Reflections on Sleeping Beauty

Frank Arntzenius

1. Introduction

Adam Elga (2000) presents a puzzle, the 'Sleeping Beauty' puzzle, which concerns the updating of belief when a person, Sleeping Beauty, finds something out about her temporal location in the world. He claims that in such cases, even though she apparently only learned something about her temporal location in the world, and nothing about the world per se, she should nonetheless change her degrees of belief in what the world is like. And Elga claims that in so doing she will violate Bas van Fraassen's 'Reflection Principle'. (See van Fraassen 1984 and van Fraassen 1995.)

After presenting Elga's argument I will present an alternative argument which has as its conclusion that Sleeping Beauty should not change her degrees of belief. I will then argue that neither of these arguments by itself is compelling, that one should distinguish degrees of belief from acceptable betting odds, and that some of the time Sleeping Beauty should not have

Analysis 62.1, January 2002, pp. 53–62.

definite degrees of belief in certain propositions. Finally I will argue that the Sleeping Beauty puzzle has to do with cognitive malfunction rather than with the updating of self-locating beliefs, but that nonetheless the updating of self-locating beliefs is interestingly different from the updating of ordinary beliefs about the world.

2. *Sleeping Beauty*

Some researchers are going to put Sleeping Beauty, SB, to sleep on Sunday night. During the two days that her sleep will last the researchers will wake her up either once, on Monday morning, or twice, on Monday morning and Tuesday morning. They will toss a fair coin Sunday night in order to determine whether she will be woken up once or twice: if it lands heads she will be woken up on Monday only, if it lands tails she will be woken up on Monday and Tuesday. After each waking, she will be asked what her degree of belief is that the outcome of the coin toss is heads. After she has given her answer she will be given a drug that erases her memory of the waking up; indeed it resets her mental state to the state that it was in on Sunday just before she was put to sleep. Then she is put to sleep again. The question now is: when she wakes up what should her degree of belief be that the outcome was heads?

Answer 1: Her degree of belief in heads should be 1/2. It was a fair coin and she learned nothing relevant by waking up.

Answer 2: Her degree of belief in heads should be 1/3. If this experiment is repeated many times, approximately 1/3 of the awakenings will be heads-awakenings, i.e. awakenings that happen on trials in which the coin landed heads.

3. *Why Sleeping Beauty should change her mind*

Elga argues that Answer 2 is the correct answer. His argument is as follows. When SB wakes up she is certain that she is in one of three predicaments:

H_1: Heads and it is Monday
T_1: Tails and it is Monday
T_2: Tails and it is Tuesday

Let P represent the degrees of belief that SB should have when she awakens. Suppose that when she wakes up she learns that the outcome is tails. By the symmetry of the situation she ought then to have degree of belief 1/2 that it is Monday, and degree of belief 1/2 that it is Tuesday. Thus $P(T_1/T_1 \text{ or } T_2) = 1/2$. And hence $P(T_1) = P(T_2)$.

Alternatively suppose that when she wakes up she learns that it is Monday. It seems that then she still ought to have degree of belief 1/2 that the outcome was heads. In fact Elga gives an apparently compelling argu-

ment for this conclusion, by considering a slightly different case. Suppose that the coin is tossed on Monday night rather than Sunday night. If it comes up heads SB will not be woken up on Tuesday, if it comes up tails she will be woken up on Tuesday. Whether the coin is tossed on Sunday night or Monday night ought to make no difference to her degrees of belief. But it is clear that if she learns that it is Monday, and she knows that a fair coin is to be tossed on Monday night, then she ought to have degree of belief 1/2 that it will come up heads. Now, learning that it is Monday amounts to learning that she is in H_1 or T_1. Thus $P(H_1/H_1 \text{ or } T_1) = 1/2$. And this entails that $P(H_1) = P(T_1)$. Altogether we therefore have $P(H_1) = P(T_1) = P(T_2) = 1/3$.

