update procedure.

Keywords: Bayesianism; *de se* beliefs; observation selection; probability; sample space

According to Lewis (1979), *de dicto* propositions describe what the world is like and *de se* propositions describe when or where a subject is in that world. When the indexicals in *de se* propositions cannot be replaced by *de dicto* propositions, they are said

to be irreducibly *de se*. An ongoing debate in the literature on the foundations of probability concerns how to update probabilities based on information that is irreducibly *de se*.

Following Lewis (1979), it is now generally assumed that the usual Bayesian approach that starts with possible worlds is sufficient for dealing with *de dicto* propositions, but some forms of indexical information are supposed to require centered worlds. As a result, there are multiple proposals for revising the Bayesian formalism and its update rule (*e.g.*, Halpern 2005, Meacham 2008, and Titelbaum 2014).

Rather than advocating for any such alternative, I want to question the assumption underlying this debate: is there such a thing as evidence that is both irreducibly *de se* and probabilistically relevant? If not, no revision of the Bayesian formalism is required. Moreover,[1] even if we accept the existence of information with an irreducibly *de se* aspect, it might still be the case that the *de se* aspect is not probabilistically relevant (and hence need not be represented in the probabilistic event space). This is indeed what I will argue for.

In a recent analysis of the Sleeping Beauty problem, I wrote that "[i]t is never necessary to conditionalize on statements with indexicals" (Wenmackers forthcoming, section 3.1). The main suggestion of that paper is that when a probabilistic scenario seems to require

[1] I thank an anonymous referee for suggesting this helpful distinction.

the addition of centered worlds, it simply has not been modelled with an adequate sample space yet.[2] Outside the context of the Sleeping Beauty problem, there might still be counterexamples to this claim. In particular, Titelbaum 'Mystery Bag' scenario (2014, pp. 235–236) seems a strong contender for offering a counterexample to the claim. The main goal of the present paper is to demonstrate that it does not succeed as a counterexample. This is achieved by specifying an adequate sample space for it.

So, although introducing a new formalism to model the phenomenon in question is prevalent in the extant literature, we find it to be unnecessary.

1. Mystery Room problem

Let n be the number of players in the scenario; then there are n–1 versus 1 balls of each colour. I rephrase and simplify the scenario here: whereas Titelbaum's original version had n=10 (2014, pp. 235–236), my minimal version has n=3.[3]

> THE MYSTERY ROOM PROBLEM (n=3)
>
> Assume you find yourself in the following situation: you are one of three externally indistinguishable players arranged in a perfectly symmetrical way – sitting in an equilateral triangle, centered in a cylindrical room. Assume that you have no context-insensitive expression to pick out who you are in this situation. The assumption includes that you do not remember your name or other things about your history. It is announced that a fair coin will be tossed and depending on its outcome a bag will be prepared: in case of heads, the bag will contain two black balls and one white ball; in case of tails, it will contain two white balls and one black ball. The three players in the room will not learn the outcome of the toss, but they will each receive one ball from the bag (distributed randomly) and only observe whether their own ball is black or white. The ball you receive is white. What is the probability that the coin toss result was heads?

Let me first make explicit a tension. On the one hand, there is always at least one person who receives a white ball and that person will receive the evidence from a first-person perspective, so it may be tempting to claim that nothing is learned here. On the other hand, in the case of heads, as compared to the tails case, it is less likely that you (or, as I will emphasize, any particular, independently selected player) will receive this evidence.

Titelbaum (2014) claims that this is a situation in which you only learn the centered evidence "I received a white ball". Yet, this evidence should increase your degree of belief in the coin having landed heads (an uncentered claim). As such, he regards this example as a counterexample against a particular extension of the Bayesian framework for updating on evidence that contains context-dependent, indexical information, due to Halpern (2005), based on preliminary work with Tuttle (Halpern&Tuttle 1993) and followed-up by Meacham (2008).[4] Titelbaum advocates a different extension of the

[2] Whether we call the elements of the suitably rich sample space 'possible worlds' or 'centered worlds' is less substantial if the usual Bayesian update rule holds.

