Relevance of the ensemble and actual theory distinction in the sleeping beauty problem

Minseong Kim*

(Dated: November 28, 2025)

Abstract

The halfer and the thirder solutions to the sleeping beauty problem are re-cast as relating to the time-averaged probability and the entanglement entropy of the Beauty in the quantum formalism. The thirder solution is correct whenever the ensemble theory view is used, which keeps the quantum nature of the coin. The halfer solution is correct for the actual theory view, which considers the quantum state of the coin to have been determined or collapsed into one outcome before Beauty wakes up for the first time in the experiment. The ensemble-actual distinction has recently become very relevant in the black hole information problem (marked by the arrival of replica wormholes) and provides a similar logical structure to the sleeping beauty problem (SBP) as well.

Keywords: sleeping beauty problem, ensemble theory, quantum entanglement, quantum-classical distinction, black hole information problem

* mkimacad@gmail.com; ORCiD:0000-0003-2115-081X

I. INTRODUCTION

The Sleeping Beauty problem (SBP), introduced by Adam Elga (Elga, 2000), presents a scenario in which self-locating uncertainty supposedly leads to change in epistemic and probabilistic beliefs or credence. Typically, we believe that a fair coin toss should not privilege one outcome over another, but SBP seems to suggest that this can change depending on self-locating uncertainty.

The purpose of this paper is to present a novel quantum physics-based perspective toward SBP. While concepts borrowed from quantum physics are used, this paper technically is not tied to quantum physics - quantum interference effects do not drive the results of this paper. Furthermore, the results are agnostic to interpretations of quantum mechanics.

At the center are 1) the time average of the entanglement entropy of the Beauty (and the role of entanglement in the quantum formalism), 2) the ensemble theory point of view against the actual theory point of view. The thirder argument $P(H) = 1/3$ (head) holds when the ensemble theory point of view is assumed, while the halfer argument holds when the actual theory point of view is taken.

In one of the simplest way to define an ensemble theory, which would suffice for this paper, it is about the evolution of an average (mean) state when evolved by different random Hamiltonians H_k. Mathematically with the common initial state $|\Psi_i\rangle$, with unitary propagator $U_k(t)$ corresponding to each Hamiltonian H_k,

$$|\Psi_{ens}(t)\rangle \propto \int dH_k \ U_k(t)|\Psi_i\rangle$$

where dH_k serves as a probability measure. In reality, only one of $H_k(t)$ is the actual Hamiltonian or the actual theory. The ensemble theory is invoked only because there is a lack of knowledge about which theory an observer is in.

In context of SBP, the ensemble theory view ignores the fact that the coin toss outcome is already observed by experimenters on Sunday and therefore has collapsed to one outcome, and this ignorance drives the non-trivial entanglement entropy trajectory. Note that in the quantum formalism, it is through entanglement that some system is connected to (or relates to) another system. That the entanglement entropy changes implies either that the probability or the connection has been modified.

By contrast, the actual theory view recognizes that the coin toss outcome is used on Sunday to pick only one out of the two candidate theories that evolve the Beauty state. The coin observation on Sunday completely eliminates the entanglement between the coin and the Beauty, and they only relate through the chosen theory. Since uncertainty experienced by Beauty only exists in the theory realized in the universe, the halfer argument $P(H) = 1/2$ holds.

A. Black hole physics inspiration

The distinction between the ensemble theory and the actual theory is inspired by black hole physics. In JT/RMT (Jackiw-Teitelboim gravity/random matrix theory) duality (Penington *et al.*, 2022; Saad *et al.*, 2019), the gravitational theory that pictures the spacetime we live in is considered to be an ensemble theory that describes an ensemble (mean configuration) of multiple theories with different evolutions of the universe state. The 'probability' assigned to the theories in an ensemble can be treated as classical. In the quantum gravitational context, an ensemble theory is an effective (field) theory that is used because some quantum gravitational degrees of freedom remain inaccessible to observers.

In normal circumstances, an ensemble theory can approximate the actual theory without a problem, but the approximation breaks down in extraordinary circumstances like black hole evaporation. The calculations show that the entanglement entropy of a JT black hole in an ensemble theory increases monotonically, while its entropy in an actual theory briefly increases and then decreases back to zero (Penington *et al.*, 2022). The mathematical details are not replicated here since this paper is not about black hole physics, but essentially the same idea is applied in this paper. No knowledge of JT/RMT or physics, other than the very basic formalism used in quantum physics, is required for the arguments of this paper.

B. Sleeping beauty problem, setup

A fair coin is thrown on Sunday (head probability $P(H) = 1/2$, tail probability $P(T) = 1/2$). Regardless of the outcome, the Beauty is put to sleep afterwards and must be awakened by others to wake up. An amnesia pill assures that she does not remember previous awakenings. If the result of the coin toss is head H, then she is awakened only on Monday Mon and then put back to sleep. Otherwise, she is awakened on both Monday Mon and Tuesday Tue. Upon awakening, she is asked for her credence that the coin toss result is H.