The surprising consequence of this argument is that she started out on Sunday with a degree of belief 1/2 in heads, and by merely waking up she has changed her mind: she has switched her degree of belief to 1/3, despite the fact that she received no new information about the world. (Here 'no new information' means: 'no new non-self-locating information'.) This violates Bas van Fraassen's 'Reflection Principle', which entails that if a person P is now certain that tomorrow P will have degree of belief x in R, while suffering no 'cognitive mishaps' between now and tomorrow, then P ought to now have degree of belief x in R. Elga's explanation of these surprising consequences is that in Sleeping Beauty type cases 'you have gone from a situation in which you count your temporal location as irrelevant to the truth of H, to one in which you count your own temporal location as relevant to the truth of H.' (145). Frankly speaking, this 'explanation' still leaves me puzzled.

4. *Why Sleeping Beauty should not change her mind*

I will now quickly argue that Sleeping Beauty should not change her mind, that is, that she should maintain degree of belief 1/2 in heads when she wakes up, and then spend some time defending this argument against objections. The argument is simple: SB gains no new information that is relevant to the outcome of the coin toss, she should not violate conditionalization and/or Reflection, and hence her degree of belief upon waking up should be what it was on Sunday, namely 1/2. (For a more elaborate and explicit version of this argument see Lewis 2001.) A first worry about this argument is that if it is compelling, then Elga's argument had better be faulty. I will postpone discussion of that issue to the next section. A second worry is that in the long run 2/3 of SB's awakenings will be tails-awakenings. Thus, if she bets according to her degree of belief of 1/2, she can be expected to lose money against a bookie, and she and the bookie know this in advance. I will argue that the resolution of this worry is that she should bet at odds that differ from her degrees of belief, and that this

is perfectly consistent with standard Bayesian lore. I will argue for this by means of some examples.

Suppose that a bookie comes to you on Sunday and allows you to bet at 1:2 odds on the toss of a fair coin on Monday, where you can nominate which side you are betting on. Here 1:2 odds means that you win $1 if it lands on the side you nominate and you lose $2 if it lands on the other side. Clearly you should not accept.

Next, the bookie gives you a slightly different offer. He still offers you a bet at 1:2 odds. But now the absolute value of the bet depends on the outcome of the toss in the following way: if the toss lands heads the bet is $1 versus $2, if the coin lands tails the bet is $2 versus $4. Now you should be indifferent as to whether you accept the offer, and if you do accept the offer you should nominate tails. For if you do that and it lands heads, then the bet is $1 versus $2, so you loose $2, while if it lands tails the bet is $2 versus $4 and you win $2. So you loose $2 or win $2, and each is equally likely.[1]

Though this case does not seem as puzzling as the Sleeping Beauty case, there is something puzzling about it. If you accept the offer on Sunday, then there is a 50% chance that you are committing yourself to bet A on Monday, and there is a 50% chance that you are committing yourself to bet B on Monday, where each of bets A and B are bad bets according to your degrees of belief. The puzzling thing is that this commitment is nonetheless not a bad commitment. This may seem less puzzling when you realize that this commitment can be re-described as a choice, made by you on Sunday, to accept bet #1 of the following two bets that you are offered:

bet#1: you bet on tails: you win $2, or loose $2
bet#2: you bet on heads: you win $1, or loose $4.

Nonetheless it remains the case that if you accept the offer on Sunday you are committing yourself to one of two possible bets on Monday, each of which is bad according to your degrees of belief.

The next example is the Sleeping Beauty example, except that each time that SB is woken up she is not asked for her degrees of belief in tails, instead she is asked whether she is willing to accept a bet at 1:2 odds on tails. SB should be indifferent, since the structure of this case is exactly the same as the case above. In the variable stakes bet, by direct stipulation, the potential gains and losses are multiplied by 2 when it is tails; in the Sleeping Beauty case the potential gains and losses are multiplied by 2 when it is tails because SB is twice asked the same question, and is bound to give the same

[1] In email exchanges I have learned that Jamie Dreyer has invented essentially the same example (his 'variable stakes casino example') before I did.

answers. This suggests that SB's degrees of belief in heads upon waking up should remain unchanged at 1/2, but that nonetheless upon waking up she should accept any bets on tails at odds of 1:2 or better. One might still be worried: doesn't betting at odds that are bad according to one's degrees of belief violate standard decision theory? Let me give yet another example to explain why it needn't.

