논리연구 13-1(2010) pp. 21~52

Sleeping Beauty's Reflection: In and Out* ** †

Hanseung Kim

【Abstract】 What van Fraassen calls 'Reflection Principle' is claimed to meet several counterexamples, one of which stands out in the form of the Sleeping Beauty problem. Adam Elga argues that what he believes is the correct answer to the Sleeping Beauty problem shows that Reflection is subject to counterexamples. David Lewis proposes a different answer which preserves Reflection intact. Recently, Nick Bostrom presents a hybrid view which is supposed to allow us to keep Reflection. In proposing his hybrid view Bostrom criticizes both Elga and Lewis while taking some 'good' parts from each. He claims that Elga's view is not entirely acceptable because it presupposes the 'Self-Indication Assumption'. I shall claim, however, that Elga could avoid Bostrom's criticisms by introducing Bostrom's notion of agent-part. I believe that several probability-related puzzles including the Sleeping Beauty problem indicate a promising view concerning the way we should regard our future selves' opinions. According to this view, whether one takes the outsider stance or insider stance makes a difference in an important way that one and the same proposition is associated with different degrees of belief by one agent.

【Key words】 Sleeping Beauty Problem, Reflection Principle, Self-Indication Assumption, Indexicality, probability, puzzles

* 접수완료: 2009.12.17 심사 및 수정완료: 2010.1.17

** This work was supported by research program 2009 of Kookmin University in Korea.

† I am deeply grateful to three anonymous referees for their helpful and valuable comments. Although I am not capable of fully responding to each of their concerns, they will sharpen my insights and lead me to a better work.

How seriously should we regard our future selves' voices? Many believe, as seriously as we do our present selves' voices. Some go further to take the following advice as our principle: Listen to our future selves' opinions and take them as guide to our present opinions. Van Fraassen calls this 'Reflection Principle', which is formulated as follows (van Fraassen 1984: 244):

[*Reflection Principle*] $P_t(A/P_{t'}(A) = r) = r$

where P_t represents a probability function that gives a real number between 0 and 1 to a proposition A at time t and t is prior to t'. According to Reflection, given that one predicts that she will believe that A has the probability r, she should also believe *now* that A has the same probability.

Several counterexamples to this principle have been suggested, one of which stands out in the form of the Sleeping Beauty problem. Adam Elga argues that what he believes is the correct answer to the Sleeping Beauty problem shows that Reflection is subject to counterexamples (Elga 2000: 144). David Lewis (2001) proposes a different answer which preserves Reflection intact. Recently, Nick Bostrom (2007) presents a hybrid view which is supposed to allow us to keep Reflection. In proposing his hybrid view Bostrom criticizes both Elga and Lewis while taking some 'good' parts from each. He claims that Elga's view is not entirely acceptable because it presupposes the 'Self Indication Assumption'. I shall claim, however, that Elga could avoid

Bostrom's criticisms by introducing Bostrom's notion of agent-part. I believe that several probability related puzzles including the Sleeping Beauty problem indicate a promising view concerning the way we should regard our future selves' opinions. According to this view, whether one takes the outsider stance or insider stance makes a difference in an important way that one and the same proposition is associated with different degrees of belief by one agent.

1. Reflection

It has been argued that Reflection Principle faces counterexamples. And I believe the suggested examples are classified into two groups: one is the *predictable irrationality* group and the other is the *predictable judgment impairment* group. To the former belong the counterexamples in which one anticipates that her rationality shall be jeopardized by some external causes. William Talbott's example is the one in which someone believes that after she drinks in a party she will believe a thing that she does not endorse at all right now (Talbott 1991: 136; Also Maher 1992: 120-122; Christensen 1991: 234-236). In the same vein van Fraassen (1995) brings up the example of Ulysses who rationally foresees that his judgment will not work properly in the near future.

In these examples it seems that rationality is not sacrificed even if the Reflection Principle is violated. Talbott, however, believes one can defend Reflection from this counterexample by denying that Reflection alone is sufficient for epistemic rationality. One might argue as follows: Reflection is not applicable to the drink party example because 'the agent knows that she will violate separate and joint requirement on epistemic rationality, *Temporal Conditionalization*,' (1991: 137) which says that probability assignment to a proposition at a future moment should conditionalize on the new evidence that will be acquired up to the moment, i.e.,

[*Temporal Conditionalization*] $P_{t'}(A) = P_t(A/E)$

Let A stand for the proposition 'I can drive my car now' and E for the evidences the agent acquire during the drink party including that the agent drinks a lot. The agent gives a low probability to $P_t(A/E)$, say, 0.1. At the moment t' when the party is over, however, she will assign a different probability to A, say, 0.9. Thus Temporal Conditionalization is violated. Because of this foreseeable irrationality due to the alcoholic influence, one may claim, the drink party example is not a genuine counterexample to Reflection.

On the other hand, van Fraassen defends Reflection by generalizing it. He (1995: 16) argues that Reflection Principle is a special case of the following general principle:

[*General Reflection Principle*] My current opinions about event E must lie in the range spanned by the possible opinions I may come to have about E at later time t, as far as my present opinion is concerned.

