Sleeping Beauty and the forgetful Bayesian

BRADLEY MONTON

1. Consider the case of Sleeping Beauty: on Sunday she is put to sleep, and she knows that on Monday experimenters will wake her up, and then put her to sleep with a memory-erasing drug that causes her to forget that waking-up. The researchers will then flip a fair coin; if the result is Heads, they will allow her to continue to sleep, and if the result is Tails, they will wake her up again on Tuesday. Thus, when she is awakened, she will not know whether it is Monday or Tuesday. On Sunday, she assigns probability 1/2 to the proposition H that the coin lands Heads. What probability should she assign to H on Monday, when she wakes up?

Adam Elga (2000) argues that the answer is 1/3. As Elga (citing Ned Hall) points out, though, this answer violates Bas van Fraassen's (1984, 1995a) Reflection Principle, which entails that

ANALYSIS 62.1, January 2002, pp. 47–53. © Bradley Monton

> Any agent who is certain that she will tomorrow have credence *x* in proposition *R* … ought *now* to have credence *x* in *R*. (Elga 2000: 146)

Elga takes the Sleeping Beauty example to provide a counter-example to Reflection, since on Sunday Beauty assigns probability 1/2 to *H*, and she is certain that on Monday she will assign probability 1/3. I will show that there is a natural way for van Fraassen to defend Reflection in the case of Sleeping Beauty, building on van Fraassen's treatment of forgetting. This will allow me to identify a lacuna in Elga's argument for 1/3. I will then argue, however, that not all is well with Reflection: there is a problem with van Fraassen's treatment of forgetting. Ultimately I will agree with Elga's 1/3 answer. David Lewis (2001) maintains that the answer is 1/2; I will argue that cases of forgetting can be used to show that the premiss of Lewis's argument for 1/2 is false.

2. It has been claimed (by for example Talbot 1991) that cases of forgetting can be seen as violations of the Reflection Principle. For example, van Fraassen had a croissant for breakfast today; he currently assigns a very high probability to the proposition *C* that he had a croissant this morning. Van Fraassen does not normally eat croissants, though, and he recognizes that a year from now he will not remember having a croissant. If he believes that he will assign a low probability to *C* a year from now, then he should assign a low probability to *C* now, on pain of violating Reflection. Van Fraassen rejects the antecedent of that conditional:

> a year from now I should say, when asked about this, that I have definite opinions about the rate I was eating croissants per week or per month in that earlier time, but no opinion (i.e., totally vague opinion) about any particular day therein. This will automatically satisfy Reflection of course. (van Fraassen 1995a: 22)

Van Fraassen's answer is that a year from now his doxastic state should not be modelled by a single probability function, but by a set of probability functions. This set contains at least one member where the probability assigned to *C* is 0, and at least one member where the probability assigned to *C* is 1. Thus, a year from now van Fraassen will assign the totally vague opinion [0, 1] to *C*. This satisfies the General Reflection Principle:

> My current opinion 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. (van Fraassen 1995a: 16)

No matter what probability van Fraassen assigns to *C* now, it will lie in the range [0, 1]; this is why Reflection is automatically satisfied.

An ideal Bayesian agent – one who structures her opinion in the form of precise numerical probabilities and changes it solely by conditionalization on evidence – automatically satisfies the General Reflection Principle (van Fraassen 1995a: 17). Such a Bayesian never forgets. Van Fraassen takes Reflection to be a more general normative principle, which one should satisfy even if she is not an ideal Bayesian agent. One who strives to be an ideal Bayesian agent, but sometimes forgets, might wonder what normative principles she should apply in situations where she is forgetful. Van Fraassen can give a partial answer: forgetful Bayesians should satisfy Reflection.

Sleeping Beauty can be viewed as a forgetful Bayesian. On Sunday, Beauty knows what day it is, but on Monday she does not know. In the croissant case, what van Fraassen forgets is a proposition about his eating a croissant on a particular day. Thus, when van Fraassen forgets, the class of possible worlds which are compatible with his beliefs (the doxastically accessible worlds) expands to include worlds where he did not eat a croissant on that day. (This modal analysis comes from Lewis 1986: 27–29.) In the Sleeping Beauty case, though, what Beauty forgets is her place within a particular world. On Sunday, the class of possible individuals-at-times who might, for all Beauty believes, be Beauty (Beauty's doxastic alternatives) only includes individuals-at-times for whom it's true that it's Sunday. If Beauty didn't forget, then on Monday Beauty's class of doxastic alternatives would only include individuals-at-times for whom it's true that it's Monday. But in fact, on Monday Beauty's class of doxastic alternatives includes individuals-at-times for whom it's true that it's Monday and individuals-at-times for whom it's true that it's Tuesday. Thus, Beauty does forget.[1]