[3] This is intended to avoid a combinatorial explosion of the sample space, which is proportional to n (when coarse-grained) or $n!$ (when fine-grained). Moreover, I consider the case where you receive a white ball, rather than a black ball, which makes the n=3 version analogous to the Sleeping Beauty problem.

[4] Titelbaum reconstructs their framework as a hierarchical procedure in which new evidence is used to redistribute one's degrees of belief firstly over centered worlds and secondly over uncentered worlds. Titelbaum points out that this approach is consistent with a "relevance-limiting thesis" (Titelbaum 2014,

Bayesian framework: the certainty-loss framework. Whatever the comparative virtues of Halpern–Tuttle–Meacham's approach and Titelbaum's alternative may be, we may ask whether it is necessary to extend the usual way of probabilistic modelling for the particular case of the Mystery Room. I argue that the answer is negative.[5]

I maintain that – in order to answer the question posed in the vignette – it is irrelevant that one of the players is supposed to be you. The relevant information is that (a) one of the players is selected beforehand, *i.e.*, independently of the outcomes of the toss and the draw and (b) the modeller can track this observer's evidence. Moreover, it is very common to use labels for indistinguishable individuals (in the broad sense) in probabilistic scenarios and we are free to apply this here as well.[6]

To disentangle the issues, it will help to consider a variant of the above problem, in which there are no players in the room, but the three balls will be distributed to the empty chairs by the same procedure as in the original scenario. There is an agent that gets information about one of the chairs (randomly selected beforehand). I will call this the "empty chair variant".

2. From 'I' to 'player *i*'

In what follows, I will distinguish between referential terms and mere labels:

- Referential terms include indexicals (such as 'I') and demonstratives (such as 'that'). They are used to establish a relation between elements of a mental or other model and objects, individuals or processes in (the agent's representation of) the actual world.
- In contrast, mere labels only serve to distinguish between various elements of an abstract structure.[7]

Labelling helps us to track in our model (at least one of) the chairs or players and evidence about them (or their evidence) through time. Because of the assumed symmetry of the situation, it is not possible to apply the labels to specific chairs or people in the room without using some demonstrative or indexical information (such as 'that chair over there' or 'me'). But we do not need to do that to arrive at the answer. We introduce (mere) labels i, j, and k to the chairs or players in the model, with i the chair about which we receive evidence or the player whose evidence we track and j and k the others.

In addition, it is helpful to think of the chairs as 1, 2, and 3, such that chair 1 will receive a black ball, the player on chair 3 will receive a white ball, and the player on chair 2 will receive a white or black ball, depending on the coin toss result. By extension, we can also think of each player on a chair as being assigned the corresponding number. You are

p. 232), such that changes in centered evidence do not influence degrees of belief assigned to uncentered possible worlds.

[5] This conclusion agrees with my analysis of the Sleeping Beauty problem (Wenmackers forthcoming).

[6] Although (re-)labelling can cause problems in situations with infinite sample spaces (see, *e.g.*, Bartha and Hitchcock, 1999), this need not worry us in this particular case.

[7] Given an abstract structure, we may give it an interpretation by establishing a relation between its labels and objects in the world, after which the labels can be used as referential terms, but on my terminology, they would then cease to be mere labels.

maximally uncertain about which number was assigned to the chair you are supposed to track or to you (or to any of the other chairs or players).

After the draw, you should update your probability assignment to the coin having landed heads to 1/3. Is there anything you learn that can be represented in an indexical- and demonstrative-free way? The answer is yes, for what you learn is not just that some chair or someone receives a white ball (which is the certain event, statistically independent of the coin toss outcome), but rather that a particular chair or person (arbitrary but fixed), selected independently of the coin toss and ball draw outcome, receives a white ball. You can compute the probability conditional on having received a white ball, but you should not conditionalize on this, unlike at least one other player. To verify that this is correct and to avoid ambiguities in these verbal descriptions, it helps to make the sample space explicit.