If I foresee that I shall have a belief I do not endorse right now, General Reflection is not violated. For, van Fraassen (1995: 20) argues, General Reflection 'forbids only opinion which is at odds with any and all opinions I think I may come to have a certain future time.'

The second group of counterexamples to Reflection involves judgment impairment due to memory loss, losing track of time or other epistemic mishaps.[1] Talbott's spaghetti example (1991: 138-140), for example, is the one in which one shall not remember, and hence shall have low credence in, that she had spaghetti today. He argues that the spaghetti example is not a violation of Temporal Conditionalization but a violation of Reflection.[2]

The second group of counterexample is not free from criticisms. Schervish et al (2004: 316) argue that there are two additional assumptions required for Reflection: that the agent should not lose any information between two moments (the *filtration* requirement) and that future times should be stopping times, i.e., for each time t when a prediction is to be made it is known to the agent whether a certain moment t' is prior to t (the *stopping times* requirement). According to Schervish et al,

[1] Arntzenius's duplication example (2003: 364-366) involves losing track of self identity.

[2] For the counterexamples which involve losing track of time, see Arntzenius (2003: 357-362) and Elga (2007: 482).

Talbott's spaghetti example does not fulfill the filtration requirement.[3] Elga's train example does not fulfill the stopping times requirement.[4]

These defenses, however, lead us to different, and even conflicting, conceptions about Reflection Principle. On one hand, the two additional requirements above are humanly impossible to be fulfilled, which makes Reflection Principle too idealistic. On the other hand, van Fraassen's General Reflection makes fulfilling Reflection relatively easy. Even if these revised versions of Reflection are accepted, I believe, it is still arguable that the counterexamples do not lose their forces. In the formulation of General Reflection, one might claim, the expression 'possible opinions' is vague. Or as to the filtration requirement, one might claim, it is not entirely clear what we should count as 'information'. Thus, I believe, there exist still good reasons for reconsidering counterexamples to Reflection and constraints upon Reflection. What is the nature of Reflection? Is it a moral principle that you should be true to your future self? Or is it an epistemic requirement, whose failure costs rationality? If Reflection requires additional assumptions, what are their philosophical implications? These questions are still moot. And I believe that reconsidering counterexamples is a good starting point

3) The counterexamples involving losing track of time are alleged to violate the stopping times requirement.

4) One might complain that defending Reflection by imposing additional requirements is nothing but allowing counterexamples to Reflection per se. Then it becomes a verbal issue whether a violation of additional requirements implies a violation of Reflection.

for sorting these questions out. I propose to begin with the Sleeping Beauty problem.

2. The Sleeping Beauty problem

This famous problem is presented in the following thought experiment: On Sunday Sleeping Beauty is told by the researchers that she will be put to sleep on Sunday night. And then a fair coin will be tossed. If it lands heads, she will be woken up once on Monday. If it lands tails, she will be woken up twice on Monday and Tuesday. When she is woken up she will be asked what credence she has in that the coin lands heads. She is told that after she answers the question she will be given a drug that has a power to erase her memory of waking up. Given that she wakes up, what credence should she give to the belief that the coin landed heads?

Let us introduce the names for propositions as follows:

HEADS: The coin lands heads.
TAILS: The coin lands tails.
H: The coin lands heads and Beauty wakes up on Monday
T1: The coin lands tails and she wakes up on Monday.
T2: The coin lands tails and she wakes up on Tuesday.

And let us use the following credence functions:

$P\_(Q)$: The credence function Beauty should give to Q on Sunday.

$P(Q)$: The credence function Beauty should give to Q upon awaking without being told what day it is.

$P_+(Q)$: The credence function Beauty should give to Q after she is told that it is Monday.

The Sleeping Beauty's problem is, what is P(HEADS)? Without further explanations the followings are straightforward:

(1) $P(HEADS) = P(H)$

(2) $P(H \vee T1 \vee T2) = 1$

(3) $P\_(HEADS) = P\_(TAILS) = 1/2$

Elga (2000) and Lewis (2001) offer different answers to the problem. Elga's answer is 1/3. His claims are:

(E1) $P(H) = P(T1) = P(T2) = 1/3$

(E2) $P_+(HEADS) = P_+(H) = P(H/H \vee T1) = 1/2$

Lewis's answer is 1/2. His claims are:

(L1) $P(H) = P(T1 \vee T2) = 1/2$

(L2) $P_+(HEADS) = P_+(H) = P(H/H \vee T1) = 2/3$

Both Elga and Lewis agrees that P(T1) is identical to P(T2). For if TAILS is true, Beauty has no reason to favor T1 over T2 or vice versa. Disagreement begins where Elga claims that $P_+(HEADS)$ should be 1/2 while Lewis claims that P(HEADS) should be identical $P\_(HEADS)$. Elga believes that it does not change the original problem even if the coin tossing occurs after Beauty's awakening on Monday. Then $P_+(HEADS)$ is just the probability of a fair coin's landing heads, i.e., 1/2. Lewis believes that P(HEADS) should be identical $P\_(HEADS)$ unless Beauty come to have new relevant information upon awakening. And Lewis claims that there is no such information.

