PERSPECTIVE DISAGREEMENTS IN SLEEPING BEAUTY PROBLEM

XIANDA GAO

ABSTRACT:I argue perspective disagreements in the setup of Sleeping Beauty Problem(SBP) is logically sound. This suggests double-halving is the correct answer. Counter examples against traditional Halfers and Thirders are also provided. As a result I propose the rejection of both Self-Sampling Assumption(SSA) and Self-Indication Assumption(SIA). Because an observer reasoning as if she is randomly selected from a reference class switched to an outsider's perspective which alters the answer. This could be important in anthropic reasoning.

THE SLEEPING BEAUTY PROBLEM

The problem can be briefly described as follows:

Sleeping Beauty Problem(SBP):

Beauty is undergoing an experiment in which she will be waken up once or twice based on a fair coin toss. If the coin landed on tails (T) she will wake up twice with a memory wipe in between so that the 2 awakenings are indistinguishable. If the coin landed on heads (H) she will not be subjected to any memory wipe so only her first awakening needs to be considered. What should her credence of H be when she wakes up not having memories of a previous awakening.

There are two main camps to the problem. Halfers think the probability should be 1/2 and thirders argue the probability is 1/3. As of right now majority of the literature seems to lean towards the thirders' view. There are some disagreements among halfers as what would the probability of H be after beauty learns that she is indeed in the first awakening. Here I will simply call those who changes their answer to 2/3 as halfers and call whose answers remains at 1/2 as double-halfers.

It is worth pointing out that halfers agrees with the Self-Sampling Assumption(SSA): that all else equal, an observer should reason as if she is randomly selected from the set of all *actually existent* observers. Thirders on the other hand agrees with the Self-Indication Assumption(SIA): that all else equal, an observer should reason as if she is randomly selected from the set of all *possible* observers. With these in mind, let us take a look at the perspective disagreement in SBP.

THE DISAGREEMENT

To see where the disagreement lies I purpose to consider the following problem:

Duplicating Beauty:

Beauty falls asleep as usual. The experimenter tosses a fair coin before she wakes up. If the coin landed on T then a perfect copy of beauty will be produced. The copy is precise enough that she cannot tell if herself is old or new. If the coin landed on H then no copy will be made . The beauty(ies) will then be put into two identical rooms respectively. What should their probability of H be once awake fully knowing the setup.

I imagine many would agree the Duplicating Beauty and the original SBP are equivalent. Both camps have literatures using cloning examples to advance their argument. I will focus on Duplicating Beauty here since it facilitate easier description and discussion on the disagreement. However my argument does not depends on this equivalency. So I will later dedicate a brief section on the procedures leading to the same disagreement in original SBP for anybody suspicious of the equivalency.

Now suppose another person is observing the experiment. He will not be shown the coin toss result nor will he ever be cloned. However after beauty(ies) are put into the two rooms he need to randomly choose one room and enter. Let us call him the selector. Now suppose after entering the random room he sees a beauty inside. Since the chosen room is not empty by very simply bayesian updating his probability of H shall be 1/3.

On the other hand, per halfers, the beauty in the same room should have p(H)=1/2 after waking up. Since her room has equal chance of being selected regardless of H or T the probability remains unchanged when she sees the selector. Now the beauty and the selector has a disagreement about p(H). The disagreement remains when they are free to communicate. Here we have two people giving different probabilities to the same proposition while having the same information, in violation of Aumann's agreement theorem.

One might use this against halfers since thirders' interpretation will not lead to such disagreement. I will address the problems with thirders' reasoning and the pitfall of such agreement later with a counter example. For now I would argue both the selector and beauty are correct in their reasoning and this perspective disagreement is valid.

We can see this perspective disagreement with a frequentist approach. What if the experiment is repeated by a large number, say 1000 times. From the selector's point of view, as long as someone undergoes the experiment and let him choose one out of two rooms it is counted as a repetition. After 1000 repetitions by his count he should expect to find about 750 non-empty rooms, 250 times following H and 500 following T. Consistent with his p(H) of 1/3. From

beauty's point of view each repetition means undergoing the experiment and wakes up again waiting for the selector's choice. So by her count, taking part in 1000 repetitions means she would recall 1000 coin tosses after waking up. In those 1000 coin tosses there should be about 500 of H and T each, about 500 times she sees the selector half following H the other half following T. This is in accordance with her p(H)=1/2. Note the beauty waking up remembering 1000 coin tosses might be created by any one of those coin tosses just as the beauty disagreeing with the selector after the first toss might be a copy. Not being physically the same person should not affect the answer. To see this we can easily modify the Duplicating Beauty problem so the original beauty is always destroyed after each toss with either one or two copies created depending on H or T. In the modified version all resulting beauties are newly created yet it is hard to argue their answer would be different from the unmodified version. Another point worth noting is if all beauties in the experiment, old and new, each recall 1000 repetitions then by the selector's count the number of repetition would be way greater. This is unsurprising since multiple beauties would be created in the process and each of them require additional tosses to complete their 1000 repetitions.