3. Sample space for the Mystery Room problem

For this part, it does not matter whether the chairs are empty or occupied. To keep it simple, we will only talk about the number of the chair.

First, we introduce some notation to represent the information from the word problem. We represent a black ball by B and a white ball by W. We assume the coin toss is certain to happen and represent the outcome 'heads' by H and 'tails' by ¬H. A fair coin is used:

$$P(H) = P(\neg H) = 1/2. \qquad (1)$$

The colour of the ball that chair 1 and chair 3 receive is independent of the coin toss result. By letting the number of the chairs correspond to the order of the letters, we can summarize the two outcomes as either BBW or BWW, corresponding to the coin landing heads (H) or tails (¬H), respectively. Hence,

$$P(BBW) = P(BWW) = 1/2.$$

At the start, you have no way of referring to 'your' chair or the others in a non-demonstrative way and you are maximally uncertain about the number assigned to it.[8] Nevertheless, you can label the chairs in your model (by i, j, and k, with i corresponding to the chair you are asked to track) and assume associated names (1, 2, and 3), but you are maximally uncertain about which labels corresponds to which name:

$$P(i=1) = P(i=2) = P(i=3) = 1/3$$
$$= P(j=1) = P(j=2) = P(j=3) = P(k=1) = P(k=2) = P(k=3). \qquad (2)$$

[8] Similarly, for the original scenario with players: at the start, prior to seeing their ball's colour, all players are maximally uncertain about which player they are or who any of the others are in the game. Moreover, the players have no way of referring to themselves or others in a non-indexical way.

In other words, initially, we can treat i as randomly selected from {1,2,3}; likewise for j and k. We track i's evidence, which is that i's ball from the bag is white. To arrive at the answer, it does not matter how we (in our role as probabilistic modellers) obtain i's new evidence as long as assumptions (a) and (b) from section 1 apply.

If all you need to do is to track evidence about a single chair, it suffices to consider $2 \times n = 6$ possibilities. The notation states the colours of the balls in the order they are distributed to the chairs (as before), using the position of i to represent whether i equals 1, 2 or 3:

$$B_iBW,\ B_iWW,\ BB_iW,\ BW_iW,\ BBW_i,\ BWW_i.$$

Listing the six possibilities in the same order as before, we can regard these events as the following intersections:

$$H \cap i{=}1,\ H \cap i{=}2,\ H \cap i{=}3,\ \neg H \cap i{=}1,\ \neg H \cap i{=}2,\ \neg H \cap i{=}3.$$

If we also want to take into account the evidence about other chairs, we need a richer sample space, with $2 \times 3! = 12$ possible outcomes:

$$\Omega = \{\, B_iB_jW_k,\ B_iB_kW_j,\ B_jB_iW_k,\ B_jB_kW_i,\ B_kB_iW_j,\ B_kB_jW_i,$$
$$B_iW_jW_k,\ B_iW_kW_j,\ B_jW_iW_k,\ B_jW_kW_i,\ B_kW_iW_j,\ B_kW_jW_i \,\}.$$

This sample space is depicted in **Fig. 1**. There are seven non-atomic partitions of Ω that may be of relevance for various questions about the problem: $\{H, \neg H\}$, $\{i{=}1, i{=}2, i{=}3\}$, $\{j{=}1, j{=}2, j{=}3\}$, $\{k{=}1, k{=}2, k{=}3\}$, $\{W_i, \neg W_i\}$, $\{W_j, \neg W_j\}$, and $\{W_k, \neg W_k\}$.[9] The twelve atomic possible outcomes are equiprobable, due to equations (1) and (2):[10] $\forall \omega \in \Omega\ \ P(\omega) = 1/12$.