Let us make three changes to the Sleeping Beauty case. In the first place let us assume that SB is offered a bet on tails at 2:3 odds (rather than 1:2 odds) each time she wakes up. Secondly, there is another person, Dormant Belle, DB, who is placed in exactly the same situation as SB is: she is offered the very same bets (on the very same coin toss) that SB is offered, and her memory will be erased after she accepts or declines a bet etc. Thirdly, the payoff structure is altered. SB gets the gains or losses of her first bet, and of DB's last bet, if there is one. DB gets the gains or losses of her first bet, and the gains or losses of SB's last bet, if there is one. DB and SB don't know each other, but they each know that there is another person in the same situation as they are, and they each know what the payoff structure is. Each is supposed to care only about the profits or losses that they themselves make. What should SB (and DB) do?

The answer is that SB upon waking should not accept a bet offered at 2:3. The profits or losses that SB makes on DB's last bet, if there is one, are independent of SB's choice whether to accept a 2:3 bet when SB wakes up. So SB needs only to decide for or against the acceptability of a single bet (on Monday). She either gains $2 or loses $3 on that bet, depending on the outcome of a toss of a fair coin. So she should not accept the bet. DB, if she is sensible, will reason the same way. It is a prisoners' dilemma.

Let us next change the example a little bit. Suppose that DB is SB's identical twin sister and that both believe that whatever the one decides, the other will decide the same. Should each accept the bet? Well, that depends on what they regard as the correct decision theory. On a simple-minded version of evidential decision theory they ought both to accept the bet, since if both accept the bet, each either wins $4 or loses $3, and each is equally likely. On the other hand, according to causal decision theory they should not accept. For holding fixed what the other does, no matter what the other does, each should expect to do better (by $0.50 per trial in the long run) by not accepting.

Now let us return to the standard Sleeping Beauty case. Let us start by supposing that SB is an evidential decision theorist, and that SB accepts that her agreeing to a particular bet on one particular awakening is good evidence that she will agree to it upon the other awakening, if there is one. Thus, her acceptance of a bet on a particular awakening has two beneficial consequences if in fact the coin lands tails: she will win that particular bet, and she will make that bet again and win it again, upon her other awak-

ening. So if SB accepts evidential decision theory she should accept a bet on tails at 2:3 odds, even though her degree of belief in tails is 1/2.

In contrast let us suppose that SB is a causal decision theorist. Clearly SB's acceptance of a bet on Tuesday does not cause her acceptance of that bet on Monday. Moreover it is implausible to claim that her acceptance of a bet on Monday causes her to accept that bet on Tuesday, since it is more plausible to claim that both acceptances have a common cause, namely SB's mental state on Sunday. So let us suppose that SB does not believe that her acceptance of a bet any one day *causes* her acceptance of such a bet on the other day. Let us furthermore suppose that SB accepts a rather strict version of causal decision theory, namely one according to which a 'dependency hypothesis' (see Lewis 1981) consist of a listing of the causal effects of SB's possible actions. Given SB's acceptance of such a causal decision theory, and given her belief that acceptance of a bet on any one day does not cause her acceptance on any other day, she should, upon waking up, not accept a bet at odds that differ from her degrees of belief.

As a final example, let us suppose that SB believes that any acceptance of any bet on any one day is counterfactually connected to such an acceptance on the other day, i.e. let us suppose that SB accepts the claim 'If I were to accept this bet now, then I would accept it upon any other such waking'. And suppose that SB accepts 'counterfactual decision theory', i.e. a causal decision theory according to which counterfactual dependencies are the relevant dependencies. Then SB will accept bets at odds that differ from her degrees of belief, e.g. a bet on tails at odds 2:3 even though her degree of belief in tails is 1/2.