It should be noted that Elga's answer implies a *prima facie* violation of Reflection while Lewis's answer is not. According to Elga, on Sunday Sleeping Beauty should assign the credence 1/2 to HEADS, while she anticipates that she should assign 1/3 to the same proposition when she is woken up. On the other hand, according to Lewis, the credence in HEADS does not change unless the new relevant information is given to Beauty. Reflection is intact.[5)]

Both Elga and Lewis agree that $P_+(H)$ is greater than P(H) by 1/6 because Beauty comes to have a new relevant information when she is told that it is Monday. On the other hand, Lewis believes that Sleeping Beauty gains no new relevant information when woken up. Elga also agrees that there is no new

---

5) One might claim that Elga's 1/3 view no more violates Reflection than Lewis's 1/2 view because Reflection is inapplicable since the memory loss is expected in this example. It is still arguable, as I said, whether Reflection requires no memory loss condition. Beside this, however, there is still a difference between Elga and Lewis regarding Reflection: Lewis would claim that Reflection is intact regardless of memory loss.

information (2000: 146) and yet holds the 1/3 view by claiming that Reflection is violated in this example. However, there is another logically possible option for the 1/3 view without undermining Reflection: one can claim that Sleeping Beauty gets new relevant information upon awakening. Let us call this option the *New Information* view and examine where this view leads us.

If one takes the New Information view, two questions are in order. (1) Is Reflection or Temporal Conditionalization violated? (2) What is the new information that Sleeping Beauty receives? For the first question: whatever this new relevant information (say X) is, there is a room for holding that Temporal Conditionalization is intact. For depending on X, it is possible to accept that P(HEADS) = P_ (HEADS/X) = 1/3 while P_(HEADS) = 1/2. However, Reflection is still violated since P_(HEADS) is not identical to P(HEADS). (Or, one can claim, Reflection is inapplicable because the filtration condition is not met.)

Now move to the second question. The only possible candidate for this information X, it seems, is expressed by the statement 'I am awake *now*.' According to Lewis (and Elga too), this information is nothing but that Beauty is situated in one of the three cases, i.e., (H v T1 v T2). Since (1), P_(H/H v T1 v T2) should be identical to P_(HEADS). Then, New Information view collapses to the 1/2 view. Thus X cannot be (H v T1 v T2).

Ruth Weintraub (2004) holds the 1/3 view by arguing that 'I am awake *now*' is new relevant information Beauty receives. And, she claims, this information is not equivalent to (H v T1 v T2), since the former is available to Beauty only if she actually wakes up while the latter is not. This difference is sometimes crucial for activating actions. She says: "the belief that it will rain *sometime* doesn't motivate me to take an umbrella, whereas the belief that it is raining *now* does." (2004: 9) I believe that New Information theorist should agree with Weintraub that X is aptly expressed as 'I am awake now'. She calls her proposal 'a simple solution'. However, her proposal does not satisfy opponents. It is well argued by Perry (1979) that indexical information is required to perform some actions. Bringing an umbrella is one of such actions. New Information theorists, however, should explain how Beauty's receiving indexical information could change her credence in the non indexical proposition, HEADS, *not* her performing an action. Weintraub offers no such explanation.

There is another proposal, according to which, a person is viewed as a collection of agent-parts. In the Sleeping Beauty case, there exist distinct agent-parts on Sunday, Monday and Tuesday. When Beauty awakes, the information she receives is to be expressed as the sentence 'I am no more Sunday agent-part.' This information is not available to her on Sunday. That is, on Sunday she anticipates that she shall have this information upon awakening but is not able to entertain it. Can New Information theorists live happily with this proposal?

3. Bostrom's Hybrid View

Bostrom suggests that we look at the Sleeping Beauty problem in terms of agent-parts (2003; 2007). But his answer to the problem is not 1/3. Instead he proposes a hybrid view, according to which P(HEADS) should be 1/2 (as in the 1/2 view) and P_+(HEADS) should also be 1/2 (as in the 1/3 view). Bostrom's crucial move is to introduce agent-parts in a finer way: an agent is divided into parts in terms of relevant epistemic states as well as relevant temporal moments and possible situations. According to him, for example, we should consider the following possible agent-parts:

h': Monday agent-part that does *not* know that it is Monday, and HEADS
h*: Monday agent-part that knows that it is Monday, and HEADS
t1': Monday agent-part that does *not* know that it is Monday, and TAILS
t1*: Monday agent-part that knows that it is Monday, and TAILS
t2': Tuesday agent-part that know does *not* that it is Tuesday, and TAILS
t2*: Tuesday agent-part that knows that it is Tuesday, and TAILS.

Reflecting these agent-parts, we need a different set of names for propositions as follows:

H': h' is the agent-part now.

H*: h* is the agent-part now.
T1': t1' is the agent-part now.
T1*: t1* is the agent-part now.
T2': t2' is the agent-part now.
T2*: t2* is the agent-part now.