To find the prior probability that chair i receives a white ball, we need to apply the law of total probability:

$$\begin{aligned} P(W_i) &= P(i{=}1) \times P(\, W_i \mid i{=}1\,) + P(i{=}2) \times P(\, W_i \mid i{=}2\,) + P(i{=}3) \times P(\, W_i \mid i{=}3\,) \\ &= 1/3 \times 0 + 1/3 \times 1/2 + 1/3 \times 1 \\ &= 1/2. \end{aligned}$$

Similarly for a black ball, and similarly for the other chairs.

[9] Two of the partitions are not represented by sets in **Fig. 1**: $\{j{=}1, j{=}2, j{=}3\}$ and $\{k{=}1, k{=}2, k{=}3\}$.

[10] For example, let's determine $P(B_iB_jW_k)$. $P(B_iB_jW_k) = P(BBW) \times P(\, i{=}1 \mid BBW\,) \times P(\, j{=}2 \mid BBW \cap i{=}1\,) \times P(\, k{=}3 \mid BBW \cap i{=}1 \cap j{=}2\,) = 1/2 \times 1/3 \times 1/2 \times 1 = 1/12$. Likewise for the other atomic events.

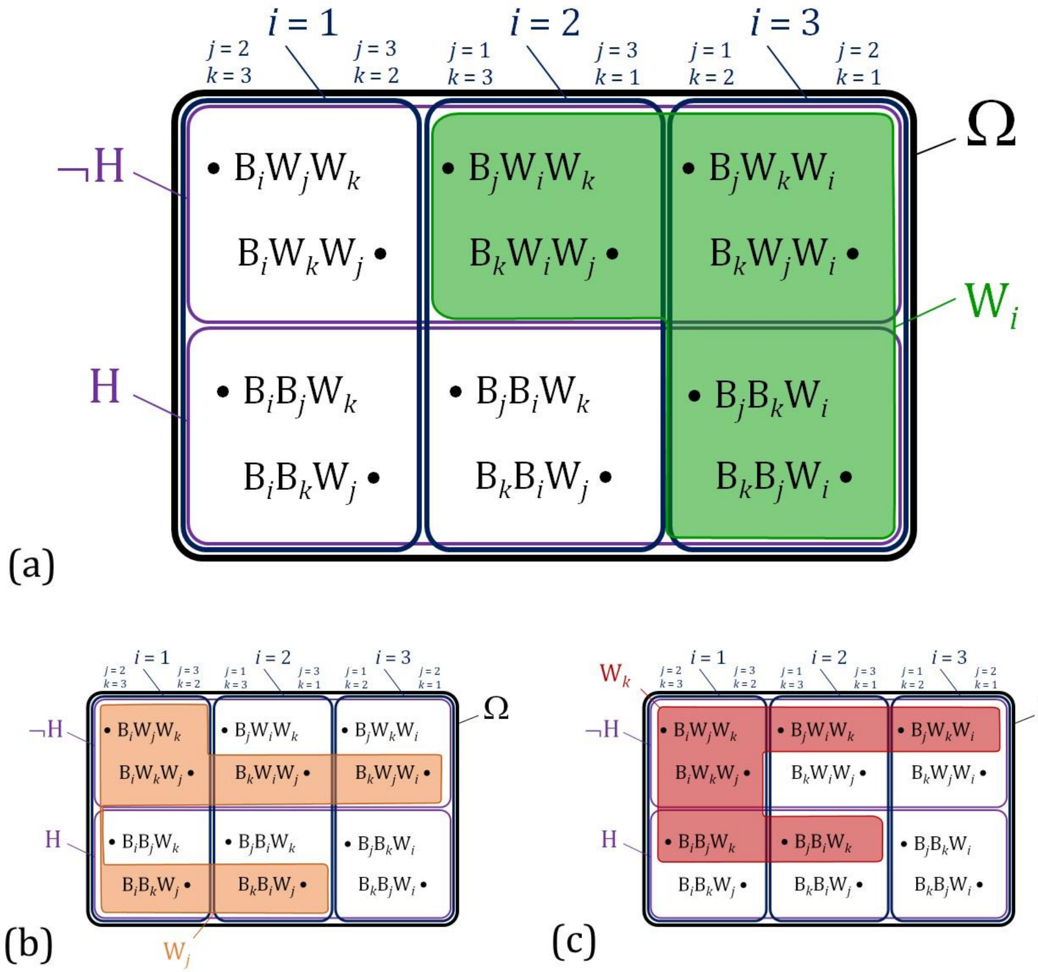

Figure 1: Euler diagram representing an adequate sample space (Ω) for analysing the Mystery Room problem. The elements are represented by twelve labelled dots. (a) W_i is the event that i receives a white ball; (b) W_j is the event that j receives a white ball; (c) W_k is the event that k receives a white ball. The complements of the events in (a)–(c) correspond to said chair or player receiving a black ball.

To compute the posterior, it is helpful to first compute:

$$P(H \cap W_i) = P(i{=}1) \times P(H \cap W_i \mid i{=}1) + P(i{=}2) \times P(H \cap W_i \mid i{=}2) + P(i{=}3) \times P(H \cap W_i \mid i{=}3)$$

$$= 1/3 \times 0 + 1/3 \times 0 + 1/3 \times 1/2 = 1/6.$$

The posterior that the problem asks for is given by the following conditional probability:

$$P(H \mid W_i) = \frac{P(H \cap W_i)}{P(W_i)} = \frac{1/6}{1/2} = 1/3.$$

Similarly for the other chairs, so:

$$P(H \mid W_i) = 1/3 = P(H \mid W_j) = P(H \mid W_k).$$

It may seem as if you can learn nothing from observing that the ball distributed to your chair is white because (i) you know beforehand that some chair will receive a white ball and (ii) when you observe a white ball, you know that there is a chair for which the evidence consists of a black ball.

