논리연구 15-1(2012) pp. 107～153

# Sleeping Beauty: a Garberian Approach*

Namjoong Kim

【Abstract】 In my previous paper (2009), I defended a solution of the Sleeping Beauty problem, according to which, on Monday, Sleeping Beauty assigns a lower credence to the coin's landing heads than 1/2. This conclusion was largely favorable to the Thirder view. However, even if my defense of the Thirder view was successful, it left one important question to be unanswered: Where did the Halfers go wrong? Their main argument was simple: Because Sleeping Beauty does not receive new and relevant evidence about how the coin lands, her credence in its landing heads should remain to be the same. But note that, if the Thirder view was right and Sleeping Beauty receives no new and relevant evidence, then this becomes a special case of the so-called old evidence problem (Glymour, 1980). In this paper, I will explain why it is rational for Sleeping Beauty to change her credence despite the lack of new evidence about the coin's landing heads. For this explanation, I will use Garber's well-known solution to the old evidence problem.

【Key Words】 Sleeping Beauty, Old evidence, Garber, De se credence, Self-locating belief

* 접수일자: 2011.12.05. 심사 및 수정 완료일: 2011.12.31. 게재확정일: 2012.02.02.

## 1. Introduction

In his (2000), Adam Elga presents a puzzling paradox: On Sunday, Sleeping Beauty, a paragon of probabilistic rationality, knows that she will go through the following experiment: On that night, she will be put to sleep by evil experimenters. Then, they toss a fair coin.

Case 1: (*HEADS*) The coin lands heads. In this case, she is awakened only once on Monday. Case 2: (*TAILS*) The coin lands tails. In this case, she is awakened twice, first time on Monday and the second time on Tuesday. Between these two awakenings, they inject a drug that erases her memory of Monday (so that she wakes up with the same memory in both Monday and Tuesday awakenings). In either case, one minute after she wakes up on Monday, she is told that (*MON*) it is Monday and, if the coin lands tails, one minute after she wakes up on Tuesday, she is told that (*TUE*) it is Tuesday. In both cases, the experiment ends on Wednesday, on which she is immediately told that it is Wednesday.

To see why this is a paradox, think about two questions: (i) "What is her credence in *HEADS* when she wakes up on Monday?" and (ii) "What is her credence in *HEADS* when she is told that it is Monday?" David Lewis argues that the right answers to (i) and (ii) are 1/2 and 2/3. Elga himself argues that they are 1/3 and 1/2. Their arguments look equally plausible, and thus a paradox.

What are the arguments for those competing views? Let s be

her last conscious moment on Sunday, *m* be the moment of wakeup on Monday, and *m*+ be one minute later. Now, think about what credence SB ought to assign at *m* to *HEADS*. Lewis argues that it is ½: Note that *HEADS* is a proposition entirely about what the actual world is like; for, it is just the proposition that it is a *HEADS*-world. However, her total evidence at *m*, which is (*WAKEUP*) "SB wakes up today with the memory up to Sunday," doesn't seem to be new evidence about what the actual world is like. For the strongest proposition about the topic entailed by *WAKEUP* is that SB wakes up some day with the memory up to Sunday, which she fully believed on Sunday night. Since SB knew on that night that the coin was fair, her credence at *s* in *HEADS* was 1/2. Because only new and relevant evidence can change one's credence, her credence at *m* in *HEADS* is also ½. (See Lewis (2001), p. 175.)

By contrast, Elga thinks that SB's credence at *m* in *HEADS* is less than ½. If she knew at *m* that it was Monday, she would assign ½ to *HEADS*. For, if she knew at that moment that it was Monday, then *HEADS* and *TAILS* would be equally probable to her. If she were sure at *m* that it was Tuesday, she would assign zero to *HEADS*. This is because waking up on Tuesday is possible only when the coin lands tails. Clearly, her actual credence at *m* in *HEADS* is the weighted average of these two values, where the weights come from her credences at *m* in *MON* and in *TUE*. Since she is not sure at *m* that it is Monday, her credence at *m* in *HEADS* must be smaller than 1/2. (See Elga (2000), pp. 144-5.)

Hence, there are equally attractive but mutually inconsistent answers to (i). Since their conclusions are inconsistent, one of the above arguments is unsound. Suppose that you are on Elga's side concerning this issue. In that case, your problem is in explaining why it is rational for SB to change her credence in *HEADS* from *s* to *m* although *WAKEUP* is not new in the relevant sense. Of course, you can try to solve this problem by somehow arguing that *WAKEUP* is new and relevant. For example, Weinstraub argues that *WAKEUP* is new evidence at *m* because SB can derive from it that it is Monday or Tuesday, which she didn't believe at *s*. (See Weintraub(2004).) However, while it is persuasive that *WAKEUP* is new evidence about *what time it is*, she fails to explain why it is new evidence *about what the world is like*, especially, about whether this is a *HEADS*-world or a *TAILS*-world.

In this paper, I will criticize Lewis's argument. However, my approach will be different from Weintraub's. For I will not contend that *WAKEUP* is new evidence about *HEADS* versus *TAILS*. Instead, I will argue against Lewis's assumption that, if *WAKEUP* isn not new evidence about that topic, then SB's credence in *HEADS* should not change at *m*. For my argument, I will use Daniel Garber's solution of the so-called problem of old evidence. Perhaps, the best-known example of the old evidence problem is the confirmation of general relativity theory by the anomalous precession of Mercury's perihelion: Scientists changed their credence in the theory although the odd movement of the planet's perihelion was not new evidence about physical matters. I

consider this to be a counterexample of the principle that, if hypothesis *H* is about topic *T* and an agent has no new evidence about *T*, then she should not change her credence in *H*. Since Lewis's assumption is an instance of this principle, we have no reason to accept his argument.

## 2. Background 1: *De Nunc* Credences

In discussing the SB problem, we talked about SB's credence in *HEADS* and her credence in *WAKEUP*. First, what do we mean by "credence"? Second, do not these two credences belong to different types of credence? In this section, I will answer these questions.

First, what is credence? It is a type of subjective probability: Some probabilities are subjective in that they are (i) distinguished from objective probabilities and (ii) dependent upon the belief states of agents. For example, think about the probability that (*C*) Oswald had a confederate. Since it is now a matter of the past history, it is intuitive to say that *C*'s objective probability is 0 or 1. However, it is equally intuitive to say that *C* has some present probability *r* that is neither 0 nor 1; furthermore, different people assign different probabilities to *C*. So the probability of *C* is subjective, in the sense that it is a form of personal judgment. Then, what is the metaphysical nature of a subjective probability? The traditional answer is that it is the degree of somebody's belief. In the above example, *r* is the degree of an agent's belief in *C*.[1)]

Second, in the SB paradox, what's the difference between her beliefs in *HEADS* and in *WAKEUP*? They have different types of contents. On the one hand, her belief in *HEADS* is totally about what possible world she is in. For we can identify her belief in *HEADS* with the belief that this is a *HEADS*-world. On the other hand, her belief in *WAKEUP* is *not* totally about what kind of possible world this is. Yes, it is partially a belief that *this world* is a place in which SB wakes up with the memory up to Sunday. However, it is more than that. For it is also the belief that *today* is a day when she wakes up with the memory up to Sunday. Therefore, SB's belief in *WAKEUP* is not only about what kind of possible world this is but it is also about what time it is.

Consequently, the contents of the above beliefs, *HEADS* and *WAKEUP,* have different truth-conditions. The truth-value of *HEADS* will remain to be fixed regardless of time: if *HEADS* is true, then it will be always true and, if it is false, then it will be always false. By contrast, *WAKEUP* comes to have different truth-values as time passes: *WAKEUP* was false on Sunday but it is true on Monday.

Moreover, the contents of some beliefs have different truth-values relative to *individuals* as well as to *time*. Think about the belief that (*B*) I am pretty now. Its truth-value varies relative to individuals: *B* is true of Jane but false of Jessica. Also, its truth-value varies relative to time: *B* was false of Jane before her plastic surgery but true of her afterwards.

---

1) See Ramsey (1926) for the classic presentation of this view.

So, we can distinguish three types of propositions that a belief can have as a content: Genuine propositions have the same truth-values to everything at any time. Tensed propositions have the same truth-value to everything but may have different truth-values at different times. Centered propositions may have different truth-values to different things at different times. Correspondingly, we can identify three types of belief: A *de dicto* belief has a genuine proposition as the content. A *de nunc* belief has a tensed proposition as the content. A *de se* belief has a centered proposition as the content. Finally, we can distinguish three types of credences: A *de dicto* credence is the degree of a *de dicto* belief. A *de nunc* credence is the degree of a *de nunc* belief. A *de se* credence is the degree of a *de se* belief. Note that each of these distinctions is not meant to be exclusive. For a genuine proposition is a tensed proposition but not vice-versa and a tensed proposition is a centered proposition but not vice-versa. Similarly for the other distinctions.[2)]

Given these notions, we can see the characteristic feature of the SB paradox. SB's total evidence at *m*, *WAKEUP*, is *de nunc*

---

2) An anonymous referee complained that, first, when a credence combines *de dicto* and *de nunc* credences, it is unclear whether the combined credence is *de nunc* or *de dicto*, and, second, *de se* or *de nunc* credences are always reducible to *de dicto* credences. First, the combined credence will have different truth-values relative to different times, and so it will be *de nunc* but not *de dicto*. Thus I see no unclear point here. Second, not all *de se* or *de nunc* beliefs or credences are reducible to *de dicto* ones. For, if they were, it would be irrational to have a non-zero *de nunc* credence that today is not Wednesday because it would be reduced to the *de dicto* credence that Wednesday is not Wednesday.

evidence, but her credence in *HEADS* is *de dicto* credence. Then, the debate is over whether a piece of *de nunc* evidence can change one's *de dicto* credence, when the evidence doesn't seem to have new information about what kind of world this is. If Lewis's view is right, SB won't change her credence in *HEADS* by receiving *WAKEUP*. If Elga's view is correct, she will.