Bostrom's claims are[6]:

(B1) P(H) = 1/2
(B2) P(HEADS/H v T1) = 2/3
(B3) P_+(HEADS) = P_+(HEADS/H* v T1*) = 1/2

According to Bostrom, when Beauty is told that it is Monday the information she gets is (H* v T1*), not (H v T1). Notice that Temporal Conditionalization fails here. That is,

$$P_+(\text{HEADS}) \neq P(\text{HEADS}/\text{H}^* \vee \text{T1}^*)$$

For P(HEADS/H* v T1*) is undefined since P(H* v T1*) is 0. On the other hand, Reflection is not violated in his hybrid view. That is,

$$P(\text{HEADS}/\ P_+(\text{HEADS}) = 1/2) = 1/2$$

Let us consider whether Bostrom's hybrid view allows that Beauty gets new relevant information when she is woken up

6) Bostrom also accepts (1) to (3) above as Elga and Lewis.

without being told which day it is. For Bostrom, P(HEADS/H v T1) is identical to P(HEADS/H' v T1'). But (H' v T1') cannot be information available to Beauty upon awakening on Monday. The only possible candidate would be (H' v T1' v T2'). This indexical information is not available to Beauty on Sunday. Neither is it available to Beauty who is told which day. On the other hand, Beauty should give credence 1 to (H' v T1' v T2') upon awakening without being told which day. That is, this information is something that Beauty can come to possess only under specific conditions but it represents a sure event.[7] Both Elga and Lewis deny that this is new relevant information. The difference is that Elga accepts, but Lewis denies, Beauty's credence change in HEAD upon awakening and thus Elga denies, but Lewis holds, Reflection. Bostrom is not different from them in this regard: Beauty gets no new relevant indexical information upon awakening.

I shall argue that Beauty's indexical information (H' v T1' v T2') is crucial enough to change her credence in HEADS. That is, I shall defend a New Information view. My argument is based upon Bostrom's notion of agent-part. This view is a live option to Bostrom if he accepts the 1/3 view to the Sleeping Beauty problem. But he does not. I shall first explain his worry that forces him not to take that option. And then I shall defend New Information view from his worry.

---

[7] The idea of indexical information that is accessible only to a specific agent or at a specific time or place is suggested and abandoned in Perry (1979) under the name of 'Limited Accessibility'.

4. SIA and the Outsider/Insider Distinction

Bostrom believes that the 1/3 view presupposes what is called *Self Indication Assumption* (SIA hereafter) and SIA is wrong. SIA states an assumption that (other things equal) we should prefer hypotheses according to which many observers exist over hypotheses on which few observers exist. Applied to the Sleeping Beauty problem, Bostrom believes, SIA implies that the credence in TAILS is greater than HEADS because the TAIL hypothesis posits twice more agent-parts than the HEADS hypothesis does. SIA is untenable, Bostrom argues, since it renders an issue of an empirical nature *a priori*. Bostrom (2007) performs the following thought experiment: There are two astronomical hypotheses H1 and H2. These two hypotheses are alike in every respect except that H1 implies that a universe contains more observers than H2. Then, according to SIA, we should conclude that H1 is more probable than H2, which makes any empirical experiments to verify the hypotheses unnecessary. Bostrom believes this is outrageous.

There is a parallel but somewhat different reasoning. Suppose that you awake in a place where you have never been and you cannot remember how you came to be there. You are told that you surely come from either city A or city B, and the population of city A is one million while that of B is only thousand. Then, you should reason that the credence in the proposition that you come from A is more than B. This reasoning looks like another application of SIA, but, according to Bostrom, it is not. The difference lies in that the populations of city A and city B is *actual* while the number of observers in the hypotheses H1 and H2 are only *hypothetical.*

Bostrom believes that the 1/3 view leads to the same absurdity as in the astronomical theory case. In the Sleeping Beauty problem Beauty on Sunday considers two hypotheses, one of which implies one awakening while the other two awakenings, and all the awakenings are only hypothetical to Beauty on Sunday. If these awakenings were all actual for some reason, then Bostrom would believe that P(HEADS) is 1/3. He proposes the 'N-fold Sleeping Beauty' problem (Bostrom 2007) which is just like the original problem except that the experiment is repeated N times on following N weeks. As N increases, Bostrom argues, P(HEADS) approximates 1/3. The Sleeping Beauty problem, he says, is a special case in which N is 1. It looks quite unnatural, however, that P(HEADS) is 1/2 when only one tossing happens and approximates 1/3 as the number of tossings increases. If there is a way to evade Bostrom's worry, it seems better to stick to either 1/2 or 1/3.

I believe that we can evade his worry that the 1/3 view leads to accepting SIA. (I shall leave aside whether SIA *per se* is tenable.) SIA implies that our existence as an observer is a key factor to choose over hypotheses H1 and H2 that imply different numbers of observers. Notice that SIA presupposes that we belong to the group of observers implied by H1 or H2. However, if we take Bostrom's notion of agent-parts, we do not have to accept that such a *belonging*-relation holds in the Sleeping Beauty case. Sunday agent-part a Beauty belongs to neither {h', h*} nor {t1', t1*, t2', t2*}. Thus, SIA does not force Beauty to accept that P_(HEADS) is 1/3. Also, technically speaking, there is no agent-part of Beauty that is considering, upon awakening, whether she belongs to {h', h*} or {t1', t1*, t2', t2*}. Thus, again, SIA does not force Beauty to accept that P(HEADS) is 1/3.