The perspective disagreement can also be shown by betting odds. To see this however beauties' money must be duplicated when she is duplicated. In another word each person's money follow each one's perspective. This way beauty will not gain any information simply by checking her wallet meaning her money mimics her epistemic state which determines her answer. For the selector, the fair payout of a 1-dollar bet on H should be 3 dollars. Remember in the 1000 repetitions , he will see a beauty and enter the bet about 750 times. In which 250 times following H. If he pays 1 dollar to enter the bet each time, a 3 dollar payout on H would make his expected return 0. On the other hand the fair payout of a 1-dollar bet on H should be 2 dollars for beauty. In her 1000 repetitions she will be seeing the selector and entering the bet about 500 times, 250 times following H andT each - a 2 dollar payout will make her expected return 0. The apparent disagreement is because the selector is doubly likely to enter the bet than beauty after T. Since from his perspective seeing any one of the two beauties is the same observation. In both cases he enters the bet. Whereas from beauties' perspective which exact beauty the selector sees leads to different observations. Although it gives her no information about the coin toss it decides whether or not she will enter the bet.

The above reasoning can be easily applied to original SBP. Although the execution would be rather cumbersome. In SBP the selector must lose track of the time himself in order to randomly select a day out of two. This can be done by using a pill that makes him sleep through beauty's first awakening. He must choose between the sleeping pill and a undistinguishable placebo and

go to sleep together with beauty when she first enters sleep. This way when he wakes up he can check if beauty is awake as well. Now he is in the same position as if he has just entered a random room in Duplicating Beauty. From beauty's perspective to repeat the experiment means her subsequent awakenings need to be shorter to fit into her current awakening. For example, if in the first experiment the two possible awakenings happen on different days, then the in the next repetition the two possible awakening can happen on morning and afternoon of the current day. Further repetitions will keep dividing the available time. It would cause each awakening to be increasingly short, making large number of repetitions impractical. However the methodology is still valid, since the length of the awakenings has no effect on beauty's answer. To correctly see the effect of betting odds mentioned above her money should follow her epistemic state as well. Therefore any change to her wallet should be reversed when the relating memory is erased. In summary beauty and the selector would also be having disagreements in SBP with valid reasons.

ARGUMENT AGAINST SELF-INDICATION ASSUMPTION

The above arguments are of little importance if beauty should assigns p(H)=1/3 to begin with. In that case although she does not change her belief upon seeing the selector, unlike her counterpart does, both of them would still agree on p(H)=1/3. This can be explained by SIA, where she shall reason finding herself exist is the same as randomly selecting a potential observer(a room in this case) and then find the observer actually exist(the room is not empty). This put beauty in the same position as the selector, causing them to agree. To point out the weakness in this reasoning, consider the following example:

The 81-Day Experiment(81D):

There is a building with a circular corridor connected to 81 rooms with identical doors. At the beginning all rooms have blue walls. Then a painter randomly selects an unknown number of rooms and paint them red. Beauty would be put into a drug induced sleep lasting 81 day, spending one day in each room. An experimenter would wake her up if the room she currently sleeps in is red and let her sleep through the day if the room is blue. Her memory of each awakening would be wiped at the end of the day. Each time after beauty wakes up she is allowed to exit her room and open some other doors in the corridor to check the colour of those rooms. Now suppose one day after opening 8 random doors she sees 2 red rooms and 6 blue rooms. How should beauty estimate the total number of red rooms(R).

For halfers, waking up in a red room does not give beauty any more information except that R>0. Randomly opening 8 doors means she took a simple random sample of size 8 from a population

of 80. In the sample 2 rooms (1/4) are red. Therefore the total number of red rooms(R) can be easily estimated as 1/4 of the 80 rooms plus her own room, 21 in total.

