Addendum to 'Relevance of the ensemble and actual theory distinction in the sleeping beauty problem': the philosophical importance of the quantum formalism (backward induction paradoxes and black hole information problem)

Minseong Kim*

(Dated: December 1, 2025)

Abstract

This is an addendum (memo) that goes beyond the sleeping beauty problem. I discuss why 1) the sleeping beauty problem (SBP) essentially has the same structure as the black hole information problem, 2) how the quantum formalism addresses irreducible or fundamental uncertainty (though this would differ from unquantifiable Knightian uncertainty) and how this addresses the question of vagueness. 3) I discuss why the backward induction paradoxes are not problematic under the quantum formalism, and whenever the concept of rationality is at question (or being questioned), rationality should be treated as a vague concept.

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

I. SBP = BLACK HOLE (BH) INFORMATION PROBLEM?

Structurally we can think of SBP as a simplification of the (black hole) information problem.

Consider observers that are inside some black hole. From the point of the black hole exterior, their time do not correspond to the exterior point of time. Coordinate-wise, the exterior time coordinate becomes spatial inside the black hole, while the exterior spatial coordinate becomes temporal. Therefore, the observers inside the black hole are essentially facing temporal self-locating uncertainty when they are trying to cast their story in terms of an observer on the black hole exterior.

We need to do more than just self-locating uncertainty to justify SBP being a simplification of the information problem, which is to consider the black hole evaporation process (though this will be drastically simplified). Observer C enters the black hole at exterior time $t = 0$. At $t = 1$, it is known that regardless of outcomes, C will not escape the black hole. At $t = 2$, C may escape the black hole with the probability of 1/2. This is essentially the SBP setup.

Now C asks that conditional on C being inside the black hole, what is the credence to be assigned for the probability of escaping a black hole at $t = 2$. If the conventional thirder solution is correct, then it is 1/3. But this is a paradox - the information paradox.

If the actual versus ensemble distinction does correctly resolve SBP, then we could see how the information problem may conceptually be resolved. The answer is holography of information, which states that the information regarding the black hole interior is already on the exterior. There is no actual uncertainty regarding C's history represented on exterior time t, and the theory that evolves C s already determined. C only experiences illusions of uncertainty because C does not have access to exterior degrees of freedom. When C knows this, C can self-correct for this and produce a coherent exterior account of C that nevertheless reflects C's uncertainty.

C's Page curve is therefore actually trivial - zero all the time. However, reflecting ensemble uncertainty, we can produce a coherent probabilistic story that reproduces the conventional Page curve with corrections (in case of SBP, probability of head from 1/3 to 1/2, and for the information problem, replica wormholes) as well.

II. VAGUENESS AND IRREDUCIBLE UNCERTAINTY ON THE QUANTUM FORMALISM

Consider some color that can either be considered as blue or green. The sorites paradox suggests why this can be problematic. A small perturbation to some color should not change the color label. Yet if this holds, then any color would turn out to be blue or green, destroying the distinction between blue and green.

We could say that some color - say skygreen - is 30% blue and 70% green. But by making such a statement, we usually mean that because this color is titled toward being green, it should be referred to as green. The probabilities do not reflect classical uncertainty - even if we keep measuring, we will not consider skygreen to be blue 30% of the time.

Furthermore, we cannot consider skygreen to consist of 30% blue and 70% green (though we could say that in terms of mixing physical paints or dyes), where percentages are not probabilities but proportions. Skygreen is a distinct color from the ideal blue and the ideal green. On top of this, we have to make sure that skygreen is vaguely similar to neighboring colors.

This problem is neatly resolved in the quantum formalism where we could simply state skygreen as:

$$|skygreen\rangle = \sqrt{0.3}|blue\rangle + \sqrt{0.7}|green\rangle \tag{1}$$

$|skygreen\rangle$ does not say that it consists of blue and green. $\langle blue|skygreen\rangle$ only establishes non-orthogonality. If we are to represent compositions, then we would have a state like $|b\rangle|b\rangle|b\rangle..|g\rangle|g\rangle$ instead, where $|b\rangle$ substitutes for $|blue\rangle$ and $|g\rangle$ for $|green\rangle$. $|skygreen\rangle$ does not have to be measured to collapse into one of blue or green, and yet maintains some relations to blue and green, reflecting non-classical and fundamental uncertainty.

When asked a binary question of if skygreen is blue or green, if we take the 50% threshold rule, then we can answer skygreen to be green as $|\langle green|skygreen\rangle|^2 > 1/2$, which is a deterministic (and non-probabilistic) inner product result. We can therefore simultaneously establish that for neighboring colors, their inner products will be close to 1 - vagueness - while escaping from some conceptual issues.

Color bluegreen-49, which may be considered 49% blue and 51% green is green, while bluegreen-51 (51% blue and 49% green) is blue despite these two colors essentially being similar, and the reason why we seem to get the sorites paradox is because we are invoking the wrong tool for addressing the color question. While it is true that $\langle bg51|bg49\rangle \approx 1$, for answering the question of whether a color is blue or green, we must take $|\langle blue|\cdot\rangle|^2 > 1/2$? instead.

You may say that the 50% rule should be already known, so I must be doing something stupid. It is not, because the reason why the above solution does not fare well classically is that classical probability is ill-suited for the 50% rule, which has so far been the main point of this section. If I say bluegreen-49 is green because I would, at 49% of time, call it green, then it seems very irrational to call bluegreen-51 which is essentially the same color to be blue, because I am pretty much thinking of this color green about the same time as bluegreen-49. So we must escape this classical uncertainty in order to justify the 50% rule.