## 3. Background 2: the Problem of Old Evidence

Ever since the Sleeping Beauty problem was first introduced to philosophers, Elga's view has been the most popular view. However, it always suffered from one difficulty: *WAKEUP* doesn't seem to be new evidence about what kind of world this is. Later in this paper, I will suggest that Garber's solution of the latter problem provides a clue. But I first need to explain what the original problem of old evidence was and how Garber solved it.

Here are details of the problem related to Mercury's perihelio n:[3] **Example 1.** After Einstein had presented his general relativity theory (hereafter: *GRT*) to the Prussian Academy of Science in 1915, *GRT* was confirmed by the already known fact that (*P*) the perihelion of Mercury rotates around Sun.[4] The problem is that *P*

---

[3] See Glymour (1983) for the original presentation of the problem.

[4] According to Newton, a small object moving around a heavier object must have a *fixed* ellipse orbit in the absence of other such objects. In the solar system, a planet's orbit can change as a result of interaction with other planets, but the observed rotation of Mercury's perihelion was greater than is predicted by Newton's theory. The general relativity theory can explain this phenomenon in terms of gravitational wave produced by Mercury.

was old evidence in 1915 in that scientists had already fully believed in $P$. According to the classical Bayesianism, $P$ couldn't confirm $GRT$. For, it states that (i) the credence of the given agent satisfies three axioms of probability theory and (ii) evidence $E$ confirms a hypothesis $H$ exactly when the agent's credence in $H$ increases by strict conditionalization upon $E$. By (ii), $P$ could not increase scientists' credence in $GRT$ because $C_{1915}(GRT)=C_{\text{before } 1915}(GRT/P)=C_{\text{before } 1915}(GRT)$.

This problem, of course, can be generalized. Usually, Scientists develop a new theory only after they find out that the old theory isn't compatible with an *already known* phenomenon. Suppose that, under some uncontroversial facts, the new theory entails a known phenomenon. In such a case, the new theory must be confirmed by old evidence. However, as long as the evidence is old, the credence in the new theory cannot be increased by strict conditionalization. This is the so-called problem of old evidence.

Then, what was Garber's solution to this problem?[5] His view was that, when the old evidence appears to confirm the new theory, actually there is other evidence that is new, and it is this new evidence that confirms the theory. What is it? It is the discovery that the evidence is logically entailed by the theory (under some other known facts). Focus on **Example 1** again. Let $K$ be the totality of scientists' background beliefs before 1915; especially, $K$ includes observations relevant to Mercury's orbit, such as its mass and distance from Sun. Then, at some time *after*

---

[5] See Garber (1983). For related discussions, also see Jeffrey (1983) and Eells(1985).

Einstein had come up with *GRT*, scientists *newly learned* that *GRT* entails *P* under *K*.[6] Perhaps, this new evidence could increase their credence in *GRT*.[7]

Unfortunately, there is a problem. According to the classic form of Bayesianism, a rational agent's credences satisfy the axioms of probability including

(NORMALITY) If X is a logical truth, then C(X)=1,

where *C* is a rational believer's credence function. So, if scientists' belief systems satisfy NORMALITY, *GRT*'s entailment of *P* under *K* is just another piece of old evidence. For, *GRT*'s entailment of *P* under *K* is a logical truth and so scientists, if rational in the above sense, must have already fully believed it. They are supposed to be logically omniscient at any moment.

This seems to be a case where usually harmless idealization backfires. Note that even the best of us are not logically omniscient and we learn new logical truths. Especially, before Einstein developed *GRT*, scientists could not have known that *GRT* entails *P* under *K*; for, they had no idea about the theory.

---

6) Einstein pointed out the connection between *GRT* and *P* in his (1916).

7) A referee claims that it is impossible to receive *logical evidence*. Perhaps he meant that it is impossible for us to receive *perceptual* evidence about logical matters. However, as a matter of fact, much of my logical knowledge was acquired by *perceptually observing* logic textbooks. If he meant that some kind of *a priori* epistemic capability plays crucial role in the learning of logic, I do not see where our disagreement lies in. Therefore, I do not see why it is impossible to receive evidence that includes information about logical matters.

This means that we need a new form of Bayesianism in which NORMALITY was removed or replaced by a weaker condition.

Garber suggested so-called local Bayesianism as such a theory. According to it, the traditional model is *global* in the following sense: Credence function $C$, representing a scientist's opinion, maps the sentences of a *global* language $L$ into [0,1]. $L$ is global in that it can express propositions related to any problem of a scientific topic.

By contrast, Garber's new model is *local* in the following sense: Let us assume a language $L$ specialized for a scientific topic $T$.[8] Then, we extend it into another language $L^*$ in this way: (i) Add sentences expressing *roughly* implicative relation between $H$ and $E$ and between $H$ and $\sim E$ in $L$, where $H$ expresses a hypothesis about $T$ and $E$ a piece of possible evidence, and then (ii) add all the boolean combinations of the resulting sentences. (The roughness of (i) is intended by Garber, enabling a broad range of application but, for his purpose, the "implicative relation" can be (and usually is) interpreted as meaning logical entailment.) Let us write sentences involving such an implicative relation in the form of "$H \vdash_K E$" or "$H \vdash_K \sim E$", where $K$ is some background knowledge, and call them "implicative sentences." (For convenience, we will usually omit the quotations marks.) The trick is that, in Garber's model, the

---

[8] Originally, Garber focuses upon a language $L$ specialized for solving a *problem*. However, we can safely apply his discussion into to a language specialized for discussing a *topic*. For, the general question, "What are true about topic $T$?," can be regarded as a *problem*. A language specialized for solving this problem will also be a language specialized for discussing $T$.

agent can be *ignorant* of the truth-value of an implicative sentence "$H \vdash_K E$" even when "$\vdash_K$" is interpreted as meaning logical entailment under $K$. To see how this helps, let's introduce these definitions (following Garber, I assume that $L^*$ is an elementary sentential language;[9] also, I assume that all its truth-functional connectives are definable from "~" and "&"):

(1) An assignment $I$ of truth-values to sentences in $L^*$ is an interpretation of $L^*$ iff $I$ uniquely assigns a truth-value ($T$ or $F$) to every sentence in $L^*$.
(2) An interpretation $I$ of $L^*$ is consistent iff, for any sentence $X$ in $L^*$, $[I(\sim X)=T$ iff $I(X)=F]$ and $[I(X\&Y)=T$ iff $I(X)=T=I(Y)]$, where $I(\alpha)$ is the truth value assigned by $I$ to sentence $\alpha$.
(3) A sentence $X$ is a logical truth in $L^*$ iff, for any consistent interpretation $I$ of $L^*$, $I$ assigns $T$ to $X$.

Then, we let $C$ be a credence function having $L^*$'s sentences as its domain. Now, we can formulate a modified version of NORMALITY:

(LOCAL NORMALITY) If $X$ is a logical truth in $L^*$, then $C(X)=1$.

LOCAL NORMALITY is weaker than NORMALITY because, even when $H \vdash_K E$ is interpreted as a logical entailment sentence, it may not be a logical truth in $L^*$ as defined by (1)-(3) and so

---

[9] Eells makes it clear that this is only for the simplicity of discussion: "Of course, the choice of making $L$ and $L^*$ *truth-functional* languages is just an example. They could instead be first-order language, where sentences containing modal logical structure, second-order quantifiers, and so on." (See Eells(1985), p. 292.)

possibly $C(H \vdash_K E)<1$. However, it works just like NORMALITY *within* the probabilistic model consisting of $C$ and $L^*$. Garber's ingenuity is in this modest way of weakening of NORMALITY while preserving the axiom's theoretical benefits. Since the resulting system is a bit weak for Garber's purpose, he also introduces the following restriction upon the agent's credence distribution $C$:

$$(K^*) \quad C[X\&Y\&(X \vdash_K Y)]=C[X\&(X \vdash_K Y)].$$

As Eells points out, $K^*$ guarantees that, if the agent fully believes $X$ and $X \vdash_K Y$ respectively, then she also fully believes in $Y$. Thus, " $\vdash$ will behave somewhat like implication in classical Bayesianism" (Eells[1985] p. 290).