For thirders, beauty's own red room is treated differently. As SIA states, finding herself awake is as if she chose a random room from all 81 rooms and find out it is red. Therefore her room and the other 8 rooms she checked are all in the same sample. This means she has a simple random sample of size 9 from a population of 81. 3 out of 9 rooms in the sample (1/3) are red. The total number of red rooms can be easily estimated as a third of the 81 rooms, 27 in total.

I believe the above calculations are straightforward. The same numbers would be obtained as the most likely cases in a bayesian analysis with uniform priors. Notice if an outside selector randomly checks 9 rooms and happens to open the exact same 9 rooms beauty knows (her own room and the 8 rooms she checked), he would estimate R=27. Because beauty and the selector has exact same information about the rooms, he would be in disagreement with beauty according to halfers and in agreement with beauty according to thirders even if the two of them are free to communicate. The disagreement/agreement pattern is the same as in SBP.

However there are contradictions in thirders' answer. First of all, before opening any doors if beauty is told that R=21 she should expect to see 2 reds if she opens 8 doors. According to thirders after opening 8 doors and actually seeing the expected number of 2 red rooms beauty must estimate R=27 instead of 21. This change of mind cannot be explained. Secondly, estimating R=27 means if beauty opens another 8 random doors she should expect to see 24/72x8=2.67 red rooms. This means after beauty saw 2 reds in the first 8 random rooms she would expect to see about 3 reds if she choose another 8 rooms. I fail to see how can this be justified. Last but not least, because beauty believes the 9 rooms beauty knows is a fair sample of all 81 rooms, it means red rooms (and blue rooms) are not systematically over- or under-represented it. Since beauty is always going to wake up in a red room, she has to conclude the other 8 rooms is not a fair sample. Red rooms have to be systematically underrepresent in those 8 rooms. This means even before beauty decides which doors she wants to open we can already predict with certain confidence that those 8 rooms is going to contains less reds than average. This supernatural predicting power is a conclusive evidence against SIA and thirders' argument.[1]

---

[1] The argument can also be structured this way. Consider the following three statements:
A: The 9 rooms is an unbiased sample of the 81 rooms.
B: Beauty is guaranteed to wake up in a red room
C: The 8 rooms beauty choose is an unbiased sample of the other 80 rooms.
These statements cannot be all true at the same time. Thirders accept A and B meaning they must reject C. In fact they must conclude the 8 rooms she choose would be biased towards blue. This contradicts the fact that the 8 rooms are randomly chosen.

It is also easy to see why beauty should not estimate R the same way as the selector does. There are about 260 billion distinct combinations to pick 9 rooms out of 81. The selector has a equal chance to see any one of those 260 billion combinations. Beauty on the other hand could only possibility see a subset of the combinations. If a combination does not contains a red room, beauty would never see it. Furthermore, the more red rooms a combination contains the more awakening it has leading to a greater chance for a beauty to select the said combination. Therefore while the same 9 rooms is a unbiased sample for the selector it is a sample biased towards red for beauty.

One might want to argue after the selector learns a beauty has the knowledge of the same 9 rooms he should lower his estimation of R to the same as beauty's. After all beauty could only know combinations in a subset biased towards red. The selector should also reason his sample is biased towards red. This argument is especially tempting for halfers since if true it means their answer also yields no disagreements. Sadly this notion is wrong, the selector ought to remain his initial estimation. To the selector a beauty knowing the same 9 rooms simply means after waking up in one of the red rooms in his sample, beauty made a particular set of random choices coinciding said sample. It offers him no new information about the other rooms.

To make this point clearer, it might be beneficial to understand how people reach to an agreement in an ordinary problem. Consider the following case:

Balls In Urns(BIU):
Suppose there is a urn filled with either 2 blue balls and 1 red ball(BBR) or 2 red balls and a blue ball(BRR) with equal chances. Andy randomly picked 2 balls from the urn and finds one ball of each colour. He correctly conclude the probability of BBR is 1/2, same as the probability of BRR. Afterwards Bob asked for a red ball and was given one, he then randomly picked 1 ball from the 2 remaining balls in the urn and saw a blue one. He correctly concluded the probability of BBR is 2/3. It turns out however Andy and Bob actually saw the exact same 2 balls. The two of them are free to communicate and argue. Suppose both of them are rational can they reach to an agreement? Who should change his answer?

