# Diachronic Dutch Book Arguments for Bounded Rationality, case study: *Sleeping Beauty*

Alistair Isaac

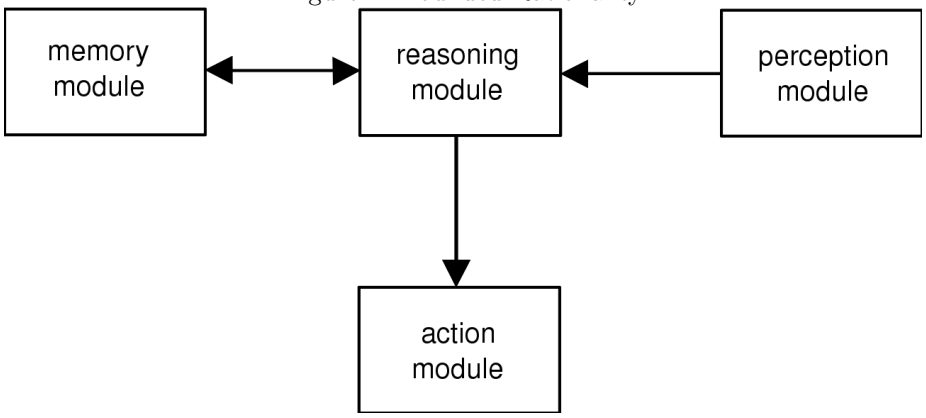
May 19, 2008

Some actions are more sensative than others to the number of times they are performed. The action of putting a coin in a piggy bank, for example, will produce the effect of $n$ coins in the bank if it is performed $n$ times. Consider, however, a TV set with dedicated on and off buttons. The action of pressing the off button will produce the effect of turning the TV off if it is performed $n$ times, so long as $n > 0$. In the case of the piggy bank, $n$ is a variable which must be taken into account when calculating the effect, while in the case of the TV set, once the threshold $n = 1$ has been reached, $n$ can be discounted in calculating the effect. There are even situations where $n$ plays a more complex role in producing an effect. Consider the case of a light switch which toggles between on and off. If the act of flipping the light switch has been performed $n$ times and we wish to calculate the net effect, we must know both i) the initial state of the switch, and ii) whether $n$ is odd or even. Of course, usually we can simply observe the effects of our actions, and we do not need to explicitly calculate them (and thus we can avoid wasting energy pushing the off button on the TV repeatedly, or undoing our intent by flipping the light switch after the lights have reached the desired state). However, first Game Theorists, and then Philosophers, have spent some effort examining a particular contrived decision theoretic problem involving an agent who passes through a number of temporally distinct, yet indistinguishable states. If the agent must chose an action to perform at each state, the relationship between the number of times that action is performed and its success conditions can play a crucial role in the agent's decision making process.

Situations in which there is some impediment to the flow of information between the agent's decision making, or reasoning powers and between his powers of perception, memory, or action have been called cases of *Bounded Rationality* [*fig. 1*].[1] When considering such cases from the decision theoretic standpoint, we stipulate that the agent is aware of the structure of his uncertainty. In the case of the Absent Minded Driver ([Pi97], [Ru98]), for example, a man driving home from a bar is uncertain about whether he has made his turn yet or not,

[1] The expression "Bounded Rationality" has been used in a number of distinct ways. The sense in which it is used here conforms most closely to the decision theoretic literature, *e.g.* Rubinstein (1998) *Modeling Bounded Rationality*.

Figure 1: Bounded Rationality




and finds it impossible to distinguish from surrounding landmarks which is the case. Crucial to the analysis of the problem, however, we stipulate that the man knows ahead of time (before leaving the bar) that he will arrive in this situation. Thus, we consider the agent's reasoning faculty to be intact, and subject to the usual norms of rationality; it is just that his memory or perceptive faculties have been disturbed. Only by stipulating that the agent knows ahead of time precisely how his memory will be disturbed can we make the problem precise from a decision theoretic standpoint.