Let's see how this helps in **Example 1**. It allows the agent to assign a credence smaller than 1 to $GRT \vdash_K P$ interpreted as $GRT$'s logical entailment of $P$ under $K$. So, it is now possible that $C_{\text{before learning}}(GRT \vdash_K P)<1$ and $C_{\text{after learning}}(GRT)=C_{\text{before learning}}(GRT/GRT \vdash_K P)>C_{\text{before learning}}(GRT)$. But how does he show that this is more than a theoretical possibility? Look at this table:

| | $GRT$ | $GRT \vdash_K P$ | $GRT \vdash_K \sim P$ | Before learning $GRT \vdash_K P$ | After learning $GRT \vdash_K P$ |
|---|---|---|---|---|---|
| $I_1$ | T | T | F | $r_1$ | $r_1/(r_1+r_3+r_4)$ |
| $I_2$ | T | F | F | $r_2$ | 0 |
| $I_3$ | F | T | T | $r_3$ | $r_3/(r_1+r_3+r_4)$ |
| $I_4$ | F | T | F | $r_4$ | $r_4/(r_1+r_3+r_4)$ |
| $I_5$ | F | F | T | $r_5$ | 0 |
| $I_6$ | F | F | F | $r_6$ | 0 |

For brevity, this table does not include any rows assigning F to $P$ and the column representing the truth-values of $P$. (Remember that the scientists fully believe(d) $P$ before and after they learn that $GRT \vdash_K P$.) Also, $I_1$-$I_6$ include only interpretations whose credences can be positive under Garber's $K^*$ restriction. After learning $GRT \vdash_K P$, scientists assign 0 to $I_2$, $I_5$ and $I_6$. By strict conditionalization, $C_{\text{before 1915}}(GRT)=r_1+r_2$ and $C_{\text{before 1915}}(GRT/GRT \vdash_K P) = r_1/(r_1+r_3+r_4)$. Hence, $C_{\text{before 1915}}(GRT)<C_{1915}(GRT)$ exactly when $r_1+r_2<r_1/(r_1+r_3+r_4)$. This condition is easily satisfied. For example, let $r_1=0.3$, $r_2=0.1$, $r_3=0.1$, $r_4=0.1$, $r_5=0.2$, and $r_6=0.2$; then, $r_1+r_2=0.4< 0.6=r_1/(r_1+r_3+r_4)$. This establishes that $GRT \vdash_K P$ *can* confirm $GRT$, depending upon the previous credences.

Here is a general lesson: For a similar case, let us call a piece of evidence $E$ which belongs to language $L$ (about topic $T$) "intrasystematic evidence" and evidence in the form of $H \vdash_K E$ "extrasystematic evidence." Then, Garber is committed to the claim that, even when an agent's total intrasystematic evidence is old about $T$, the extrasystematic evidence may be new about the topic of implicative relations between the hypothesis and the intrasystematic evidence. In such a case, the extrasystematic evidence can legitimately update the agent's credence about $T$.

This is Garber's solution of the problem of old evidence, which is quite much the standard view at present. Now, I point out some important features of his view and ask related questions: First, Garber focuses upon the evidential role of learning a hypothesis $H$'s entailment of the intrasystematic evidence $E_k$. But what if the agent learns something new about $E$'s entailment of $H$?

Second, Garber discusses only cases in which extrasystematic evidence takes a very simple logical form. However, what if the extrasystematic evidence has a more complicated logical structure, such as disjunction?

Third, the model of local Bayesianism includes LOCAL NORMALITY instead of NORMALITY, which allows a rational agent to *newly* learn $H \vdash_K E$ despite its being a logical truth. However, what if we interpret a sentence including "$\vdash_K$" as not expressing logical entailment (under $K$) but as expressing some contingent but still implicative relation between $H$ and $E$?

I answer these questions all at once: there is a case in which a rational agent receives some (i) *contingent*, (ii) *disjunctive* extrasystematic evidence about some implicative relation (iii) *from evidence to the hypotheses*. Where can we find such a case? In the Sleeping Beauty problem!

## 4. Strategy

Let us call the topic of what kind of world this is "*de dicto* matters" and that of what time it is "*de nunc* matters." Then, here is the difficulty of Elga's view: how can *WAKEUP*, SB's evidence at *m*, change her credence in *HEADS*, although *WAKEUP* is not new evidence about *de dicto* matters but *HEADS* is a hypothesis entirely about *de dicto* matters?

Look into this problem more carefully. It presupposes that *HEADS* is a hypothesis about *de dicto* matters and *WAKEUP* is old evidence about *de dicto* matters. This looks plausible: *HEADS*

is a proposition about what world this is; for, it is the proposition that this is a *HEADS*-world. *WAKEUP* seems to be old evidence about *de dicto* matters; for, *HEADS* and *TAILS* are the two possibilities about *de dicto* matters open to SB before *m* but *WAKEUP* removes neither.

However, even if this is true, why is it a problem that *WAKEUP* changes SB's credence in *HEADS*? Consider this very intuitive principle:

> (PRESERVATION) If hypothesis $H$ is about topic $T$ but the agent's total present evidence $E$ is not new evidence about $T$, then her credence in $H$ should not change.

Remember that *HEADS* is a hypothesis about *de dicto* matters but *WAKEUP* is not new about *de dicto* matters. Thus, it follows from PRESERVATION that SB's credence at m in *HEADS* must be the same as her previous credence, ½. This is exactly Lewis's argument.

However, PRESERVATION has a counterexample. Think about **Example 1**. According to Garber, $GRT \vdash_K P$ is extrasystematic evidence and scientists change their credence in *GRT* on the basis of this evidence. In that example, *GRT* is clearly a hypothesis about the topic of *physical matters* but $GRT \vdash_K P$ is not new evidence about any empirical matter. Rather, it is new evidence about a *logical matter*, namely, *GRT*'s entailment of $P$ under $K$.

If PRESERVATION is false, we have no reason to believe that *WAKEUP* can never change the credence in *HEADS*. Despite its intuitiveness, it is possible that an agent having only old about

about *T* can change her credence in a hypothesis about *T*. In principle, this can solve the problem of old evidence that appears in the SB problem.

Nevertheless, the devil is in the details. Here, the question is "How can we build a concrete model applicable to the SB problem from Garber's so-called local Bayesianism?" Without this question being answered, it will not be so persuasive to just point out the theoretical *possibility*. This is more so when we consider this fact: at first sight, the SB problem appears to have nothing to do with Garber's ideas, because it is unclear that SB at *m* has any evidence about an implicative relation between her evidence *WAKEUP* and hypothesis *HEADS*.

But note that she has new evidence about *de nunc* matters; for, she newly learns at *m* that it is Monday or Tuesday, not Sunday as she believed before. Surprisingly, this happens to be equivalent with a form of extrasystematic evidence, evidence about a roughly implicative relation between *WAKEUP* and *HEADS*. Here I sketch an argument:

> Let "*K*" name SB's total background knowledge at *m* and "*T*" abbreviate "the (not yet known) truth about what day it is between Monday and Tuesday". Next, define $Ex_{\mathrm{MON}}$ as the claim that *WAKEUP* logically entails neither of *HEADS* and *TAILS*, and $Ex_{\mathrm{TUE}}$ as the claim that *WAKEUP* entails *TAILS*, respectively under *K* combined with *T*.
>
> (P1) SB receives *MON*v*TUE* as evidence at *m*.
>
> (P2) To SB at *m*, $Ex_{\mathrm{MON}}$v$Ex_{\mathrm{TUE}}$ is equivalent with *MON*v*TUE*.
>
> (P3) If received as evidence at *m*, $Ex_{\mathrm{MON}}$v$Ex_{\mathrm{TUE}}$ will be (i) contingent (ii) disjunctive evidence about an implicative relation (iii) from *WAKEUP* to *HEADS*.
>
> (P4) If P1-P3 are true, SB has some extrasystematic evidence at *m*.
>
> (C) Therefore, SB has some extrasysematic evidence at *m*.

If this argument is sound, we have a potential explanation of how SB's credence changes from *s* to *m*: At *m*, SB has intrasystematic evidence *WAKEUP* and extrasystematic evidence $Ex_{\text{MON}} \vee Ex_{\text{TUE}}$. Although *WAKEUP* is not new evidence about *de dicto* matters, she has $Ex_{\text{MON}} \vee Ex_{\text{TUE}}$, extrasystematic evidence about an implicative relation between *WAKEUP* and *HEADS*. Perhaps, just as $GRT \vdash_K P$ can change scientists' credence in *GRT* despite *P* not being new evidence about physical matters, $Ex_{\text{MON}} \vee Ex_{\text{TUE}}$ may change SB's credence in *HEADS* despite *WAKEUP* not being new evidence about *de dicto* matters.

Still, it is not perfectly clear that, when she receives $Ex_{\text{MON}} \vee Ex_{\text{TUE}}$, SB will change her credence in *HEADS* somehow similarly to how scientists changed their credence in *GRT*. First, (i)-(iii) in P3 constitute important differences between $Ex_{MON} \vee Ex_{TUE}$ and $GRT \vdash_K P$. Second, even if those differences turn out to be inessential, it only means that we *can* apply a Garberian approach to the SB paradox, not that it leads to Elga's view about the SB problem.

To solve these problems, I will proceed in this order: In Section 5, I will contend that, if an agent receives a piece of *disjunctive* evidence about the logical entailment *from evidence to a hypothesis*, it can still be a form of extrasystematic evidence. In Section 6, I will claim that *contingent* extrasystematic evidence is possible and that Garber's approach is applicable to a logically omniscient agent with respect to such evidence. In Section 7, I will discuss several principles about the newness and oldness about a certain topic. In Section 8, I will suggest a broadly

Garberian approach to the SB problem. As a result, Lewis's argument will shown to be unsound. In Section 9, I will admit that a similar criticism is possible for Elga's argument when interpreted in a certain way, but I will argue that there exists an alternative interpretation which avoids the criticism

## 5. Disjunctive Extrasystematic Evidence about Evidence-to-Hypothesis Entailment

In this section, I will discuss the possibility of disjunctive evidence about the entailment from intrasytstematic evidence to a hypothesis. Such evidence can be regarded as extrasystematic, because the agent receiving it can rationally change her credence in the hypothesis despite the intrasystematic evidence's oldness about the given topic.