Other than her forgetting, Beauty is 'a paragon of probabilistic rationality' (Lewis 2001); this (I maintain) is what makes her a forgetful Bayesian. Van Fraassen believes that Beauty should satisfy Reflection, and utilizing van Fraassen's defence of Reflection in the case of forgetting gives Beauty a way to do so. On Sunday, Beauty recognizes that on Monday she will not know whether it's Monday or Tuesday. If she knew it was Monday, she would assign probability 1/2 to H; if she knew it was Tuesday, she would assign probability 0 to H. Thus, her doxastic state on Monday

[1] For those who are not convinced, consider the following rather different sort of argument. Suppose that Beauty falls asleep on Sunday at 11 p.m., is awoken Monday at 9 a.m. and put to sleep at 10 a.m., and if Tails is awoken Tuesday at 9 a.m. and put to sleep at 10 a.m. Instead of dividing time into 24 hour days, we can divide time into 24 hour Tays. The first Tay, called Tay1, starts at 11 a.m. Sunday and ends at 11 a.m. Monday. When Beauty falls asleep on Sunday at 11 p.m., she believes it is Tay1. When she wakes up on Monday at 9 a.m., she no longer believes that it is Tay1. Thus, she has forgotten what Tay it is. My subsequent analysis will still go through: if she knew it was Tay1, she would assign probability 1/2 to H, while if knew it was Tay2, she would assign probability 0 to H.

should be represented by a set of probability functions such that her opinion for H is the vague [0, 1/2]. It follows that she can assign probability 1/2 to H on Sunday and satisfy Reflection.

Elga does not just *say* that the answer to the question 'what should Beauty's opinion about H be on Monday, when she wakes up?' is 1/3; he argues for it. But there is a lacuna in Elga's argument, such that it cannot be used to show that the [0, 1/2] answer is incorrect. Elga (2000: 144) starts his argument by saying 'Let P be the credence function [Beauty] ought to have upon first awakening.' He then shows that $P(H_1) = P(T_1) = P(T_2)$, where H_1 is the proposition that the coin lands Heads and it is Monday, T_1 is Tails and Monday, and T_2 is Tails and Tuesday. Since these are mutually exclusive and exhaustive, $P(H_1) = 1/3$. The problem here is that Elga simply *assumes* that Beauty's doxastic state on Monday can be modelled by P, which is a single probability function; this assumption automatically rules out vague opinions. Without this assumption, Elga's argument doesn't get off the ground, since each step in his argument relies on there being a single probability function.

3. I will now argue that the justification for the [0, 1/2] answer is flawed, because there is a problem with van Fraassen's defence of Reflection in the case of forgetting. Go back to the croissant case: a year from now van Fraassen assigns opinion [0, 1] to C. Suppose that he then discovers that a stalker (a Fraassen fan) has been videotaping him for years; the videos have a time/date stamp and van Fraassen can confirm that they are reliable. Van Fraassen sees himself eating a croissant for breakfast on a video stamped with today's date. When he conditionalizes on this evidence, we would expect him to end up assigning a high probability to C. But in fact this will not happen. The standard way of doing conditionalization with vague opinions, which van Fraassen (1989: 194; 1998: 215) endorses, is to conditionalize on each probability function in the set of probability functions which represents one's opinion.[2] But conditionalization cannot raise the probability of a proposition that has probability 0, nor lower the probability of a proposition that has probability 1. Thus, after conditionaliza-

[2] For a very different model of updating with vague opinions, see Walley 1991, especially 217–27. This would not be a good model for van Fraassen to use in the case of Sleeping Beauty though. One reason for this is that Beauty having opinion [0, 1/2] for H entails, according to Walley (1991: 3), that no matter what the payoff, there is no bet for H Beauty would be willing to accept. (Beauty would be willing to accept a bet against H as long as the payoff were equal to or better than 1 to 1.) But in fact it would be unreasonable for Beauty, when she wakes up, to reject a bet for H with payoff 1,000,000 to 1, for example. Perhaps there is some alternative to both Walley's approach and the standard approach that would be successful, but I have not been able to come up with one.

tion on the videotape evidence, van Fraassen will still assign opinion [0, 1] to C.[3]

The same problem arises in the Sleeping Beauty case. We would expect that, if Beauty were to learn that it is Monday, then her opinion for *H* would be non-vague and non-zero. (Elga says 1/2; Lewis says 2/3.) If Beauty assigns opinion [0, 1/2] to *H* on Monday when she wakes up, though, then after conditionalization her opinion for *H* will continue to be vague over an interval which includes 0, or her opinion will be a sharp 0.

Van Fraassen is not a Bayesian; he could reply that since Beauty is a paragon of probabilistic rationality, the way for her to update on the evidence that it is Monday is *not* via conditionalization. In fact, van Fraassen (1995b, 1995c) has invented a model of belief revision which allows one to change from zero to non-zero probability assignments via updating which is not conditionalization. Van Fraassen admits, though, that this model 'is mobilized in rather serious epistemic contexts, and also does not answer to a good deal of common usage with lower standards' (1998: 219). More importantly, in this model the only way to change from a zero to a non-zero probability assignment is by updating on a proposition that was assigned zero prior probability. In the croissant case, though, while van Fraassen's prior probability that a stalker has been videotaping him is surely very low, it need not be zero.