Using our notation, (i) says:

$$P(H) = 1/2 = P(H \mid W_i \cup W_j \cup W_k).$$

The latter equality holds because the event $W_i \cup W_j \cup W_k$ (representing the event "*i*, *j*, or *k* receives a white ball") equals the entire sample space.

And (ii) says (applying the sum rule):

$$P(B_j \cup B_k \mid W_i) = P(B_j \mid W_i) + P(B_k \mid W_i) - P(B_j \cap B_k \mid W_i) = 1.$$

Alternatively, (ii) can be interpreted as follows: by receiving a white ball, you learn that *i* isn't 1. So, you also learn that one of the others must be chair 1, which definitely receives a black ball:

$$P(i{=}1 \mid W_i) = \frac{P(i{=}1 \cap W_i)}{P(W_i)} = \frac{P(\emptyset)}{P(W_i)} = 0,$$

$$P(j{=}1 \mid W_i) = 1/2 = P(k{=}1 \mid W_i).$$

The main point is that when you learn that *i* receives a white ball, you are allowed to conditionalize on "W_i". No more, no less. Conditionalizing further on "$\neg W_j \cup \neg W_k$" will not alter any of your probability assignments, and you do not have the evidence of "$\neg W_j$" nor "$\neg W_k$" (though you can compute the conditional probabilities).

Crucially, these conclusions are the same whether you are an external observer in the empty chair variant, or whether you are player *i* in the original scenario.

4. From conditional probability to conditioning

Observe that the previous section only deals with determining conditional probabilities. Except for the final paragraph, updating of probabilities (conditioning) is not mentioned. In the analysis of the original scenario and in the empty chair variant, there are two steps in the analysis, of which section 3 focused on the first one:

(1) The first step consists of a model in which any conditional probabilities can be computed using Bayes' theorem (*i.e.*, the mathematical theorem that states $P(A|B) = P(B|A) \times P(A)/P(B)$ for any events A and B, with $P(B) \neq 0$). In this step, we use mere inter-model labels, to keep track of different individuals *etc*.