Consider this example: **Example 2**. In John's house, Ann is having a dinner with John. She is thinking about whether (*CD*) the wine will be chardonnay or (*PN*) it will be pinot noir. Superficially, her total evidence is that ($STEAK_A$) steak is Ann's favorite, which was previously fully believed and so is old evidence. The set $K_A$ of her background beliefs include these sentences (in addition to $STEAK_A$): "*CD* or *PN* but not both", "The host, John, never chooses a wrong wine for the entrée", "The entrée would be a steak or a fish", "Chardonnay is a good choice for a fish but a wrong choice for a steak", and "Pinot noir is good for both steak and fish", "If John remembers his only girlfriend's birthday, he will serve her favorite", and "John

remembers Ann's birthday". Note that neither *CD* nor *PN* follows from $STEAK_A$ under $K_A$, which doesn't include "Ann is John's only girlfriend". Nevertheless, Ann, who received C in Intro Logic 101, is not sure of this. After a few minutes of thought, she gets confident that either ($Ex_1$) "$STEAK_A$ entails neither *CD* nor *PN* under $K_A$" is true or ($Ex_2$) "$STEAK_A$ entails *PN* under $K_A$" is true but not sure of which. Assuming that she makes no other logical mistakes, what is Ann's new credence in *CD*?

Depending upon her previous credence distribution, her credence in *CD* can decrease. Formally, let $L_{\text{ANN'S DINNER}}$ be the minimal language closed under Boolean algebra which includes sentences *CD*, *PN*, $STEAK_A$, and the sentences in $K_A$. Let $L_{\text{ANN'S DINNER}}$* be the minimal extension of $L_{\text{ANN'S DINNER}}$ also closed under Boolean algebra which includes $STEAK_A \vdash_{KA} CD$ and $STEAK_A \vdash_{KA} PN$, where "$\vdash_{KA}$" is interpreted as logical entailment under $K_A$. As a result, $L_{\text{ANN'S DINNER}}$* includes $Ex_1 \vee Ex_2$, or more formally $[\sim(STEAK_A \vdash_{KA} CD,PN)] \vee [STEAK_A \vdash_{KA} PN]$.[10] We assume that LOCAL NORMALITY holds in Ann's credence function with respect to $L_{\text{ANN'S DINNER}}$* but NORMALITY doesn't. So, Ann can be not fully confident of (correct) $Ex_1$. Also, we assume that Ann's credence function obeys Garber's *K** restriction. Hence, she doesn't assign a positive credence to an interpretation assigning T to all of *CD*, $STEAK_A$, and $STEAK_A \vdash_{KA} PN$. Now, check this table:

---

[10] $\sim(STEAK_A \vdash_{KA} CD,PN)$ is the abbreviation of $[\sim(STEAK_A \vdash_{KA} CD)] \& [\sim(STEAK_A \vdash_{KA} PN)]$. Similarly for later examples.

| | $CD$ | $STEAK_{\mathrm{A}} \vdash_{KA} CD$ | $STEAK_{\mathrm{A}} \vdash_{KA} PN$ | Before learning $Ex_1 \mathrm{v} Ex_2$ | After learning $Ex_1 \mathrm{v} Ex_2$ |
|---|---|---|---|---|---|
| $I_1$ | T | T | F | $r_1$ | 0 |
| $I_2$ | T | F | F | $r_2$ | $r_2/(r_2+r_3+r_4)$ |
| $I_3$ | F | F | T | $r_3$ | $r_3/(r_2+r_3+r_4)$ |
| $I_4$ | F | F | F | $r_4$ | $r_4/(r_2+r_3+r_4)$ |

Again, the table includes only $I_1$-$I_4$ , the interpretations whose credences can be positive under the restriction of $K^*$. If Ann is a Strict Conditionalizer, her credence in $CD$ changes just when $r_1+r_2 > r_2/(r_2+r_3+r_4)$ or $r_1+r_2 < r_2/(r_2+r_3+r_4)$. This is so when $r_1=r_2=r_3=r_4=0.25$. (In this particular case, $C_{\text{after learning}}(CD)=1/3<1/2=C_{\text{before learning}}(CD)$.) Hence, receiving $Ex_1 \mathrm{v} Ex_2$ changes Ann's credence in $CD$ despite the oldness of $STEAK_{\mathrm{A}}$ in *some* case.

Obviously, $Ex_1 \mathrm{v} Ex_2$ is disjunctive and it is about the entailment relation (under $K_{\mathrm{A}}$) from evidence to a hypothesis. Since it can change the credence in the hypothesis just as $GRT \vdash_K P$ could **in Example 1**. I conclude that disjunctive extrasystematic evidence is possible and it can change the credence in a hypothesis even when the total intrasystematic evidence is old.

However, this seems restricted to cases in which the agent's intelligence is less than perfect. Is there any type of extrasystematic evidence such that, when she receives it, even a logically omniscient agent will change her credence despite the oldness of her intrasystematic evidence?

## 6. Contingent Extrasystematic Evidence

In his discussion, Garber doesn't provide a specific interpretation of an implicative sentence (or a sentence including a turnstile). Nevertheless, it is natural in most of his works to interpret extrasystematic evidence as expressing logical entailment between a hypothesis and evidence. This makes it difficult to apply his view to the SB problem, in which the agent is assumed to be "a paradigm of probabilistic rationality" (Lewis (2001)). To remove this obstacle, I will discuss the possibility of contingent extrasystematic evidence.

When we write "$H \vdash_K E$", it expresses a tertiary relation among a hypothesis, evidence, and background knowledge. Admittedly, this relation is logical and so the whole sentence expresses a necessary truth *once the referents of* "$H_i$""$E_i$", *and* "$K$" *are fixed*. However, what if a non-rigid expression is occupying one of the three places? For an example, read Richard Jeffrey's discussion below, in which he conjectures that Newton had an odd form of extrasystematic evidence for his gravity theory:

> ⋯ where E reports the facts about the tides that Newton explained, it seems correct to say that his explanation gave them the status of evidence supporting his explanatory hypotheses, H ⋯ I suppose that he hoped to be able to show that
>
> (T) H implies the true member of E,
>
> where H was his theory (together with auxiliary data) and E was a set of mutually exclusive propositions, the members of which make various claims about the tides, and one of which is true. (Jeffrey[1983b], p. 148.)

Here, Jeffrey is claiming that Newton could confirm his gravity theory $H$ (H above minus the auxiliary data $K$) with extrasystematic evidence "$H \vdash_K$the true member of E" even though he didn't know which of E is the relevant intrasystematic evidence about tides.[11] Here, what interests me is not his conclusion but T, the example sentence. Note that Jeffrey doesn't demand that $H$ should entail every member of E, only that it should entail *some* true member of E. Hence, E can be $\{E,F\}$ such that $H$ entails $E$ but $H$ doesn't entail $F$, respectively under $K$. In that case, the truth of $T$ depends upon which of $E$ and $F$ is the truth about tides. This is an empirical, contingent matter. Therefore, T is an example of *contingent extrasystematic evidence*.

Note that the contingency of T is due to the description in it with a contingent satisfaction condition. My idea is that an implicative sentence may include such a description in the position of the background knowledge. Hence, consider a sentence $Ex_T$ in the form of "$E_k \vdash_{K \cup \{\text{the true member of } \mathbf{E}\}} H_k$". Suppose that E is $\{T_1,T_2\}$ such that $E_k \vdash_{K \cup \{T1\}} H_k$ but $\sim(E_k \vdash_{K \cup \{T2\}} H_k)$ and that it is a contingent matter which of $T_1$ and $T_2$ is true. Then, $Ex_T$ is contingent. Consequently, a logically omniscient agent may not know its truth-value.

This allows us to apply Garber's approach to a logically omniscient agent. Think about this example: **Example 3**. In the dinner table of John's house, Beth, a logically omniscient

---

11) Jeffrey claims that Newton was in a position to know this implicative sentence's truth without necessarily knowing which of is entailed by $H$ under $K$. Also, he attributes this claim to David Lewis without footnoting the reference. Perhaps, from an oral conversation? See Jeffrey (1983) p. 149.

professor of Advanced Logic 505, is thinking about whether *CD* or *PN*. Her evidence includes the fact that ($STEAK_B$) steak is Beth's favorite, which is old news to her. Her background knowledge $K_B$ includes all propositions expressed by the sentences in $K_A$ after the replacement of "Ann" with "Beth".[12] Now, John asks her to be his girlfriend and Beth agrees. So, she is sure that (*HISGIRL*) Beth is John's girlfriend. This is logically equivalent with this claim: (*OTHERS*) John has many girlfriends, one of whom is Beth, or (*ONLYME*) Beth is John's only girlfriend. In this case, what's her new credence in *CD*?

To answer, we define define $Ex_{OTHERS}$ and $Ex_{ONLYME}$ as follows:

($Ex_{OTHERS}$) $STEAK_B$ entails neither *CD* nor *PN* under $K_B$ and whichever is true between *ONLYME* and *OTHERS*.