A. Ship of Theseus

The ship of Theseus requires a different vagueness treatment on the quantum formalism. It actually asks how different microstates share the same effective state - the question of a microstate theory (or say, quantum gravity) and an effective theory (effective quantum field theory, such as the Standard Model, or even a classical theory).

The ship of Theseus only becomes problematic when we insist 'constitution' (that is, relevant to 'consists of,' not some legal books that nobody seems to agree on) to be discussed strictly in the microstate theory. Or more precisely, the question of whether some microstate constitutes some effective state cannot be addressed in both the microstate theory (because it does not directly deal with emergent approximate states) and the effective theory (because it cannot deal with microstates), and we have to use meta-theoretical understandings. This issue is what causes the ship of Theseus problem in the quantum formalism.

Identity is largely an emergent phenomenon to be discussed in the effective theory.

B. Backward induction paradoxes

1. *Chainstore paradox and vagueness of rationality*

We can easily see why the chainstore paradox is actually an example of vagueness. A monopolist facing N potential competitors sequentially is better off by not taking the rational backward induction solution. But the issue is when the monopolist is compelled to follow the rational backward induction solution. After all, when facing the Nth competitor, no one would credibly believe that the monopolist will choose to fight when the competitor enters the market while suffering the last loss. Therefore, as the monopolist approaches the competitors close to the Nth one, it would have to be thought that even if the monopolist can choose to go irrational, its probability becomes lower.

The very definition of 'rationality' is actually a vague concept. The backward induction solution is supposedly rational, but why is it so? Answering this often leads to a logically tautological one or essentially amounts to the sorites paradox. The monopolist faces the $N-1$th competitor, which should be indistinguishable from facing the Nth competitor. But the $N-2$th competitor should be indistinguishable from the $N-1$th competitor for the monopolist, and so forth - the sorites paradox.

Therefore, we have to re-cast rationality as a vague concept that gets called by some rule like the 70% rule or something. Consider the following quantum state representing some game strategy:

$$|\psi_t\rangle = a_{0t}|BI\rangle + \sum_{k=1}^{K} a_{kt}|IR_k\rangle \tag{2}$$

where $|BI\rangle$ is a backward induction strategy state with $\langle BI|IR_k\rangle = 0$ and $\sum_k |a_k|^2 = 1$, with t indexing time or some competitor. Then sequentially in reverse, the monopolist asks competitors: 'I know that at $t = N$, the last competitor, $a_{0t} = 1$. Now at $t = N-1$, I set $a_{0t} < 1$ but $a_{0t} \approx 1$. $\langle\psi_{t=N-1}|\psi_{t=N}\rangle \approx 1$. Now I also set $\langle\psi_{t-1}|\psi_t\rangle \approx 1$. Would you call my strategies irrational, if not optimal?' The monopolist invokes the idea of irrationality being vague, thereby forcing the sorites paradox to competitors without actually causing a paradox, since competitors will surely say that the monopolist cannot be considered irrational. And when optimized over all possible 'non-irrational' strategies, we get some probabilistic deterrence strategy profile that is dominant over the purely backward induction strategy profile.

In this case, the quantum formalism drastically simplifies the required formalism for addressing this question via the inner product, though the importance of the quantum formalism really has to trace back to SBP and the bluegreen-skygreen problem discussed before.

This is not Timothy Williamson's margin for error, since the Williamson theory is about 'knowing.' Agents in this chainstore problem are not really allowing for errors. Rather, they are asking what are acceptable rational strategies we can consider. The conventional backward induction strategy simply throws some of the strategies to be unacceptable, when it is the concept of rationality and acceptability that is at question here. And vagueness around the concept of rationality allows the monopolist to exploit the sorites paradox validly.

From the skygreen problem you may think that settling vagueness that way would mean that for time $t < t'$, strategies that are not the pure backward induction strategy will be considered irrational. This is an invalid analysis, because the chainstore problem asks how a forward-looking potential competitor would think based on future actions the monopolist may carry out. We are not evaluating a particular monopolist strategy at some time t independently from future times, and asks to label it as rational or irrational. We can do that, but it is irrelevant for the competitors. All the forward-looking competitor at time t care are what it would believe if the monopolist makes some action at time $t + 1$.

In contrast to the typical 'backward-looking competitor' support for the deterrence strategy in the chainstore problem, it is actually this forward-looking behavior that supports the deterrence monopolist strategy. The backward-looking logic actually confuses readers to think that the monopolist strategy at each time t should be judged individually. After all, if we are to evaluate the monopolist strategy at $t = 1$ without considering future actions, then one is forced to go 'skygreen' and destroy the sequential information of the game in handling the question.

2. Surprise examination paradox, identical to the chanstore paradox

The surprise examination paradox can be handled similarly because it is identical to the chainstore paradox, though the question of 'surprise' is a much more tricky one. From the chainstore paradox, we know that the exam cannot be held on Friday. However, from that time on (backward), the possibility of the exam being held can increase so that Monday has the maximum probability of the exam being held.

The surprise examination paradox then can twist this around, and the students with this probabilistic knowledge, may expect the exam to be held on Monday. However, there are flaws with this reasoning, because the probability of the exam is just highest on Monday, not 1. Therefore, while a perfect surprise is impossible, a decent amount of surprise would probabilistically be possible.

Again, the quantum formalism is not really fundamental (though very convenient and useful due to the inner product structure) here, and its importance really traces back to SBP and the skygreen (or the glue color) problem, where the concepts of vagueness and uncertainty are fundamentally being tested.