(2) In the second step, a Bayesian agent applies that model to (certain aspects) of their own situation. This requires establishing a relation between labels internal to the probabilistic model and symbols for relevant aspects of their actual situation. It is in this second step that the agent's own perspective becomes crucial, for it determines which evidence they actually have. Such agents can still apply hypothetical reasoning and Bayes' theorem for any kind of counterfactual evidence (as in step 1), but they have to apply Bayes' rule, *i.e.*, to revise their probability assignments based on the maximal evidence E they actually receive: from P(A) to P(A|E) for all events A (provided $P(E) \neq 0$).

The evidence serves two functions associated with the two steps:

(1) The structure of possible pieces of evidence codetermines the appropriate sample space and the application of Bayes' theorem. In this step, only the *de dicto* aspects of the evidence are relevant. All that is probabilistically relevant about the evidence can be represented from an external perspective, even if the evidence is received in a *de se* way. For the Mystery Room, this was achieved by the move from I to player *i*.[11]

(2) The *de se* aspects of the evidence are crucially tied to the perspective of an agent: it informs them which pieces of possible evidence are actually received and hence need to be taken into account when applying Bayes' rule.

So, for the actual computation of numerical values, Bayes' theorem in step 1 is sufficient, and we only need the usual Bayes' rule in step 2. Provided that the sample space was rich enough to begin with, the *de se* aspect of the evidence does not influence any of the conditional probabilities; it only influences which conditional probabilities are relevant for updating.

So, in step 2, an actual player in a Mystery Room scenario has to establish a relation between the proposed model and the actual situation in order to apply the formal result in practice. The player has to recognize that the overall situation agrees with the abstract structure of the model. Then, the player may choose to connect the indexical "I"

[11] It might be suspected that – when moving to an external perspective – the modeller's first-person perspective acts as a stand-in for the player that is addressed as "you" in the vignette. However, observe that both the modeller and any of the players can compute any hypothetical player–evidence combination. Yet it is not true that the modeller actually learns what the player is said to learn in the vignette: the modeller shouldn't update, as the described player should. Only when the modeller becomes an actual observer, as in the empty-chairs variant, the modeller is required to update. Moreover, it is common in probability problems that we are being asked about a posterior probability assuming some evidence was received (by us or a particular agent in the scenario).

with the label "*i*". By itself this choice is inconsequential: picking *j* or *k* would work just as well, because of the symmetry apparent in **Fig. 1**. Of course, after this referential choice has been made, the player has to stick to it. Importantly, this connection between the indexical "I" and the label "*i*" is external to the probabilistic model (described in section 3). As such, it does not complicate the updating within the model: if the sample space is rich enough, standard Bayesian updating suffices. Hence, I conclude that the probabilistically relevant aspects of the evidence received by the player are not irreducibly *de se* and do not require any modification to the Bayesian formalism, merely diligent application of it.

Therefore, it seems that the move to 'centered worlds' was an unnecessary complication introduced by philosophers, which, moreover, is not fully general. By only making a distinction between *de se* and *de dicto*, it does not allow us to reason hypothetically about self-locating uncertainty of multiple agents (in step 1). Yet, this can be done with 'native' probability theory, by picking a rich enough sample space,[12] which prevents the need of any special update rule (in step 2) altogether.

5. Analogy between player *i* and Sleeping Beauty

Let us compare this analysis to that of the Sleeping Beauty problem. Karlander and Spectre (2010) agree with Weintraub (2004) that Sleeping Beauty learns something by awakening. In addition, they observe that during the first awakening, one can demonstratively refer to the awakening as 'this awakening': "It is customary to take this name as rigidly designating [an awakening], in this case [the first one], even if one does not know [which awakening] it is. Moreover, one knows by being awake that the rigidly designated [moment] is [a waking moment during the spell]". It is interesting that this rigid connection between a name and an 'individual' (object, person, location, or moment, in the Sleeping Beauty case) is also crucial in the Mystery Room example, although I argued that both indexicals and demonstratives are dispensable in step 1.