I conclude that Reflection should not be a normative principle when forgetting is involved. Thus, it would not be surprising if Elga were correct in claiming that Beauty violates Reflection, since Beauty in essence forgets what day it is.

4. Though I believe that the 1/3 answer is correct, I have not directly argued for it. To provide further evidence that 1/3 is correct, I will now argue that cases of forgetting can be used to show that the premiss of Lewis's argument for 1/2 is false.

Lewis's premiss is:

> Only new relevant evidence, centred or uncentred, produces a change in credence; and the evidence ($H_1 \vee T_1 \vee T_2$) is not relevant to HEADS versus TAILS. (Lewis 2001: 174)

Uncentred evidence is evidence about what world is actual, while centred evidence is evidence about where one is within a world.

Consider Fred, who wakes up after a very long sleep in a windowless room and picks up his clock, which reads 11.00 a.m. Fred believes that his clock is reliable, so he assigns a high probability to the proposition *L* that

[3] This criticism is analogous to one of the criticisms raised by Hájek 1998 and Monton 1998 for van Fraassen's analysis of agnosticism.

it is now light outside. Fred then drops his clock, and in the excitement of its shattering Fred forgets whether the clock read 11 a.m. or 11 p.m. If it were 11 p.m. it would not be light outside, so Fred lowers the probability he assigns to *L*. Let *A* be the proposition that within the past half-minute or so, the clock read 11 a.m. Let *P* be the proposition that within the past half-minute or so the clock read 11 p.m. Learning $(A \vee P)$ is, according to Lewis, gaining new evidence. But that's not what produces the change in credence. If upon dropping the clock Fred simply added the belief $(A \vee P)$ to his set of beliefs, nothing would change, since Fred already believes *A*. What is necessary for the change in credence is *forgetting A*. Thus, it is not the case that only new relevant evidence can produce a change in credence; the losing of evidence can do so as well.

5. I will close by assuming that the 1/3 answer is correct, and examining what justifies Beauty's change in credence from Sunday to Monday.

Elga says that what justifies the change in credence is that (thinking of yourself as Beauty):

> you have gone from a situation in which you count your own 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*. (Elga 2000: 145)

Contra Elga, on Sunday Beauty *does* count her temporal location as relevant to the truth of *H*. If Beauty did not believe that it was Sunday, and instead thought that it was Tuesday, Beauty would assign 0 to *H*. If Beauty assigned probability 1/7 to each of the seven propositions that today is Monday, today is Tuesday, and so on, then she would assign to *H* probability $(1/2 \times 6 + 0)/7 = 3/7$. Thus, Beauty's belief that it is Sunday is relevant to the probability she assigns to *H*.

So what does justify the change in credence? On Sunday Beauty knows her temporal location, and because of that knowledge she assigns probability 1/2 to *H*. If Beauty were to know her temporal location on Monday, she would assign probability 1/2 to *H*. This suggests that it is Beauty's forgetting her temporal location which justifies her change in credence.[4]

University of Kentucky
Lexington, KY 40506, USA
bmonton@uky.edu

[4] I thank Sandy Goldberg, Brian Kierland, David Lewis and Bas van Fraassen for helpful discussion.

References

Elga, A. 2000. Self-locating belief and the Sleeping Beauty problem. *Analysis* 60: 143–47.

Hájek, A. 1998. Agnosticism meets Bayesianism. *Analysis* 58: 199–206.

Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, D. 2001. Sleeping Beauty: reply to Elga. *Analysis* 61: 171–76.

Monton, B. 1998. Bayesian agnosticism and constructive empiricism. *Analysis* 58: 207–12.

Talbot, W. J. 1991. Two principles of Bayesian epistemology. *Philosophical Studies* 62: 135–50.

van Fraassen, B. 1984. Belief and the will. *Journal of Philosophy* 81: 235–56.

van Fraassen, B. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.

van Fraassen, B. 1995a. Belief and the problem of Ulysses and the sirens. *Philosophical Studies* 77: 7–37.

van Fraassen, B. 1995b. Fine-grained opinion, conditional probability, and the logic of belief. *Journal of Philosophical Logic* 24: 349–77.

van Fraassen, B. 1995c. Science, probability, and the proposition. *PSA* 1994 *Volume* 2: 339–48, ed. D. Hull, M. Forbes and R. M. Burian. East Lansing, MI: Philosophy of Science Association.

van Fraassen, B. 1998. The agnostic subtly probabilified. *Analysis* 58: 212–20.

Walley, P. 1991. *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

ANALYSIS 62.1, January 2002, pp. 53–62. © Frank Arntzenius