II. THE ENSEMBLE THEORY VIEW (THIRDER ARGUMENT)

The conventional thirder (Elga, 2000) argument can be cast in terms of the time average of the Beauty entanglement entropy and the conditional probability of head H in the ensemble theory.

The ensemble theory consists of two potentially time-dependent Hamiltonians (or actual theories) $H_{head}(t)$ and $H_{tail}(t)$, which evolves the state of the Beauty from Sunday to Wednesday. The Hilbert space of the universe goes as:

$$\mathcal{H} = \mathcal{H}_C \otimes \mathcal{H}_B$$

where $\mathcal{H}_C$ refers to the Hilbert space of the coin toss outcome and $\mathcal{H}_B$ is the Hilbert space of the Beauty.

The ensemble theory view ignores the coin toss outcome collapse after being measured-observed to select only one of the actual theories on Sunday. The Hamiltonian H_{ens} of the ensemble theory is then given as:

$$H_{ens}(t) = |H\rangle\langle H| \otimes H_{head}(t) + |T\rangle\langle T| \otimes H_{tail}(t)$$

where $|H\rangle, |T\rangle$ refer to the coin outcomes and H_{head}, H_{tail} refer to the Beauty Hamiltonians ('actual theories').

The initial state of the universe at the end of the Sunday is given as:

$$|\Psi^{Sleep}_{Sun}\rangle = \frac{1}{\sqrt{2}}\left(|H\rangle|Sleep\rangle + |T\rangle|Sleep\rangle\right)$$

where $|Sleep\rangle$ refers to the state of Beauty asleep. Right after Beauty wakes up on Monday, the universe state is:

$$|\Psi^{Awake}_{Mon}\rangle = \frac{1}{\sqrt{2}}\left(|H\rangle|Awake\rangle + |T\rangle|Awake\rangle\right)$$

The entanglement entropy of Beauty is zero in $|\Psi^{Awake}_{Mon}\rangle$.

At the time when Beauty might be awakened on Tuesday, the universe state is:

$$|\Psi^{Awake?}_{Tue}\rangle = \frac{1}{\sqrt{2}}(|H\rangle|Sleep\rangle + |T\rangle|Awake\rangle)$$

The entanglement entropy of the Beauty is non-zero - it is $\ln 2$. The time average of the entanglement entropy (on time Monday and Tuesday) S_{beauty} is therefore:

$$\mathbb{E}_{time}[S_{Beauty}] \equiv \langle S_{Beauty}\rangle_{time} = \ln 2/2$$

It is through this change in entanglement that the necessity of credence update arises in the ensemble theory. The time-averaged entropy is to be consistent with other time-averaged observables, including the case where we consider the time-averaged probability of head $|H\rangle$ conditioned on state $|Awake\rangle$:

$$P_{time}(H|Awake) = \lim_{N\to\infty} \frac{\#HeadMon}{(\#AwakeMon + \#AwakeTues)} = \frac{1}{3}$$

where N is the number of trials, and P_{time} refers to the time-averaged probability.

III. THE ACTUAL THEORY VIEW (HALFER VIEW)

When we switch to the actual theory point of view, the analysis becomes entirely different. First, the entanglement entropy of Beauty is always zero, since one theory - whether $H_{head}(t)$ or $H_{tail}(t)$ - is already chosen on Sunday:

$$|\Psi^{Sleep}_{Sun}\rangle = |H \text{ } or \text{ } T\rangle|Sleep\rangle$$

$$|\Psi^{Awake}_{Mon}\rangle = |H \text{ } or \text{ } T\rangle|Awake\rangle$$

$$|\Psi^{Awake}_{Tue(H_{tail(t)})}\rangle = |T\rangle|Awake\rangle$$

$$|\Psi^{Sleep}_{Tue(H_{head(t)})}\rangle = |H\rangle|Sleep\rangle$$

This means that randomness from the Beauty point of view only exists in her uncertainty over what theory evolves her state - $H_{head}(t)$ or $H_{tail}(t)$. Fairness of the coin toss means $P(H_{head}(t)) = 1/2$, which should be used as the credence for head $P(H|Awake) = 1/2$ - the halfer argument (Lewis, 2001).

More precisely, the same time average calculation but under the actual theory view yields the following due to the lack of entanglement:

$$P_{time}(H|Awake, H_{head}(t)) = \lim_{N\to\infty} \frac{\#HeadMon}{(\#AwakeMon + \#AwakeTues)} = 1$$

$$P_{time}(H|Awake, H_{tail}(t)) = \lim_{N\to\infty} \frac{\#HeadMon}{(\#AwakeMon + \#AwakeTues)} = 0$$

By the Bayes rule, $P(H|Awake) = 1/2$ is confirmed.