In many applications of probability theory, labels are used to refer to selected 'individuals', and my claim is that in the Mystery Room using an arbitrary label to track individual players suffices, just like labels can be applied to individual waking moments in the Sleeping Beauty problem. The rigid connection that we need functions as a pointer within the model (what I called mere labels), not between the model and individuals in reality (as is the case for referential terms).

Labels allow us to refer to and to track particulars within the model, even if we cannot point to the relevant object, person, location, or moment in reality. Moreover, labels allow us to do just this (and nothing more) in situations where we *can* point to something in reality. This is relevant for Sleeping Beauty, who can represent her awakening now as the n^{th} awakening, without knowing whether n equals 1 or 2. And it is relevant for the players in the Mystery Room, who can refer to themselves by a label '*i*', without being able to specify in a non-demonstrative way to which of the players it refers. Knowing in addition that n feels like now and i feels like me may be interesting for philosophers, but it is quite irrelevant to solving the probabilistic questions at hand.

[12] A sample space that is 'rich enough' might be very similar to a collection of centered worlds, but in some cases it may have to be more fine-grained to allow for representing possible evidence from multiple perspectives (as in the example of **Fig. 1**).

Unlike for the Sleeping Beauty problem, so far, no authors have claimed that the Titelbaum's (2014) Mystery Bag problem requires a modification of the Bayesian update rule.[13] Presumably, the same holds for the Mystery Room scenario (as presented in section 1). The isomorphism between the latter and the Sleeping Beauty problem may strengthen the support for the claim that no new update rule is needed for the latter either (as I defended in Wenmackers forthcoming).

6. Closing remarks

Toy problems such as the Sleeping Beauty problem and the Mystery Room demonstrate that centered evidence can be probabilistically relevant to uncentered events. While other authors have taken this to show that the Bayesian formalism has to be extended, I have shown that no new formalism is required.

I believe it is crucial to tease out the role of mere labels that are internal to models as opposed to indexicals or demonstratives that aim to point out particular individuals in reality. Doing so in toy problem such as the Mystery Room may help us to address observation selection and anthropic reasoning (see, *e.g.*, section 3.2 of Friederich 2017) adequately in the context of physical theories.

Acknowledgments

I am grateful to an anonymous referee of a previous paper (Wenmackers forthcoming) for pointing me towards the Mystery Bag scenario, as well as to an anonymous referee of the current paper for constructive comments that helped me to clarify my proposal. Part of this research was supported by the Research Foundation Flanders (Fonds Wetenschappelijk Onderzoek, FWO), Grant No. G0B8616N.

References

Bartha P. and Hitchcock C. "The shooting-room paradox and conditionalizing on measurably challenged sets" *Synthese*, **118** (1999) pp. 403–437.

Friederich S. "Fine-tuning" *Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), E.N. Zalta (ed.), URL: https://plato.stanford.edu/archives/win2018/entries/fine-tuning.

Halpern J.Y. "Sleeping Beauty reconsidered: Conditioning and reflection in asynchronous systems" In: T. Gendler and J. Hawthorne (Eds.) *Oxford Studies in Epistemology* Vol. 1, Oxford University Press (2005) pp. 111–142.

Halpern J.Y. and Tuttle M.R. "Knowledge, Probability, and Adversaries", *Journal of the ACM*, **40** (1993) pp. 917–962.

Karlander K. and Spectre L. "Sleeping Beauty meets Monday" *Synthese*, **174** (2010) pp. 397–412.

Lewis D. "Attitudes de dicto and de se" *The Philosophical Review*, **88** (1979) pp. 513–543.

Meacham C.J.G. "Sleeping Beauty and the dynamics of de se beliefs" *Philosophical Studies*, **138** (2008) pp. 245–269.

Titelbaum M.G. *Quitting Certainties.* Oxford University Press (2014) 2nd ed.

Weintraub R. "Sleeping Beauty: A simple solution" *Analysis*, **64** (2004) pp. 8–10.

Wenmackers S. "The Snow White problem" *Synthese* (forthcoming) DOI: 10.1007/s11229-017-1647-x.

[13] I thank an anonymous referee for bringing up this difference.