What leads to the 1/3 view is not SIA but (internal) empirical constraints that subjectively identical experiences should have the same role in justifying one's belief.[8] The agent-parts h', t1' and t2' shall have the same experience which plays the same role in assigning the credence in HEADS. And it is this experience that yields indexical information 'I am awake *now*', and this information is equivalent to (H' v T1' v T2'). Given that SIA is not a threat to the 1/3 view, we do not have to embrace Bostrom's hybrid view that implies unnatural probability change in the N-fold Sleeping Beauty problem.

The next step is to show that receiving indexical information such as 'I am awake now' (not only sometimes motivates an action but also) sometimes yields a change in the credence in a non-indexical proposition. My argument depends upon what I shall call the '*outsider/insider* distinction'. The distinction comes from a very natural observation that we human beings have the imaginative power to see ourselves as others that we watch, meet and interact with. That is, we can take the outsider stance

8) Gupta (2006) calls such constraints Equivalence (epistemic equivalence of subjectively identical experience) and Reliability (reliability of all experience). See Chapter 2 of Gupta (2006).

towards ourselves. We also have the imaginative power that works in the opposite direction: we can imagine ourselves to take someone else's viewpoint. An agent-part can be an *insider* with respect to experience E either by having E or by imagining herself to have E. On the other hand, an agent-part can be an outsider with respect to E either by not having E or by failing (or refusing) to imagine herself to have E. That is, there are two criteria to decide whether an agent-part is an insider or outsider with respect to E: the *actuality* criterion and the *imagination* criterion. One can be an insider according to one but an outsider according to the other.

Beauty, or her agent-part, can be an insider with respect to the experience of awakening: either by actually awakening or by imagining herself to being woken up. Beauty's agent-part on Sunday remains an outsider with respect to the experience of awakening unless she imagines herself to awake. Notice that imagining oneself to have an experience is not the same as imagining oneself to see someone else or one's different agent-part have the experience.[9] Beauty's agent-part on Sunday may imagine herself to see her later agent-part h' awake. But this imagining does not make Sunday agent-part an insider with respect to the experience of awakening. As far as Beauty remains as an outsider with respect to experience of awakening, she should assign 1/2 to HEADS. On the other hand, as far as she remains as an insider with respect to the same experience, she should assign 1/3 to HEADS. As far as Beauty remains as either

[9] For more on this distinction, see Walton (1990).

an outsider or insider, Temporal Conditional and Reflection hold only if the experience of awakening (say AWAKE) is not new relevant information. That is,

> [As an outsider with respect to AWAKE]
> P_(HEADS) = P_(HEADS/AWAKE) = P(HEADS) = 1/2
> [As an insider with respect to AWAKE]
> P_(HEADS) = P_(HEADS/AWAKE) = P(HEADS) = 1/3

Only when Beauty changes her stance from outsider to insider, the credence in HEADS changes from 1/2 to 1/3 and the experience of awakening is regarded as yielding new relevant information.

It is important to notice that one is not entirely free to choose the insider or outsider stance. When Beauty is told that it is Monday, she cannot be an outsider with respect to the experience of being told that it is Monday without falling into absurdity. Also, depending on circumstances, taking the outsider stance rather than the insider stance is relatively easier or more natural, and vice versa. For instance, if Beauty has suspicion about the whole experiment she would refuse to be an insider with respect to AWAKE.

Some information is relevant enough to change credence even though there is no outsider/insider change. Being told that it is Monday yields such information. Perceptual experiences yielding indexical information such as being woken up require the outsider/insider change in order to provide new relevant information. I believe that the outsider/insider distinction represents ways in which our agent-parts are evaluated. Sometimes they are treated as our full fledged extensions and sometimes as separate agent-parts disconnected from the present agent-part. Next, I shall show that the outsider/insider distinction plays a crucial role in resolving other probability related puzzles.

5. More Probability Puzzles

Cable guy paradox[10]

Suppose you and your friend are awaiting the cable guy who shall visit your place sometime between 8:00 am and 4:00 pm tomorrow. Since you and your friend do not know the exact visit time, it seems, you should give the same credence to the proposition that the cable guy arrives between 8:00 am and noon (BEFORE, for short) as to the proposition that the cable guy arrives between noon and 4:00 pm (AFTER, for short). Thus, when you put 10 dollars on BEFORE and you friend put 10 dollars on AFTER, it seems a fair game. But you realize that you have put the money on AFTER for the following reason: Suppose it is 8:00 am now. At that moment the credence in BEFORE is identical to the credence in AFTER. As the time goes, however, the credence in BEFORE is getting smaller. At 11:59 am the probability of your winning the game is only 1/240. This sounds bizarre: Once the game begins your winning

[10] This paradox is originally given in Hájek (2005).

probability decreases, which forces you to prefer AFTER to BEFORE. However, the credence in BEFORE is identical to the credence in AFTER before the game begins.