Another analysis focuses on entanglement. Since there is no entanglement between Beauty and the coin, there is no connection or information the Beauty has (in order) to update her initial knowledge that the coin toss has been fair. In other words, her state $|Awake\rangle$ provides no information because there is no entanglement between the coin and Beauty. Therefore, $P(H|Awake) = 1/2$.

IV. DISCUSSIONS/CONCLUSION

For representing the uncertainty that Beauty actually experiences, the actual theory point of view should be used on Monday/Tuesday and thus the halfer argument: $P(H|Awake) = 1/2$. However, observers on Sunday (who are not the Beauty on Monday and Tuesday) before they observe the coin toss outcome should follow the ensemble theory view and the thirder argument, thereby assigning $P(H|Awake) = 1/3$, since a theory is yet to be determined.

This suggests a major difference between quantum probability (or quantum credence) and classical probability (or classical credence). If we leave the coin toss outcome to be in a quantum superposition on Sunday without observers observing, then the ensemble theory remains valid, and the thirder argument will continue to hold. The only measurement in this experiment would be Beauty waking up, and the standard collapse or the many-worlds arguments may be used to support the thirder argument (Groisman *et al.*, 2013; Papineau and Dura-Vila, 2009; Sebens and Carroll, 2018; Vaidman, 2001).

The key issue that has gone neglected is that a measurement-observation occurs for the coin toss outcome in the conventional setup, which changes the nature of the experiment. There is no longer actual quantum wavefunction collapse or world branching (in case of the many-worlds Everettian interpretations) after the coin toss outcome is measured, unlike many of the setups that utilize the quantum formalism to tackle SBP. For example, (Groisman *et al.*, 2013; Papineau and Dura-Vila, 2009; Sebens and Carroll, 2018; Vaidman, 2001) (the list is only partial) utilize some form of a quantum coin to support the thirder argument - indeed this can be trivially seen, given that they need quantumness to invoke the Everettian or the many-worlds interpretations - but the result is obtained by assuming quantumness of the coin. Whenever the coin is fully quantum, it is fully consistent with the ensemble theory point of view and the actual theory point of view should not be used. However, when the coin outcome is already realized and the probability of the coin is already classical *before* the Beauty is awakened on Monday, the actual theory point of view is relevant for the uncertainty faced by Beauty in the actual world.

Again, nothing in this paper actually contradicts the mathematical and logical aspects of the cited Everettian papers, as long as the coin is assumed to remain quantum. It is just that the actual experiment setup of conventional SBP is different from what these papers present. This could have been less important in preceding years, but after the arrival of replica wormholes in quantum gravity where the difference between an ensemble theory and an actual theory is critical (Penington *et al.*, 2022), the distinction can no longer be safely ignored. After all, this distinction entirely changes whether the entanglement entropy of a black hole ends up going back to zero (thereby maintaining the pure state evolution of the universe after a complete black hole evaporation) in the actual theory or not in case of the ensemble theory. SBP (sleeping beauty problem) demonstrates another case where this distinction may be critical.

DATA AVAILABILITY AND DECLARATION OF INTERESTS

The author(s) have no funding source to declare. Furthermore, there is no conflict of interests.

REFERENCES

Elga, Adam (2000), "Self-locating belief and the sleeping beauty problem," Analysis **60** (2), 143–147.

Groisman, Berry, Na'ama Hallakoun, and Lev Vaidman (2013), "The measure of existence of a quantum world and the Sleeping Beauty Problem," Analysis **73**, 695–706.

Lewis, David (2001), "Sleeping Beauty: reply to Elga," Analysis **61** (3), 171–176.

Papineau, David, and Victor Dura-Vila (2009), "A thirder and an Everettian: a reply to Lewis's 'Quantum Sleeping Beauty'," Analysis **69**, 10.1093/analys/ann012.

Penington, Geoff, Stephen H. Shenker, Douglas Stanford, and Zhenbin Yang (2022), "Replica wormholes and the black hole interior," Journal of High Energy Physics **2022** (3), 205.

Saad, Phil, Stephen H. Shenker, and Douglas Stanford (2019), "JT gravity as a matrix integral," arXiv e-prints arXiv:1903.11115 [hep-th].

Sebens, Charles T, and Sean M. Carroll (2018), "Self-locating Uncertainty and the Origin of Probability in Everettian Quantum Mechanics," The British Journal for the Philosophy of Science **69** (1), 25–74.

Vaidman, Lev (2001), "Probability and the many-worlds interpretation of quantum theory," arXiv e-prints arXiv:quant-ph/0111072 [quant-ph].