Again we can use a frequentist approach to solve this problem. Suppose the experiment is repeated many times with equal number of BBR and BRR. We can count the total number of occurrences when they both see the same 1 red and 1 blue balls. The relative frequency of BBR and BRR among these occurrences would indicate the correct probability. Both Andy and Bob should have no problem agreeing with this method. Here it becomes apparent that the exact

procedure of the experiment determines whose initial probability is correct and who need to adjust his answer. More specifically how is the red ball given to Bob determined. Scenario 1: Bob is always given the red ball that has been picked by Andy. In this case Andy is correct, Bob should adjust his answer of BBR from 2/3 to 1/2. This is because for both BBR and BRR Andy has the same chance to pick a red and a blue ball. Given the same red ball Andy had picked Bob would have equal chance to pick the same blue ball again regardless the colour of the last ball left. The relative frequency of BBR and BRR given the occurrences would be about the same. Another case would be Scenario 2: Any red ball in the urn can be given to Bob. In this case Bob's initial judgement is correct and Andy would have to change his probability for BBR to 2/3. Because all else equal, Bob is twice more likely to have the same red ball as Andy if there is only 1 red ball in the urn. The relative frequency of BBR to BRR with the occurrences would be 2:1. Since only one out of the two Scenarios can be true Andy and Bob must agree with each other as long as there is no ambiguity about he experiment procedure.

However, if we duplicate Bob (either by cloning or memory wiping) in the case of BRR and give each Bob a different red ball suddenly both Scenarios become true. To Andy the red ball he picked will always be given to Bob. To Bob the red ball given to him can be any one from the urn. Neither person would have reason to adjust his own probability fully knowing the experiment procedure. In this case the the two person having exact same information will remain in disagreement even if they are free to communicate. I believe the parallel between BIU, 81D and SBP is obvious enough to show why the selector, just as Andy, shall not adjust his probability.

ARGUMENTS AGAINST SELF-SAMPLING ASSUMPTION

I think the most convincing argument against SSA was presented by Elga(2000). He purpose the coin toss could happen after the first awakening. Beauty's answer ought to remain the same regardless the timing of the toss. As SSA states, an observer should reason as if she is randomly selected from the set of all actual observers. If the selector randomly choose a day among all waking day(s) he is guaranteed to pick Monday if the coin landed on H but only has half the chance if T. From the selector's perspective clearly a bayesian updating should be performed upon learning it is Monday. A simple calculation tells us his credence of H must be 1/3. As SSA dictates this is also beauty's answer. Now beauty is predicting a fair coin toss yet to happen would most likely land on T. This supernatural predicting power is a conclusive evidence against SSA.

However, if we recognize the importance of perspective disagreement then beauty is not bound to give the same answer as the Selector. In fact I would argue she should not perform a bayesian update base on the new information. This can be explained in two ways.

One way is to put the new information into the frequentist approach mentioned above. In Duplicating Beauties, when a beauty wakes up and remembering 1000 repetitions she shall reason there are about 500 of H and T each among those 1000 tosses. The same conclusion would be reached by all beauties without knowing if she is physically the original or created somewhere along the way. Now suppose a beauty learns she is indeed the original. She would simply reason as the original beauty who wakes up remembering 1000 tosses. These 1000 tosses would still contain about 500 of H and T each. Meaning her answer shall remain at 1/2.

Another way to see why beauty should not perform a bayesian update is to see the agreement/ disagreement pattern between her and the selector. It is worth noting that beauty and the selector will be in agreement once knowing she is the original. As stated earlier, one way to understand the disagreement is after T, seeing either beauty is the same observation for the selector while it is different observations for beauties. This in turn causes the selector to enter twice more bets than beauty. However once we distinguish the two beauties by stating which one is the original the selector's observation would also be different depending on which beauty he sees. To put it in a different way, if a bet is only set between the selector and the original beauty then the selector would no longer be twice more likely to enter a bet in case of T. He and the original Beauty would enter the bet with equal chances. Meaning their betting odds ought to be the same, they must be in agreement regarding the credence of H.