Once the outsider/insider distinction is introduced, we can have better understanding about this paradox. When you are awaiting the cable guy in your place between 8:00 am and noon, you are an insider with respect to the experience that yields the indexical information, 'my present agent-part is the one who does not see the cable guy visit my place' (NO VISIT, for short). Before the game begins, you (and your friend) are an outsider with respect to NO VISIT. Let us call P_(Q) the credence function you give to Q before the game begins and P(Q) the credence function you give to Q after the game begins. As far as you remain an outsider, your credence in BEFORE is identical to that in AFTER. That is,

$$P\_(BEFORE) = P\_(AFTER) = P(BEFORE/NO\ VISIT) = P(BEFORE) = P(AFTER) = 1/2$$

For suppose that you and your friend play the same game somewhere *outside* your place, under which circumstance it is natural for you to take an outsider stance with respect to NO VISIT. However, as far as you remain as an insider with respect to NO VISIT, your credence in BEFORE is smaller than that in AFTER. That is,

P_(BEFORE) = P_(BEFORE/NO VISIT) < P_(AFTER)
P(BEFORE) = P(BEFORE/NO VISIT) < P(AFTER)

Only when you change your stance from outsider to insider, your credence in BEFORE change and NO VISIT yields new relevant information. That is,

P_(BEFORE) = 1/2 > P_(BEFORE/NO VISIT) = P(BEFORE)

Two daughter puzzle[11]

In a conversation with Smith, who you meet first today, you are told that he is a father of two. He says that at least one of them is a daughter. Then, you reason that the probability that Smith has two daughters is 2/3. For suppose that you choose in a random way one hundred fathers with two kids. Then you expect that about 25 fathers have two daughters, 25 have two sons, and 50 have one son and one daughter. Since Smith has at least one daughter, the conditional probability is 25/75 (= 2/3). Now suppose he lets you meet one of his two kids. Then, it seems, the conditional probability seems change into 1/2. For the conditional probability that he has two daughters given you meet one daughter is identical to the probability that the other one is a daughter, which is 1/2. This change of probability sounds strange since it seems that no new relevant information is given when

---

11) The original version of this puzzle is introduced in Bar-Hillel & Falk (1982).

you meet Smith's daughter.

One may respond that there is a difference in information since the order of two kids is not mentioned in the proposition that at least one of two kids is a daughter while the order is fixed when you actually witness one of them. This response is based upon the following reasoning: Let us call Smith's two kids 'Ari' and 'Bori'. (These are provisional names.) And let 'P' stand for the proposition that Ari is a girl and 'Q' for the proposition that Bori is a girl. The information that at least one of the kids is a daughter is expressed as (P v Q). And the information you receive when one daughter is introduced is either P or Q. If the information is P, it implies, and yet is not implied by, (P v Q). And if the information is Q, it implies, and yet is not implied by, (P v Q) too. Thus, either way, there is a difference in information between the two ways of introducing a daughter.

According to this account, the paradoxical element of this puzzle is only illusionary. But I do not believe that the puzzle is completely explained away. One might be still able to claim that the information you receive when you meet one daughter is (P v Q) too. For, since 'Ari' and 'Bori' are merely provisional names, you are free to assign either one to the girl you meet. Thus, one might argue, it is not right to express the information you receive as P or as Q.

The paradoxical element of this puzzle also lies in the difficulty of drawing a sharp line between the two ways of introducing a daughter in the example. There are many different ways of introduction: A picture of a daughter of his can be presented. Or her name, a note written by her, a pair of shoes bought for her. The proposed account is not able to distinguish the ways of introduction which yield the information (P v Q) from those which yield the information P or the information Q.

Again, the outsider/insider distinction can shed light on understanding the paradoxical structure of the puzzle. You can be an outsider or insider with respect to the experience of meeting a daughter of Smith's (GIRL for short). When you are told by Smith that at least one of his kids is a girl, it is natural for you to be an outsider with respect to GIRL. And when you actually meet one daughter of his, it is natural for you to be an insider. As far as you remain as an outsider with respect to GIRL, the conditional probability in question is 2/3. On the other hand, as far as you remain as an insider with respect to GIRL, it is 1/2. Only if you change your stance from outsider to insider, the probability changes from 2/3 to 1/2, and the experience GIRL provides new relevant information.

The inverse gambler's fallacy[12)]

Suppose you visit Albert, who is very fond of playing with dices. Upon entering you see Albert roll four dices to get four sixes, which makes you believe that it is very likely that Albert has been rolling the dices to get four sixes for a long time prior

12) It is Ian Hacking who claims first that there is such a fallacy. See Hacking (1987). Hacking believes that certain criticisms against Design Argument commit the same fallacy. There are controversies which arguments fall into this category of fallacy. See Leslie (1988).

to your visit. Ian Hacking (2001: 45) claims, however, that you commit the inverse gambler's fallacy in this case. For the probability of Albert's rolling four sixes that you observed is independent of the outcomes of his previous trials. Now consider a slightly different situation: You are told that Albert rolled four sixes last night. Then you think that Albert must have rolled the dices many times to get four sixes. In this case, Hacking argues, you do not commit the fallacy. For a rolling four sixes in 1000 tosses is much more probable than a rolling four sixes in 10 tosses. But it seems that there is no crucial difference between the two cases. Suppose that you watch a video tape that shows Albert rolling four sixes. And you do not know the exact time when the tape is recorded. Is it right for you to assign a high credence in that Albert must have rolled the dices many times before?