($Ex_{ONLYME}$) $STEAK_B$ entails *PN* under $K_B$ and whichever is true between *ONLYME* and *OTHERS*.

I want to point out three facts. First, $Ex_{OTHERS} v Ex_{ONLYME}$ is obviously disjunctive and describes facts about a broadly implicative relation between $STEAK_B$ and *ONLYME*/*OTHERS*. Second, $Ex_{OTHERS} v Ex_{ONLYME}$ is equivalent to *OTHERSvONLYME*. For, it is easy to show the equivalence between *OTHERS* and $Ex_{OTHERS}$ and that between *ONLYME* and $Ex_{ONLYME}$.[13] Third,

---

12) We are allowed to use propositions as the domain of credence function because, given Beth's logical omniscience, we don't have to replace NORMALITY with LOCAL NORMALITY, which was incompatible with proposition as the bearer of credence.

13) It suffices to show that (i) *OTHERS* entails $Ex_{OTHERS}$, (ii) ~*OTHERS* entails

$Ex_{\text{OTHERS}} \vee Ex_{\text{ONLYME}}$ is a contingent proposition. If *HISGIRL* is false, then nothing satisfies "whichever is true between *ONLEME* and *OTHERS*", in which case neither disjunct is true. If *HISGIRL* is true, one of the disjuncts is true because [$STEAK_{\text{B}}$ entails *PN* under $K_{\text{B}}$ combined with *ONLYME*] and [$STEAK_{\text{B}}$ entails neither *CD* nor *PN* under $K_{\text{B}}$ combined with *OTHERS*]. Since Beth's being John's girlfriend is a contingent matter, so is $Ex_{\text{ONLYME}} \vee Ex_{\text{OTHERS}}$.

Since $Ex_{\text{ONLYME}} \vee Ex_{\text{OTHERS}}$ is equivalent to *HISGIRL*, we can consider Beth to be receiving the former as evidence. Then, what should be her resulting credence? Let $\mathsf{E}$ be {*ONLYME*, *OTHERS*} and "*T*" be the abbreviation of "the true member of $\mathsf{E}$". Let $S_{\text{BETH'S DINNER}}$ be the minimal class of propositions closed under Boolean operation that includes *CD*, *PN*, $STEAK_{\text{B}}$, and the propositions in $K_{\text{B}}$ and $\mathsf{E}$. So, we can formally express $Ex_{\text{ONLYME}}$ as $STEAK_{\text{B}} \vdash_{KB \cup \{T\}} PN$ and $Ex_{\text{OTHERS}}$ as $\sim(STEAK_{\text{B}} \vdash_{KB \cup \{T\}} CD,PN)$. Let $S_{\text{BETH'S DINNER}}$* be the minimal Boolean-operation-closed superset of $S_{\text{BETH'S DINNER}}$ which also includes $Ex_{\text{ONLYME}}$ and $Ex_{\text{OTHERS}}$. Since Beth is logically omniscient, we assume

---

$\sim Ex_{\text{OTHERS}}$, (iii) *ONLYME* entails $Ex_{\text{ONLYME}}$, and (iv) $\sim$*ONLYME* entails $\sim Ex_{\text{ONLYME}}$. For (i), suppose that *OTHERS* is true. Hence, it is *OTHERS* that is true between *ONLYME* and *OTHERS*. Since neither *CD* nor *PN* follows from $STEAK_{\text{B}}$ under $K_{\text{B}}$ and *OTHERS*, $Ex_{\text{OTHERS}}$ is true. For (ii), suppose that *OTHERS* is false. Since either *ONLYME* or *OTHERS*, *ONLYME* is true. Then, $Ex_{\text{ONLYME}}$ follows by (iii), which will be shown below. Since $Ex_{\text{OTHERS}}$ is incompatible with $Ex_{\text{ONLYME}}$, $Ex_{\text{OTHERS}}$ is false. For (iii), the proof is similar to that of (i). (Use $STEAK_{\text{B}}$'s entailment of *PN* under $K_{\text{B}}$ and *ONLYME*.) For (iv), the proof is similar to that of (ii). (Use (i).) Note: These proofs are not circularly dependent.

NORMALITY, not LOCAL NORMALITY. Then, her credence distributions look like this, before and after learning *HISGIRL* or, equivalently, $Ex_{\text{ONLYME}} \vee Ex_{\text{OTHERS}}$:

| | *CD* | *HISGIRL* | $STEAK_B \vdash_{KB \cup \{T\}} CD$ | $STEAK_B \vdash_{KB \cup \{T\}} PN$ | Before learning *HISGIRL* | After learning *HISGIRL* |
|---|---|---|---|---|---|---|
| $I_1$ | T | F | T | F | $r_1$ | 0 |
| $I_2$ | T | T | F | F | $r_2$ | $r_2/(r_2+r_3+r_4)$ |
| $I_3$ | F | T | F | T | $r_3$ | $r_3/(r_2+r_3+r_4)$ |
| $I_4$ | F | T | F | F | $r_4$ | $r_4/(r_2+r_3+r_4)$ |

Except that it includes a column representing *HISGIRL*'s truth-value, this table almost looks like that in **Example 2**. Similarly to in that example, Beth's credence in *CD* changes when $r_1+r_2 > r_2/(r_2+r_3+r_4)$ or $r_1+r_2 < r_2/(r_2+r_3+r_4)$. This condition is satisfied when $r_1=r_2=r_3=r_4=0.25$.

Let me ask a question: "To explain Beth's decreasing credence in *CD*, must we use Garber's approach?" Perhaps not. For, we can explain the decrease as a result of conditioning upon *HISGIRL*, not upon $Ex_{\text{ONLYME}} \vee Ex_{\text{OTHERS}}$. Still, it is equally true that we *can* explain the decrease as a result of conditioning upon the disjunctive evidence apparently about a broadly implicative relation between evidence $STEAK_B$ and hypothesis *CD*. This is enough to show that Garber's approach can be adopted for similar cases.

At this point, some readers may complain. Sure, the disjunctive evidence describes a fact about an implicative, quasi-logical relation between $STEAK_B$ and *CD*/*PN*. Even so, the concept of

contingent extrasystematic evidence may be too exotic for some people to swallow. They may insist on restricting the notion of extrasystematic evidence to information about logical relations. However, as I wrote earlier, Garber left the correct interpretation of an implicative sentence to be a widely open matter. Furthermore, we already saw that a well-known philosopher introduced *contingent* extrasystematic evidence to his discussion.[14)]

Therefore, if we allow extrasystematic evidence to have a slightly different logical structure, we can apply a broadly Garberian approach to a logically omniscient agent. Especially, such evidence can be *contingent*, *disjunctive*, and about a relation *from evidence to a hypothesis*. I will argue that, waking up on Monday, SB receives this type of evidence.

## 7. Principles of Newness and Oldness

In this section, I will first discuss two principles which describe when one is allowed to assign an (un)conditional credence that differs from the previous unconditional credence. However, as I will explain soon, these principles can be rationally violated in the presence of new and relevant extrasystematic evidence. Because I need similar principles that apply even when an agent receives such evidence, I will generalize those principles

---

14) David Lewis and John Etchemendy seem also committed to such a contingent type of extrasystematic evidence. For, Jeffrey cites them in claiming that Newton was in the position to know the truth of T without knowing the referent of “the true member of E”. See Jeffrey (1983), pp. 149-150.

for such cases.

To begin, I present two principles:

> (PRESERVATION) If hypothesis $H$ is about topic $T$ but total evidence $E$ is old about $T$, then a rational agent does not change her credence in $H$.
> (CONDITIONAL PRESERVATION) If hypothesis $H$ is purely about topic $T$ but neither total evidence $E$ includes new information about $T$ nor does a condition $F$, then a rational agent's present conditional credence in $H$ on $F$ is the same as her previous credence in $H$.

Basically, PRESERVATION means that no evidence can influence one's credence more than once. Similarly, CONDITIONAL PRESERVATION says that if some evidence was already received, then conditioning on it alone cannot give you a different value from the agent's previous unconditional credence.

Here is an explanation by Ellery Eells of why evidence usually confirms a hypothesis only once:

> One of the central tenets of Bayesian confirmation theory is that confirmation is a relation between three things: a piece of evidence, a hypothesis or a theory, *and a set of background beliefs*. As background beliefs change over time ⋯ *so does what confirms what* ⋯ it is quite natural to say simply that, because of the change in our background beliefs, $E$ simply does not confirm $T$ at the later time, after its evidential impact on $T$ has already been "absorbed." (See Eells (1985), p. 286.)

In this passage, Eells is making the following claims: When an agent receives a piece of evidence, two events usually occur. First, the evidence increases or decreases her credence in any

hypothesis to which it is relevant with its optimal evidential impact. Second, the evidence becomes a member of the set of the agent's background beliefs. If these two events always occur together, a piece of evidence's being one of the background beliefs will always mean that its optimal evidential impact upon the hypothesis has been already *absorbed* to the credences in the hypotheses. Hence, if the agent receives the same evidence again, it will not confirm or disconfirm any of the hypotheses whose credences it altered in the past.