To be specific, the disagreements/agreements pattern can be summarized as follows. If the selector randomly chooses one of the two rooms as described by SIA. Upon seeing a beauty in the room the selector's probability for H will change from 1/2 to 1/3. Beauty's probability remains at 1/2 as described above. The two of them would be in disagreement. Once they learns the beauty is the original, the selector's probability of H increases back to 1/2 from 1/3 by simple bayesian updating while beauty's probability still remains at 1/2. This way the two would be in agreement. The selector can also randomly chooses one beauty from all existing beauty(ies) as described by SSA (here the total number of beauties should be shielded from the selector to not reveal the coin toss result). In this case seeing a beauty gives the selector no new information so his probability for H would remain unchanged at 1/2. On the other hand, from beauty's perspective she is twice more likely to be chosen if there exist only one beauty instead of two. Therefore upon seeing the selector her credence of H would increase to 2/3. The two of them

would also be in disagreement. Once they learn the beauty is the original the selector's credence for H would increase to 2/3 by bayesian updating. Again beauty would not update her probability and it remains at 2/3. This way the two would agree with each other again.

As shown above, for the two to reach an agreement beauty must not perform a bayesian update upon the new information. This holds true in both cases regardless how the selection is structured.

Beauty's antibayesianism, just like the perspective disagreement, is quite unusual. I think this is due to the fact that there is no random event determining which beauty is original/clone. While the coin toss may create a new copy of beauty, nothing could ever turn herself into the copy. The original beauty would eventually be the original beauty. It is simply tautology. There is no random soul jumping between the two bodies. Beauty's uncertainty is because of the structure of the experiment which is *purely* due to lack of information. Compare this to the selector's situation. The event of him choosing a room is random. Therefore learning the beauty in the chosen room is original gives new information about the random event. From his perspective a bayesian update should be performed.Where as from beauty's perspective, learning she is the original does not give new information about a random event, for the simply fact there is no random event to begin with. It only gives information about her own perspective. So she should not performing a bayesian update as the selector did.

DISCUSSIONS

With the above arguments in mind we can clearly see the importance of perspectivism in SBP. It does not matter whether the selector follows SSA or SIA, his answer could not always correctly reflect beauty's. Once beauty switch to selector's perspective her answers would change. Therefore it is important for us to consciously track our reasoning process to make sure it contains no change of perspective. As shown above, beauty's credence of H should be 1/2 when she wakes up and remains at 1/2 once learned it is Monday, aka double-halfers are correct. In another word, learning she exists does not confirm scenarios with more observers, learning she is the first does not confirm scenarios with less observers. Applying the same logic means we should reject Doomsday Argument while disagree with the Presumptuous Philosopher.

There are more interesting implication for perspective reasonings. One thing worth noting is that disagreements can also arise among beauties themselves. Consider DB with 2 repetitions for all beauties. Beauty wakes up in a room remembering taking part of 2 coin tosses. Now the experimenter tells her that including herself there are currently 3 beauties exist in total. Very

soon the beauty figured out to get three beauties after 2 rounds means the first coin toss must be T. One of the resulting beauties would then experience another T while the other resulting beauty would experience another H. For ease of discussion lets randomly label the two beauties after the first toss T1 and T2 since they both experienced a T. Suppose T1 experience a H and T2 experience another T on the second round. We can label the the beauty who experienced a H asT1H1, and randomly label the two beauties experienced another T as T2T1 and T2T2. By indifference principle beauty shall reason she is equally likely to be T1 or T2 at her first awakening. Because T1H1 is a direct product of T1, her probability of being T1H1 at second awakening is the same as being T1 at the first awakening which is 1/2. Because T2T1 and T2T2 are random labels for the direct products of T2, she shall reason the combine probability of her being either T2T1 or T2T2 is the same as being T2 at the first awakening. So her probability of being T2T1 and T2T2 must be 1/4 each. All three beauties would reach to the same conclusion since they have the same information. Notice here they would be in disagreement regarding their probability of being T1H1. Each of them would think herself is twice more likely to be T1H1 (1/2) than the other two beauties(1/4 each).This disagreement might seems alarming. However it is also valid. As discussed above, in problems involving duplications people with the same information can have different probabilities. The reason this disagreement seems more suspicious is because the resulting beauties appears to be in symmetrical positions thus should not disagree with each other. However this symmetry is only valid from an selector's perspective. For the selector, any resulting beauty has an equal chance of being picked. In another word T1H1, T2T1 and T2T2 are only different in names during a random selection process. Therefore it is correct a randomly chosen beauty has equal chance of being T1H1, T2T1 or T2T2(1/3). From beauty's perspective however such symmetry does not exist. The three labels, T1H1, T2T1 and T2T2, are not just different in name. If the coin landed H then she would be labeled H1 for certain. If the coin landed on T she would be randomly labeled T1 or T2. Therefore from beauty's perspective being labeled as H1 is twice likely to happen than being label T1.