So, one can hold on to van Fraassen's Reflection Principle, and to Bayesian updating by conditionalization, as long as one does not adhere to a strict version of causal decision theory. Does this mean that we have an argument from the premiss of 'van Fraassen Reflection', or from the premiss of 'Bayesian conditionalization', to the conclusion that strict causal decision theory is false? That would be surprising! Or should one instead conclude that causal decision theory is correct, that SB's degree of belief in heads should be 1/2, but that Sleeping Beauty is one of those strange cases where a causal decision theorist is punished for her rationality, and predictably so? That would be interesting! But before jumping to conclusions, let us take a step back.

5. *Why Sleeping Beauty may, or may not, change her mind*

If the coin lands tails on Monday night SB's degrees of belief will be reset to what they were on Sunday night. Thus, if the coin lands tails, then SB, from a Bayesian point of view, will have a cognitive malfunction: she will violate conditionalization. Since she knows this in advance (and all the

time) the obvious question is: what should she do to avoid 'damage' as much as she can? Before discussing what the best damage control strategy is for SB, let me, in the light of the noted cognitive malfunction, explain what is wrong with the views expressed in the previous two sections.

What is wrong with Elga's argument that SB should change her mind upon waking up? Well, Elga's argument assumes that if SB were to learn that it is Monday she should arrive at her new degrees of belief by conditionalizing the degrees of belief she has when she wakes up. And it assumes that, were she to learn that it is tails, she should arrive at her new degrees of belief by conditionalization from the degrees of belief she has when she wakes up. This argument would be fine if indeed SB could be a good Bayesian throughout. But the conclusion of Elga's argument is that SB should not be a good Bayesian, since she should violate Reflection, and should violate conditionalization on the transition from Sunday evening to Monday morning after wake-up. Moreover, it is inevitable that, if the coin lands tails, then on the transition from Monday to Tuesday SB will violate conditionalization, since her degrees of belief will be artificially reset. What's more, if SB follows Elga's advice for her degrees of belief and she happens to be an evidential decision theorist she can expect to loose money. So Elga's argument is not compelling.

What is wrong with my argument that SB should not change her mind upon waking up? In essence, it has the same problems as Elga's argument. I argued that SB should not change her mind, by assuming the validity of van Fraassen's Reflection Principle, and by assuming that her degrees of belief upon waking up come from her degrees of belief on Sunday by conditionalization. Since there is nothing (relevant) to conditionalize upon when she wakes up, her non-self-locational degrees of belief must remain the same upon waking up. This would be a fine argument if SB could always update by conditionalization. But she can't. Moreover if she accepts my argument and happens to be a strict causal decision theorist, she can expect to lose money. So my argument is not compelling.

So let us look afresh at the question as to what SB should do. As should be obvious, there are several options. For instance, even while admitting that Elga's argument and my argument for the above two views are inconclusive, SB could still decide to adopt one or the other of those degrees of belief on different grounds. But let me also mention some other options, before lurching towards resolution.

When she wakes up she could do exactly what she would do on Monday morning if she had no worries about cognitive malfunctions, namely update her location belief to 'it is now Monday', and leave her degrees of belief in heads the same. If she does this she will have no cognitive damage on Monday: on Monday morning she will have exactly the beliefs that a person P would have who does not have to worry about possible cognitive

malfunction. Note that this procedure is not the same procedure as advocated in the previous section. For on the view advocated in the previous section SB should not have degree of belief 1 in 'it is Monday' upon waking up, while on the view currently under consideration she should. The current view has the advantage that SB suffers no cognitive damage whatsoever on Monday. But then, of course, on a Tuesday morning wake-up, if there is one, she would have severe cognitive damage: she would have the very same degrees of belief as she had on Monday, and those are not the degrees of belief that a person who functions cognitively perfectly throughout, would have on a Tuesday wake-up. One can think of many such schemes, and each will include some cognitive damage as compared to a perfectly functioning individual. So let us look more closely at the issue of damage control in order to get some grip on a preferred scheme.

On Sunday Sleeping Beauty can contemplate all the bets that she might be offered on future days. Suppose that she is forced to accept bets when she wakes up, for example, because a gun is pointed at her head, and she is told that she has to state the odds at which she is indifferent as to which side of the bet she takes. Now suppose that these bets are offered by a bookie who adjusts the offered bets according to the outcome of the toss, which he knows. Then, of course, the bookie can make SB lose money, and there is no relevant damage control possible. But suppose this is done by a bookie who does not adjust the offered bets to the outcomes. Then, given SB's degrees of belief on Sunday, she ought to be able to figure out what the best strategy should be with respect to such possible future bets.