I think this is a very natural picture of Bayesian confirmation. Importantly, it explains why scientists had to violate PRESERVATION in updating their credence in *GRT*. One may think that, when scientists first observed *P* (a strange movement of Mercury's perihelion) in the 19$^{th}$ century, it should have influenced their credence in *GRT*. However, it did not, because they had no idea about *GRT*. Needless to mention, they did not have any degree of belief in it. This means that, although *P* entered into the set *K* of their background beliefs, *P*'s evidential impact upon *GRT* was never absorbed into their credence in *GRT*. Furthermore, even when Einstein finally invented *GRT*, scientists must have been initially unaware that *GRT* entails *P* under *K*, which means that *P* played no role in setting their initial credence in *GRT*. Hence, when they later learned $GRT \vdash_K P$, it was actually epistemologically obligatory for them to change their credence in *GRT*.

In addition, Eells's view suggests that CONDITIONAL PRESERVATION can be also rationally violated under some

conditions. For the sake of explanation, I assume Ramsey's well-known thesis:

> If two people are arguing "If *p* will *q*?" and are both in doubt as to *p*, they are adding *p* hypothetically to their stock of knowledge and are arguing on that basis about *q*; so that in a sense "If *p, q*" and "If *p, ~q*" are contradictories. (See Ramsey(1929), p. 155.)

This thesis can be easily adopted for conditional probabilities: Suppose that an agent is trying to figure out the probability of *X* conditional on *F*. Then, according to the probabilistic version of Ramsey's thesis, the agent can calculate its value by first adding *F* hypothetically to her stock of knowledge and then judging on that basis about *X*'s probability.

If we accept this thesis, it is easy to explain why CONDITIONAL PRESERVA-TION is usually a rational norm of credal updating. Suppose that *F* was already known information about topic *T* and so a member of the agent's belief set. Then, even if she adds *F* to the stock of knowledge and then judge on that basis, the agent's thus judged subjective probability of *X* cannot differ from the one she assigned to *X* before. For the evidential impact of *F* was already incorporated to the agent's credal opinion. Moreover, this explains why it can be sometimes rational to violate CONDITIONAL PRESERVATION without any new evidence about *T* if *F* contains new information about an implicative relation between *H* and some old evidence *E* about *T*. For, if such *F* is true and includes such extrasystematic information, it means that *E*'s evidential impact about *T* was not

properly incorporated into the agent's probabilistic opinion about *H*.

Given these diagnoses, I first suggest a new version of PRESERVATION. Consider a language *L* containing various hypotheses *H*s and many possible pieces of evidence or conditions *E*s. Next, extend *L* into *L** by first adding various sentences *Ex* about whether a specific type of implicative relation holds between some *H* and *E* and then making the resulting language closed under Boolean combination.[15] We assume that the agent's total evidence is analyzable to *E*&*Ex*, where *E* belongs to *L* and *Ex* belongs to *L** but not to *L*. First, I suggest extending the notions of newness and oldness about a topic in this way:

(4) *E*&*Ex* is *locally new* about topic *T* iff [*E* is new about *T*] or [*Ex* is new about an implicative relation between *E** and *H*, for some sentence *E** and some hypothesis *H* about *T* that *Ex* mentions].

(5) *E*&*Ex* is *locally old* about topic *T* iff [*E* is old about *T*] and [*Ex* is old about the implicative relation between *E** and *H*, for any sentence *E** and any hypothesis *H* about *T* that *Ex* mentions].

Let me clarify what I mean by "*Ex* is new about ⋯" and by "*Ex* is old about ⋯" By the first clause, I mean that *Ex* includes some information about an implicative relation between *E** and *H*

15) If the agent is assumed to be logically omniscient, then we can use sets *S* and *S** of propositions, where *S* is extended to *S** as *L* is extended to *L** above. Of course, the agent may receive evidence about various types of implicative relations between evidence and hypotheses (remember that the turnstile can be interpreted in various ways) but, for simplicity, we assume that *L** includes only implicative sentences about just one type of such a relation.

that she didn't fully believe to hold between them. By the second clause, I mean that *Ex* may or may not include some information about an implicative relation between *E** and *H* but, even if it does, the agent already knew such a connection between them. So, the agent must have already updated her credence in *H* accordingly, when she noticed the connection. Given the above definitions, I suggest that the following principle is plausible:

> (GENERAL PRESERVATION) If *H* is a hypothesis about topic *T* but the agent's total evidence *E*&*Ex* is locally old about *T*, then $C_{pres}(H)=C_{prev}(H)$, (where $C_{pres}$ is the agent's present credence function and $C_{prev}$ the previous one.)

Suppose that the antecedent is true. By the definition of local oldness, *E* includes only old information about *T* and *Ex* includes at best old information about implicative relations. Then, I just cannot see how the agent can rationally change her credence in *H*. Also, note that this principle is invulnerable to a counterexample like **Example 1**. For, even though the total intrasystematic evidence *P* was old when Einstein published *GRT*, $P\&(GRT \vdash_K P)$ was not locally old about physical matters.

Next, I suggest a generalized version of CONDITIONAL PRESERVATION:

> (GENERAL CONDITIONAL PRESERVATION) If *H* is a hypothesis about topic *T* but neither the agent's present total evidence *E*&*Ex* is locally new about *T* nor is a condition *F*&*Fx*, then $C_{pres}(H/F\&Fx)=C_{prev}(H)$.

This means that, in the presence of total evidence *E*&*Ex*, the

agent's present credence in $H$, conditional on $F\&Fx$, should be the same as her previous credence in $H$ if neither $E\&Ex$ is locally old about $T$ nor is $F\&Fx$. I consider this to be a natural expansion of GENERAL PRESERVATION. To see why, we can appeal to Ramsey's thesis again: Assume that an agent is trying to figure out the probability of $X$ given $F\&Fx$. Thus she adds $F\&Fx$ and judges the probability of $X$. Neither her total evidence $E\&Ex$ includes any information directly about $T$ or any relevant extrasystematic information, nor does $F\&Fx$. Hence, she has no reason to assign to $X$ a different value from her previous credence in $X$. This means that her present credence in $X$ given $F\&Fx$ has the same value as her previous credence in $X$.

In this section, we discussed four principles which govern credal updating. The first two principles do not apply to a case in which the agent receives extrasystematic evidence. Thus it will not be surprising that they produce a contradiction when misused for such a case. For such a case, we need to use the last two principles instead.

## 8. Why are the Halfers Wrong?

Finally, it is the time to discuss the Sleeping Beauty problem. For this discussion, I ask two questions: First, when she wakes up on Monday, does SB receive any kind of extrasystematic evidence? Second, if she does, how does it affect the Halfers' argument that her credence in *HEADS* does not change from Sunday to Monday?

To start, I argue that SB has a form of extrasystematic evidence when she wakes up on Monday. Here is the argument that I sketched in Section 4:

(P1) SB receives $MON \vee TUE$ as evidence at $m$.
(P2) To SB at $m$, $Ex_{MON} \vee Ex_{TUE}$ is equivalent with $MON \vee TUE$.[16]
(P3) If received as evidence at $m$, $Ex_{MON} \vee Ex_{TUE}$ will be (i) contingent (ii) disjunctive evidence about an implicative relation (iii) from *WAKEUP* to *HEADS*.
(P4) If P1-P3 are true, SB has some extrasystematic evidence at $m$.
(C) Therefore, SB has some extrasysematic evidence at $m$.

Since this argument is obviously valid, it suffices to defend the premises.

Before this defense, we need to clarify some important points about $Ex_{MON}$ and $Ex_{TUE}$. Remember that they are defined as follows:

($Ex_{MON}$) *WAKEUP* entails neither *HEADS* nor *TAILS*, under SB's background beliefs $K$ and the true one between

16) A referee argues that, if $MON \vee TUE$ is old evidence (i.e., not new evidence), then $Ex_{MON} \vee Ex_{TUE}$ should also be old given the equivalence. So the latter cannot result in a credal change that I defend in this paper, the referee says. But the whole point of this paper is that, while $Ex_{MON} \vee Ex_{TUE}$ is *not* new evidence about *de dicto* matters, it includes new information about some kind of implicative relation between *WAKEUP* and *HEADS*, and so can result in such a credal change. Since I do not claim that $Ex_{MON} \vee Ex_{TUE}$ *is* new evidence about *de dicto* matters, the referee fails to disagree with me at that point. Because I have constructed model of a credal change without new evidence about *de dicto* matters, she also fails to convince me that such a change is impossible without such evidence.

*MON* and *TUE*.

($Ex_{TUE}$) *WAKEUP* entails *TAILS*, under SB's background beliefs *K* and the true one between *MON* and *TUE*.[17]

One important point is that we can consider "entails ⋯ under *K* and the true one between *MON* and *TUE*" as expressing a broadly implicative relation between evidence and a hypothesis.[18] Another important point is the truth-conditions of $Ex_{MON}$ and $Ex_{TUE}$. Because "the true one between *MON* and *TUE*" is a definite description satisfied only when it is Monday or Tuesday, $Ex_{MON}$ and $Ex_{TUE}$ were false on Sunday, and as such, neither was believed on Sunday.

As already said, the above argument is still sketchy. For a rigorous discussion, we need a precise definition or characterization of the symbols in the argument. Thus, define *K*

---

17) A referee argues that $Ex_{MON}$ and $Ex_{TUE}$ cannot be atomic and so we cannot apply Garber's model here. But why did he need to treat sentences like $H \vdash E$ as atomic? It was because, if the agent recognized its logical structure of $H \vdash E$, then the agent should've assigned unit credence to it, so that it is impossible to learn it as new information. Fortunately, such an artificial technique is unnecessary for $Ex_{MON}$ and $Ex_{TUE}$. For note that they were simply false on Sunday. Hence, Sleeping Beauty did not (indeed, could not) know $Ex_{MON} \vee Ex_{TUE}$, and so it is entirely possible that it is newly learned by Sleeping Beauty on Monday.