The difference between the two cases might be explained in terms of the outsider/insider distinction. Consider the experience of watching Albert's rolling four sixes (FOUR for short). Let 'MANY' stand for the proposition that Albert rolled the dice many times to get four sixes, and 'ONCE' for the proposition that he rolled the dices only once to get four sixes. As far as you remain as an insider with respect to FOUR, you should not assign more credence to MANY than to ONCE. As far as you remain as an outsider with respect to FOUR, on the other hand, your credence in MANY should be higher than in ONCE. If you watch actually Albert roll four sixes, you are forced to be an insider with respect to FOUR. If you watch the video tape, you have more flexibility to choose between insider stance and outsider stance.

6. The Third Child Puzzle

Now back to the original question: How seriously should we regard our future selves' voices? My answer is somewhat different from what Reflection recommends. According to Reflection, we should follow our future selves' opinions. I believe that we can deliberately deny future selves' opinions, without violating Reflection, even though our future selves are very likely to have those opinions on solid grounds. My reasoning depends upon the notions already introduced: agent-part and the outsider/insider distinction. Let us consider the following puzzling situation:

The third child puzzle

Mr. Kim, a middle-aged father with two kids, has no plan as of now for the third child. He is busy with having a great time with his two kids. Furthermore, the responsibility for financial support for his family forces him to judge that it is not thoughtful to have another child. Nevertheless, he often talks with his wife about the third child. They believe that if they were to have the third child, they would even say, "We would be greatly regretful if wave if we did not bring this baby into the world!" Given their health states Mr. and Mrs. Kim have a good chance of getting another baby in a near future if they are willing to do so. The question is, if this is really what they believe and desire, on which sense it is still rational for them to decide not to have the third child?

I am not suggesting that all the parents with two kids would judge the same way. Situations of this type, however, are common-place. This situation is not the one in which an agent is struggling with two conflicting wills. For Mr. Kim surely knows that he does *not* want the third child. Thus, most probably, Mr. Kim's future self under consideration shall not be a *real* future self. And yet he believes that his future self with the third child would believe that the future self is happier than the current self.

Future selves understood as agent-parts can be addressed from two different perspectives. Mr. Kim can be an outsider with respect to the experience of having the third child and then he regards his future self just as one of his current peer members. Or his current agent-part can be an insider and then regards his future self with the third child as if it is a real future extension of the current agent-part. When Mr. Kim is led to imagine that he has the third child, he is forced to change his stance from the outsider to the insider. Then the imagined experience of having the third child offers indexical information.

One of the distinct features of Mr. Kim's case is that the judgment from the insider's perspective is deliberately ignored by the current agent-part of Mr. Kim. That is, Mr. Kim's outsider dominates the insider. There are various reasons for this dominance. He might take it a policy to give special weight to his present experiences of his present agent-part. He might be afraid of unexpected difficulties that could arise if he has another baby. Anyhow, he is refusing to accept the judgment of his future agent-part and yet is not subject to irrationality. Another feature of the third child puzzle is that Reflection seems to be violated and yet it satisfies additional requirements of Reflection: the filtration requirement and the stopping times requirement. Thus, we cannot attribute a violation of Reflection to the failure to satisfy either requirement.[13]

7. Conclusion: Degrees of Indexicality

I have proposed a version of New Information view to solve the Sleeping Beauty problem and other puzzles. According to this view, paradoxical results arise from that the agent-part's stance changes from the outsider to the insider. As far as the agent-part remains as an outsider or as an insider with respect to a specific experience, there shall be no credence change, and Reflection is intact. And if the agent-part changes her perspective from the outsider to the insider, there shall be a credence change which endangers Reflection. I have argued, however, that Reflection could be defended by treating future agent-parts just as one of

[13] I am not claiming that this puzzle is a genuine counterexample of Reflection. Reflection could be saved again unless an agent regards his future selves as his extensions.

peers. Also I have argued that the credence change from shifting stances is based upon new relevant indexical information, by which Temporal Conditionalization is saved.

According to my account, the content of experience might be different depending on whether an agent-part is an outsider or an insider with respect to the very experience. For instance, the content of Beauty's experience that she shall undergo upon awakening is different from that she gets from the outsider stance. When Beauty gets this experience from the insider's perspective, this experience comes to have content that plays an epistemic role in evaluating personal probabilities. It depends upon various circumstantial elements whether an agent should be an outsider or an insider with respect to a certain experience. Having the very experience most naturally forces the agent to take the insider's perspective, but not necessarily.[14)]

It has been argued by many that entertaining indexical beliefs in appropriate circumstances is the key to performing an action at the right time at the right place. It is not entertaining indexical beliefs, however, that plays a crucial role in evaluating credence in events that involves an agent's (potential) experiences. It is rather shifting one's perspectives or imagining from the inside

---

14) Weintraub believes that an agent cannot take indexical information, say 'it is raining now', without being situated at the very moments. This is what is called the 'Limited Accessibility' approach' by Perry, who claims that this view betrays his belief 'in a common actual world' (Perry 1979). According to Perry, indexicality lies in the semantic feature of an agent's belief state. Weintraub, however, believes that indexicality of information lies in the metaphysical feature of the agent's situations. According to my account, indexicality is an epistemic matter.