Now, let us ensure that SB does not learn anything about which day it is from the bet that she is offered, by assuming that SB is certain that whatever bet she is offered she will be offered the same bet on Monday as on Tuesday. Then it should be clear which bets SB should accept upon waking up. For she knows that in the long run, if the situation were repeated, there will be 1/3 heads-awakenings, and 2/3 tails-awakenings. Since she will always accept or reject the same bets, she has good reason to accept only bets on tails at 1:2 odds (gain 1 if it is tails, lose 2 if it is heads), or better.

So she knows what bets to accept upon waking up. What about her degrees of belief upon waking up though? Well, we have seen above that if SB has degree of belief 1/3 in heads, and accepts strict causal decision theory, she will indeed accept bets on tails at 1:2 odds or better. But we have also seen that if SB has degree of belief 1/2 in heads, and she accepts strict evidential decision theory, then she will also accept bets on tails at 1:2 odds or better. And there are some other possible combinations that will have her accept bets on tails at 1:2 odds or better. One possible suggestion therefore is that SB should, upon waking up, simply set her degrees of belief to whatever they need to be, given the decision theory and depen-

dencies that she accepts, in order for her to find acceptable bets on tails at 1:2 odds or better.

However, it seems rather odd that SB's degrees of belief would depend on the decision theory that she accepts. Surely if she changes her mind about which decision theory is correct she should not thereby be forced to change her epistemic state with respect to heads. Surely changing her mind about decision theory does not entail changing her mind as to what the world is like with respect to outcomes of coin tosses. Thus it seems more plausible to say that her epistemic state upon waking up should not include a definite degree of belief in heads. She should of course know which bets she should accept and which ones she should not, but her epistemic state does not include a natural candidate for a context-independent degree of belief in heads. In fact her epistemic state upon waking up is best described by saying that she believes that she is in the situation described in the Sleeping Beauty story. Not to have a definite degree of belief in heads might be strange, but it might be the best that she can do given the forced irrationality that is inflicted upon her.

Finally, as should be clear by now, on my view self-locating learning plays no relevant role in the Sleeping Beauty case. The real issue is how one deals with known, unavoidable, cognitive malfunction. But I still think that there is something interesting and new about self-locating learning. For instance, suppose that one knows exactly what the world is like and exactly where one is in the world. Then let some time lapse. If one is not looking at a clock, one will typically no longer be certain what time it is, and hence where one is in the world, and one would not normally call this a cognitive malfunction. Nonetheless, this change in one's cognitive state is not due to conditionalization. Moreover, it includes a transition from total certainty to uncertainty. That is a new form of cognitive change incompatible with conditionalization, which is not really 'learning', but nonetheless in a rather literal sense consists of 'updating'. This means that standard Bayesian lore, after all, does have to be modified in order to deal with self-locating beliefs.

However, that is not the main moral of the Sleeping Beauty story. The main moral of that story is that in the face of forced irrational changes in one's degrees of belief one might do best simply to jettison them altogether.[2]

Rutgers University
New Brunswick, NJ 08901-2882, USA
arntzeni@rci.rutgers.edu

[2] Many thanks to Adam Elga, Brad Monton, Ned Hall, Jamie Dreyer and Bas van Fraassen for some useful email exchanges on the Sleeping Beauty puzzle.

References

Elga, A. 2000. Self-locating belief and the Sleeping Beauty problem. *Analysis* 60: 143–47.
Lewis, D. 1981. Causal decision theory. *Australasian Journal of Philosophy* 50: 5–30.
Lewis, D. 2001. Sleeping Beauty: reply to Elga. *Analysis* 61: 171–76.
van Fraassen, B. 1984. Belief and the will. *Journal of Philosophy* 81: 235–56.
van Fraassen, B. 1995. Belief and the problem of Ulysses and the Sirens. *Philosophical Studies* 77: 7–37.