18) As I will argue later, this is a *contingent* relation. Hence, one may protest that the Sleeping Beauty case is entirely different from, for example, the confirmation of general relativity based on the learning of a logical relation. But this protest misses a very important point: Garber himself never wished to confine his model to the cases in which a logical truth is newly learned. Indeed, Jeffrey applies Garber's model to a case where extrasystematic evidece is contingent, as we saw earlier.

to be the set of SB's background beliefs at *m* and *S* to be the minimal Boolean-operation-closed set of tensed propositions that includes *HEADS*, *TAILS*, *WAKEUP*, *MON*, *TUE*, and all members of *K*. Next, define E to be {*MON*, *TUE*} and let "*T*" be the abbreviation of "the true member of E". Hence, we can symbolize $Ex_{MON}$ into ~(WAKEUP $\vdash_{K\cup\{T\}}$ *HEADS*, *TAILS*) and $Ex_{TUE}$ into *WAKEUP* $\vdash_{K\cup\{T\}}$*TAILS*. Next, we expand *S* into $S_*$, the minimal boolean-operation-closed superset of *S* that includes $Ex_{MON}$ and $Ex_{TUE}$ also as members. Now, I am ready to defend the premises of the above argument.

First, I defend P1. According to her background beliefs about what would happen during the experiment, she can wakeup with the memory up to Sunday only on Monday or on Tuesday. Hence, she can deduce *MONvTUE* from *K* and *WAKEUP*.

Second, I defend P2. It is easy to show the equivalence at *m* between *MON* and $Ex_{MON}$ and that between *TUE* and $Ex_{TUE}$.[19] Hence, to SB at *m*, $Ex_{MON}$v$Ex_{TUE}$ is equivalent to *MON*v*TUE*.

Third, I defend P3. Obviously, $Ex_{MON}$v$Ex_{TUE}$ is *disjunctive* evidence about a broadly implicative relation *from evidence to a hypothesis*. Hence, it suffices to show the contingency of $Ex_{MON}$v$Ex_{TUE}$. On Sunday, "whichever is true between *MON* and *TUE*" has no referent and so both disjuncts are false. On Tuesday, $Ex_{TUE}$ amounts to the claim that waking up entails the coin's landing tails under her background beliefs and its being

---

[19] The proof is analogous to that in footnote 13, for the equivalence between *OHTERS* and $Ex_{OTHERS}$ and that between *ONLYME* and $Ex_{ONLYME}$ in **Example 3**. (Substitute *MON* for *OTHERS*, *TUE* for *ONLYME*, $Ex_{MON}$ for $Ex_{OTHERS}$, and $Ex_{TUE}$ for $Ex_{ONLYME}$.)

Tuesday, which is certainly true. Hence, the disjunction's truth depends upon SB's temporal location, which is a contingent matter.

Fourth, I defend P4. Given P1-P3, is $Ex_{MON} \vee Ex_{TUE}$ extrasystematic evidence? It seems to be. For observe its similarity to $Ex_{OTHERS} \vee Ex_{ONLYME}$ in **Example 3**. In its logical structure, $Ex_{MON} \vee Ex_{TUE}$ is similar to $Ex_{OTHERS} \vee Ex_{ONLYME}$, which we already accepted as extrasystematic evidence. Since Garber provides no necessary and sufficient condition of being extrasystematic evidence, this similarity, in my opinion, provides a good justification for considering it to be, at least until we find a reason to think otherwise.

This opens a room for challenging Lewis's argument for his Halfer view. For, even if her intrasystematic evidence is old about *de dicto* matters, maybe she can change her credence in *HEADS* because she has some relevant extrasystematic evidence. Given this possibility, let us reevaluate Lewis's original argument:

(L1) At $m$, *WAKEUP* is old evidence about *de dicto* matters.
(L2) If L1 is true, $C_{MON\ WAKEUP}(HEADS)=C_{SUN\ NIGHT}(HEADS)=1/2$.
(LC) Therefore, $C_{MON\ WAKEUP}(HEADS)=1/2$.

Since this argument is valid, it suffices to evaluate each premise.

First, L1 is true. To see this fact, think about the strongest information that *WAKEUP* has about what the world is like. It is that SB wakes up at some time with the memory up to Sunday. SB fully believed this on Sunday. Thus, when she receives *WAKEUP* on Monday, it does not include any new information

about what the world is like.

Second, there exists no strong reason to think that L2 is true, because SB receives extrasystematic evidence at *m*. Note that L2 is an instance of PRESERVATION. As already discussed, the principle applies only when the agent receives no new extrasystematic evidence. So it fails to support L2 in this case, because SB receives new extrasystematic evidence at *m*. And I see no other reason to accept L2.[20)]

Therefore, Lewis's argument is unsound. However, one may complain that, while his original argument is based on the misuse of PRESERVATION, it is possible to construct a similar argument on the basis of GENERAL PRESERVATION instead. Since the latter principle holds in the presence of extrasystematic evidence, the argument based on it may be more successful in proving the Halfer view. So think about this modified version of Lewis's argument:

(L1*) At *m*, $WAKEUP\&(Ex_{\text{MON}} \vee Ex_{\text{TUE}})$ is locally old about *de dicto* matters.
(L2*) If L1* is true, $C_{\text{MON WAKEUP}}(HEADS)=C_{\text{SUN NIGHT}}(HEADS)=1/2$.
(LC) Therefore, $C_{\text{MON WAKEUP}}(HEADS)=1/2$.

This argument is also valid, and so it suffices to discuss whether its premises are all true.

---

20) The principal principle may provide an independent support for L2. According to the principle, given that *X*'s objective chance is *r*, a rational agent's credence in *X* ought to be also *r* unless she has some kind of inadmissible evidence. However, the correct condition of inadmissibility is another issue of a big controversy. See Lewis (1980) for a discussion of the principal principle.

First, L2* is true. As we already discussed, a rational agent does not change her credence about topic *T* if she neither has new intrasystematic evidence about *T* nor has new extrasystematic evidence suitably related to *T*. This principle, GENERAL PRESERVATION, applies even when the given agent receives some extrasystematic evidence. And L2* is an instance of it.

Second, however, L1* is false. By definition, if $WAKEUP\&(Ex_{MON} \vee Ex_{TUE})$ is to be locally old evidence about *de dicto* matters, *WAKEUP* not only has to be old intrasystematic evidence about those matters but $Ex_{MON} \vee Ex_{TUE}$ also has to be old extrasystematic evidence. Now, remember that neither $Ex_{MON}$ nor $Ex_{TUE}$ was fully believed on Sunday. Neither was $Ex_{MON} \vee Ex_{TUE}$. Hence, when SB receives it as evidence on Monday, it is new evidence. Although it is not directly about *de dicto* matters, it is extrasystematic evidence about an implicative relation between *WAKEUP* and *HEADS*/*TAILS*. Therefore, the above argument is unsound.

In summary, I argued that although Sleeping Beauty receives no new evidence about what the world is like, she does receive new evidence about a broadly implicative relation between *WAKEUP* and *HEADS*/*TAILS*. Therefore, I conclude that Lewis's original argument and the modified one are both unsound.

## 9. Are the Thirders also Wrong?

At this point, the Halfers will perhaps appeal to a *tu quoque* strategy: Even if it is true that Sleeping Beauty receives new

extrasystematic evidence waking up on Monday, this fact not only implies the unsoundness of Lewis's argument but it also implies that of Elga's. In this section, I will defend Elga's view against this charge.

To begin, I formulate Elga's argument as follows:

(E1) $C_{\text{MON WAKEUP}}(HEADS/WAKEUP\&MON)=C_{\text{SUN NIGHT}}(HEADS)=1/2$.
(E2) $C_{\text{MON WAKEUP}}(HEADS/WAKEUP\&TUE)=0$.
(E3) If E1 and E2 are true, then $0<C_{\text{MON WAKEUP}}(HEADS)<1$.
(EC) Therefore, $0<C_{\text{MON WAKEUP}}(HEADS)<1$.

This argument falls short of showing that, when she wakes up on Monday, SB believes *HEADS* to the degree of 1/3. Elga argues that this credence must be precisely 1/3, based on an additional assumption. Since I am not sympathetic with his assumption, I will be content with showing that at that moment, she believes *HEADS* to a degree less than 1/2. Anyway, the above argument is clearly valid, and so it suffices to defend each of E1-E3.

First, one may defend E1 by appealing to the principle of CONDITIONAL PRESERVATION: On Sunday, SB fully expected that she would wake up on Monday. Hence, even if she is now waking up on Monday, this fact cannot be new information about de *dicto* matters. Since *HEADS* is a proposition entirely about those matters, E1 follows from the mentioned fact and the principle's relevant instance.

Second, E2 is uncontroversial. For her background knowledge at *m* logically entails that she wakes up on Tuesday only if the coin lands on tails.