In the philosophical literature, these issues have been discussed as variants of the Sleeping Beauty problem ([El00]). Beauty is the subject of an experiment wherein a binary random event occurs (say, a fair coin is flipped), and then, ignorant of the outcome, Beauty goes to sleep Sunday night. If the outcome of the random event is Heads, she will be awoken once (on Monday), if the outcome is Tails, she will be awoken twice (on Monday and Tuesday), but after each awakening, her memory will be erased before she is put back to sleep. So, Beauty's uncertainty parition will contain three indistinguishable states: [Heads & Monday], [Tails & Monday], and [Tails & Tuesday]. Elga's original question concerning this setup was "when [Beauty] is first awakened, to what degree ought [she] to believe that the outcome of the coin toss is Heads?" (143). It should be noted that this question makes no sense for two reasons: first, the three awakenings in Beauty's uncertainty partition are indistinguishable, so from the standpoint of Beauty as a reasoning agent, there is no "first" awakening, there are only awakenings within the uncertainty partition. Second, as Ramsey so elegantly argues, "the kind of measurement of belief with which probability is concerned ... is a measurement of belief *qua* basis of action" ([Ra26], 171). In the Beauty problem, no available actions have been stipulated (unlike in the case of the Absent Minded Driver, where he must decide to turn or not).

One may easily imagine appending some structure of possible actions to the Sleeping Beauty problem in order to measure her appropriate degree of belief

within the uncertainty partition. The obvious choice, perhaps, is to offer Beauty a bet:

> The old-established way of measuring a person's belief is to propose a bet, and see what are the lowest odds which he will accept. This method I regard as fundamentally sound; but it suffers from being insufficiently general . . . ([Ra26], 172)

If bets can be placed against one such that one always loses, the set of bets is called a Dutch Book. The strategy of arguing that a particular method for assigning probabilities is rational because it avoids a Dutch Book was introduced by de Finneti, 10 years after Ramsey's above proposal. Is Ramsey correct in his complaint that betting is "insufficiently general"? Well, in some ways he is obviously correct, as there are plenty of real world situations in which one might refuse to bet. Furthermore, the diminishing marginal utility of money (as Ramsey discusses) introduces a vagueness into the metric. However, we are concerned here with a particular type of situation for which Dutch Book arguments may be insufficiently general, namely cases of Bounded Rationality. Our solution here is simple, however. We replace the notion of betting with one of betting "as if." To see how this might work, let's consider some Dutch Book arguments for a generalized form of Sleeping Beauty.

There are several obvious ways to generalize Sleeping Beauty. One way ([Whi06]) is to allow Beauty to assign any probability to the initial binary random event (this will break the symmetry introduced by stipulating that it be a *fair* coin toss). So, instead of calling them Heads and Tails, let's call the outcomes $H$ and $T$ and allow $P(H)$ to vary arbitrarily, stipulating that $P(T) = 1 - P(H)$. A second way to generalize Sleeping Beauty ([Di07]) is to change the number of days Beauty is awoken if $T$ from 2 to $n$. Although this has not yet been proposed in the literature, it is natural to extend this generalization further and allow both the number of wakings if $H$ and the number of wakings if $T$ to be arbitrary. We designate the number of $H$ awakenings by $H_A$ and the number of $T$ awakenings by $T_A$. Now, more elaborate examples will also be special cases of our discusion. Suppose, for instance, the experimentors role a fair die. If the die comes up 6 or 1, they will awaken Beauty 53 indistinguishable times, if the die comes up 2, 3, 4, or 5, they will awaken her 72 indistinguishable times. When she awakes within the uncertainty partition, what should Beauty's degree of belief in the outcome {6 or 1} be?