(potential) experiences. As the two daughter puzzle and others suggest, there shall be no clear cut distinction between the circumstances in which only the outsider's perspective is admissible and the ones in which only the insider's perspective is admissible. Puzzles arise since these two perspectives are equally available. It is a matter of degrees whether one should take the insider's perspective or the outsider's perspective. The third child puzzle indicates that an agent tends to favor his current agent-part over future agent-parts. The dominance of current agent-part, however, is not a matter of irrationality but a matter of empathy or imagining future selves' experiences.

REFERENCES

Arntzenius, Frank.(2003) "Some Problems for Conditionalization and Reflection", *Journal of Philosophy* 100, pp. 356-371.

Bar-Hillel, Maya & Falk, Ruma.(1982) "Some Teasers Concerning Conditional Probabilities", *Cognition* 11, pp. 109-122.

Bostrom, Nick.(2003) "The Mysteries of Self-Locating Belief and Anthropic Reasoning", *Harvard Review of Philosophy* 11, pp. 59-74.

____________(2007) "Sleeping Beauty and Self-Location: a Hybrid View", *Synthese* 157, pp. 59-78.

Christensen, David.(1991) "Clever Bookies and Coherent Beliefs", *Philosophical Review* 100, pp. 229-247.

Dorr, Cian.(2002) "Sleeping Beauty: In Defence of Elga", *Analysis* 62, pp. 292-296.

Elga, Adam.(2000) "Self locating Belief and the Sleeping Beauty Problem", *Analysis* 60, pp. 143-147.

__________.(2007) "Reflection and Disagreement", *Noûs* 41, pp. 478-502.

Gupta, Anil.(2006) *Empiricism and Experience*. Oxford: Oxford University Press.

Hájek, Alan.(2005) "The Cable Guy Paradox", *Analysis* 65, pp. 115-119.

Hacking, Ian.(1987) "The Inverse Gambler's Fallacy: the Argument from Design The Anthropic Principle Applied to Wheeler Universes", *Mind* 97, pp. 331-340.

____________(2001) *An Introduction to Probability and Inductive Logic*. New York, NY: Cambridge University Press.

Leslie, John.(1988) “No Inverse Gambler’s Fallacy in Cosmology”, *Mind* 97, pp. 269-272.

Lewis, David.(2001) “Sleeping Beauty: Reply to Elga”, *Analysis* 61, pp. 171-176.

Maher, Patrick.(1992) “Diachronic Rationality”, *Philosophy of Science* 59, pp. 120-141.

Perry, John.(1979) “The Problem of Essential Indexical Belief”, *Noûs* 13, pp. 3-21.

Schervish, M. J. et al.(2004) “Stopping to Reflect”, *Journal of Philosophy* 101, pp 315-322.

Talbott, William.(1991) “Two Principles of Bayesian Epistemology”, *Philosophical Studie*s, 62, pp. 135-150.

van Fraassen, Bas C.(1984) “Belief and the Will”, *Journal of Philosophy* 81, pp. 235-256.

__________________(1995) “Belief and the Problem of Ulysses and the Sirens”, *Philosophical Studies* 77, pp 7-37.

Walton, Kendall.(1990) *Mimesis as Make-Believe*. Cambridge, MA: Harvard University Press.

Weintraub, Ruth.(2004) “Sleeping Beauty: A Simple Solution”, *Analysis* 64, pp 8-10.

국민대학교 교양과정부

Email: hanskim@kookim.ac.kr

잠자는 미녀의 숙고: 안과 밖

김한승

반프라센의 숙고 원리는 반례에 직면한다는 주장들이 있는데 이 중 두드러지는 것이 잠자는 미녀 문제가 그런 반례라는 주장이다. 엘가는 잠자는 미녀 문제에 대한 자신의 대답이 숙고 원리의 반례를 보여준다고 주장한다. 반면 잠자는 미녀 문제에 대한 루이스의 대답은 숙고 원리를 보존한다. 최근 보스트롬은 숙고의 원리를 유지하면서 엘가의 입장과 루이스의 입장을 절충하려는 혼성 이론을 제시한 바 있다. 그는 엘가의 입장이 '자기 표시 가정'을 전제하고 있으며 이 가정은 옳지 못하다고 주장한다. 하지만 나는 엘가가 보스트롬의 행위자-부분 개념을 이용하면 오히려 보스트롬의 비판을 피할 수 있다고 본다. 또한 잠자는 미녀 문제를 포함하여 몇 가지 확률 관련 퍼즐들을 살펴봄으로써 미래 자아의 견해를 다루는 올바른 입장을 제시한다. 이 새롭게 제안된 입장에 따르면, 행위자의 한 명제에 대한 믿음의 정도에서 중요한 것은, 그가 외부자 입장과 내부자 입장 중 어떤 것을 취하는가 하는 것이다.

주요어: 잠자는 미녀 문제, 숙고 원리, 자기 표시 가정, 지표성, 확률, 퍼즐