Third, E3 is also uncontroversial. For, waking up on Monday,

she cannot be sure about whether it is Monday or Tuesday. Clearly, her then credence in *HEADS* should be the weighted average of the two conditional credences in E1 and E2, with the weights being her credences at *m* in *MON* and in *TUE*. From these facts, it follows that her credence at *m* in *HEADS* should be somewhere between those two conditional credences exclusive. If E1 and E2 are true, those conditional credences are 0 and ½. Therefore, E3 is true.

So is the above argument sound? No, because the above defense of E1 was faulty. Remember that CONDITIONAL PRESERVATION holds only when the given agent does not receive new extrasystematic evidence. In the previous section, I argued that SB receives some new extrasystematic evidence, waking up on Monday. Hence, E1 cannot be defended on the basis of the principle.

At this point, someone may suggest reconstructing Elga's argument in terms of extrasystematic evidence. For, even in the presence of extrasystematic evidence, a rational agent ought to obey GENERAL CONDITIONAL PRESERVATION, at least. Here is the thus reconstructed argument:

(E1*) $C_{\text{MON WAKEUP}}(HEADS/WAKEUP\&Ex_{\text{MON}}) = C_{\text{MON WAKEUP}}(HEADS) = 1/2$.

(E2*) $C_{\text{MON WAKEUP}}(HEADS/WAKEUP\&Ex_{\text{TUE}}) = 0$.

(E3*) If E1* and E2* are true, then $0 < C_{\text{MON WAKEUP}}(HEADS) < 1$.

(EC) Therefore, $0 < C_{\text{MON WAKEUP}}(HEADS) < 1$.

This argument is obviously valid and so it suffices to evaluate the premises.

First, E2* is uncontroversial. By definition, $Ex_{TUE}$ states that *WAKEUP* entails *TAILS* under her stock of knowledge and whichever is true between *MON* and *TUE*. Since she knows at that moment that *TAILS* implies the negation of *HEADS,* the conditional credence in E2* must be zero.

Second, E3* is also indisputable. To see this, remember that Sleeping Beauty is sure at *m* that *MON*v*TUE* is true. So she is sure at *m* that $Ex_{MON}$v$Ex_{TUE}$. And the rest of the demonstration is analogous with that of E3.

Third, it is however difficult to defend E1*. One may think that we can defend it by appealing to GENERAL CONDITIONAL PRESERVATION. According to the principle,

(6) If *WAKEUP*&$Ex_{MON}$ is locally old at *m* about *de dicto* matters, $C_{MON\ WAKEUP}(HEADS/WAKEUP\&Ex_{MON})=C_{MON\ WAKEUP}(HEADS)$.

So if the antecedent of the above claim is true, its consequent, E1*, is also true. However, the antecedent is false, because SB did not fully believe $Ex_{MON}$ on Sunday night.

Until now, we have discussed four arguments. Each of them included a premise that is either an instance of one of the principles discussed in Section 7 or seemingly defensible by them. However, so far, none of them have succeeded in settling the debate. This suggests that we cannot settle the debate with an argument formulated in that way.

Nevertheless, I think that we are now in a significantly better position. First, before this paper, philosophers had two equally attractive arguments, which jointly led to a contradiction. In this

paper, I argued, in my opinion successfully, that neither of them is sound. And it is better to have no answer than to have a contradictory one.

Second, even if it is difficult or even impossible to settle the debate by appealing to a proposition's being old news, it remains to be an open possibility that someone can settle the debate in a completely different way. Indeed, I defended a solution of the SB problem elsewhere that is largely favorable to the Thirder view and incompatible with the Halfer view, using a quite different argument (AUTHOR YEAR).

Third, although my earlier defense of E1* based on the alleged local oldness of $WAKEUP\&Ex_{MON}$ was unsuccessful, I believe that it is still an open possibility to defend E1* on different grounds. Here is my basic idea: Although $Ex_{MON}$ is new extrasystematic evidence, it cannot drive SB's opinion at *m* towards *HEADS* or towards *TAILS* because it includes only negative information about the relevant implicative relations.

Since the last point will be quite important if true, let me elaborate this idea before finishing this section. Suppose that, waking up on Monday, SB wants to figure out what value to assign as her conditional credence in HEADS given $WAKEUP\&Ex_{MON}$. For this purpose, she adds the conjunction hypothetically to her stock of knowledge and tries to calculate an optimal credence in *HEADS* on that basis. The first conjunct is old information about de *dicto* matters and so cannot drive her opinion towards *HEADS* or towards *TAILS*. To see why the second conjunct cannot do it either, remember that it was defined

as follows:

($Ex_{MON}$) *WAKEUP* entails neither *HEADS* nor *TAILS*, under SB's background beliefs *K* and the true one between *MON* and *TUE*.

On Sunday night, SB did not fully believe this proposition, and so waking up on Monday, it will be new extrasystematic information to her. However, note that $Ex_{MON}$ is purely negative about the implicative relations that it mentions. Since she was neutral between *HEADS* and *TAILS* before and she now receives purely negative information those relations, $Ex_{MON}$ cannot drive her opinion towards *HEADS* or *TAILS* in this situation. Hence, neither conjunct of $WAKEUP\&Ex_{MON}$ can change her credence in *HEADS* as a result of this hypothetical calculation. By Ramsey's thesis, SB's conditional credence at *m* in *HEADS* given the conjunction should be the same as her unconditional credence at *s* in *HEADS*. Therefore, E1* is true. Since no other premises were controversial, the argument consisting of E1*-E3* and EC proves Elga's view.

In summary, both Elga's original argument and the revised version of it will be unsuccessful if the first premise of each argument is defended by CONDITIONAL PRESERVATION or GENERAL CONDITIONAL PRESERVATION, but it is possible to defend the latter argument's first premise by taking different matters into consideration. Therefore, Lewis's argument was shown to be unsound and Elga's view was defended.

## 10. Conclusion

To the proponents of Elga's view, there has been a very difficult question to answer: "How can SB change her credence in *HEADS* on awakening on Monday, although she has no new evidence about what the world is like?" If I have been right, here is the right answer: "Although SB has no new *intrasystematic* evidence about de dicto matters, she has some *extrasystematic* evidence capable of changing her credence in *HEADS*." Therefore, I believe that the Sleeping Beauty problem is just another case of the problem of old evidence and that as such, we can apply Garber's solution to the former.

## References

Einstein, A. (1997), "The Foundation of the General Theory of Relativity", *the Collected Papers of Albert Einstein*, translated by Alfred Engel, Princeton University Press, Princeton. pp. 146-201.

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

Eells, E. (1985), "Problems of Old Evidence", *Pacific Philosophical Quarterly*, Vol.66. pp. 283-302.

Garber, D. (1983), "Old Evidence and Logical Omniscience", *Testing Scientific Theories*, University of Minnesota Press, Minneapolis. pp. 99-131.

Glymour, C. (1980), *Theory and Evidence,* Princeton University Press.

Jeffrey, R. (1983), "Bayesianism with a Human Face", *Testing Scientific Theories*, University of Minnesota Press, Minneapolis, pp. 133-156.

Kim, N. (2009), "Sleeping Beauty and Shifted Jeffrey Conditionalization", *Synthese*, Vol.168, 2, pp. 295-312.

Lewis, D. K. (1986), "A Subjectivist's Guide to Objective Chance", *Studies in Inductive Logic and Probability,* Vol. Ⅱ, University of California Press, Berkeley, pp. 263-293.

Lewis, D. K. (2001), "Sleeping Beauty: Reply to Elga", *Analysis*, Vol. 61, 271. pp. 171-176.

Ramsey, F. P. (1929), "General Propositions and Causality", in his *Philosophical Papers*, Cambridge University Press,

Cambridge, pp. 145-163.

Ramsey, F. P. (1931), “Truth and Probability”, in *Foundations of Mathematics and Other Logical Essays*, Harcourt, Brace and Company, New York, pp. 199-211.

Weintraub, R. (2004), “Sleeping Beauty: a Simple Solution”, *Analysis*, Vol.64, 1, pp. 8-10.

성균관대학교 비판적사고와 문화연구소
The Research Institute of Critical Thinking and Culture, Sungkyunkwan University
newitx@gmail.com

# 잠자는 미녀 문제에 대한 가버식 해결책

김 남 중

이전 논문(2009)에서 나는 잠자는 미녀 역설에 대한 한 가지 해결책을 제시하였는데, 그에 의하면 미녀가 동전 앞면에 월요일에 부여하는 확률은 1/2보다 낮아야 한다. 이것은 물론 1/3주의에 유리한 결론이다. 그렇지만 내가 1/3주의를 성공적으로 옹호했다고 할지라도, 한 가지 중요한 물음이 남는다: 왜 1/2주의는 틀렸는가? 그들의 주요 논변은 간단하다: 잠자는 미녀는 동전이 어떻게 땅에 떨어지는지에 대한 새로운 증거를 받지 못했기 때문에, 그녀가 그 가능성에 부여하는 확률은 이전과 같아야 한다. 이제 다음 사실에 주목해 보자: 만일 1/3입장이 옳다면 잠자는 미녀 역설은 이른바 오래된 증거 문제의 새로운 예가 될 것이다. 이 논문에서 나는 새롭고 직접적으로 관련된 증거가 없음에도 왜 잠자는 미녀가 그녀의 믿음의 정도를 바꿀 수 있는지 대니얼 가버(1983)가 오래된 증거 문제에 대해 내놓은 해결책을 가지고 설명할 것이다.

주요어: 잠자는 미녀, 오래된 증거, 가버, 자기위치 믿음