Within this generalized Sleeping Beauty setup, let us consider the appropriate generalization of the Dutch Book argument given in Hitchcock (2004). Hitchcock imagines that a cunning bettor is put to sleep along with Beauty; this ensures that he is in the same epistemic state as her. Both before they are put to sleep the first time, and within the uncertainty parition, the bettor can make bets with Beauty concerning the outcome of the coin toss. Although the memories of both Beauty and the bettor are erased, Hitchcock assumes that the bets made within the partition persist somehow through time:

> The [bettor], like Beauty, will awaken having no idea whether it

> is the first or second awakening ... Thus he must sell the same bet to Beauty whenever they both wake up. Under this arrangement, the [bettor] will end up selling two follow-up bets to Beauty if they wake up together twice; this will happen precisely if the outcome of the coin toss is tails. ([Hi04], 412)[2]

But remember, it is a crucial part of the entire setup that awakenings within the partition are indistinguishable. Thus, if the bets made in the partition are to persist through time, they must be notated somewhere. Beauty and the cunning bettor who participates in the experiment with her cannot be exchanging cash as the different amounts in their respective pockets would give them information about where in the uncertainty partition they were located. Thus, we must assume an additional party, the banker. The banker need not be a person, it may be some mechanical betting device, or simply a storage device where bets written down on paper can be locked away until after the experiment. Returning to our opening examples, the banker may even be a simple piggy bank into which slips of paper notating bets are stuffed (so long as the piggy bank were constructed in a clever fashion such that no clues could be obtained about whether or not any slips were already in it). With these caveats, we can construct a Hitckcock-style Dutch Book argument for generalized Sleeping Beauty.

> *Generalized Sleeping Beauty 1*: A random procedure generates one of two outcomes, call them $H$ and $T$. Beauty assigns the probability $P(T)$ to $T$ and $P(H) = (1 - P(T))$ to $H$. If the outcome is $H$, Beauty will be awoken $H_A$ indistinguishable times. If the outcome is $T$, Beauty will be awoken $T_A$ indistinguishable times. A bettor may place bets with Beauty both before she is put to sleep the first time, and within the uncertainty partition. The bettor is also unable to distinguish states within the uncertainty partition, and does not know the outcome of the coin toss. A banker will store the bets for Beauty and the bettor; money does not change hands unless the banker has stored the bet. Within the uncertainty parition, the banker can store bets until after the experiment without revealing any information to Beauty or the bettor about their location within the partition. Thus, the bettor can pick two bets to place, one which he may place once before Beauty is put to sleep, the other which he may only place $H_A$ times if the coin came up heads and $T_A$ times if the coin came up tails.
>
> *Claim*: in order to avoid a Dutch Book, once Beauty is inside the uncertainty partition, she must bet *as if* her belief in $H$ is $\frac{P(H)H_A}{P(H)H_A+P(T)T_A}$.

*Proof*: Without loss of generality, we assume that $T_A > H_A$. Suppose first that once in the information partition, Beauty bets as if her belief in heads is

[2] I have changed Hitchcock's terminology to conform to that of the main text. Instead of "bettor," he uses "bookie."

Table 1: Dutch Book Strategy for GSB1 [r *greater than suggested*]

| bets | payoff | cost | if H | if T |
|---|---|---|---|---|
| 1 | $\frac{1}{(1-P(T))}\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right)$ if T | $\frac{P(T)}{(1-P(T))}\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right)$ | $-\frac{P(T)}{(1-P(T))}\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right)$ | $\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}$ |
| 2 | $T_A$ if H (each) | $rT_A$ (each) | $(1-r)\,T_AH_A$ | $-r\,(T_A)^2$ |
| comb'd | | | $(1-r)\,T_AH_A-\frac{P(T)}{(1-P(T))}\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right)$ | |
| payoffs | | | | $\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right)-r\,(T_A)^2$ |

Table 2: Dutch Book Strategy for GSB1 [r *less than suggested*]

| bets | payoff | cost | if H | if T |
|---|---|---|---|---|
| 1 | $\frac{1}{(1-P(H))}\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right)$ if H | $\frac{P(H)}{(1-P(H))}\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right)$ | $\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}$ | $-\frac{P(H)}{(1-P(H))}\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right)$ |
| 2 | $H_A$ if T (each) | $(1-r)\,H_A$ (each) | $-(1-r)\,H_A^2$ | $rH_AT_A$ |
| comb'd | | | $\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right)-(1-r)\,H_A^2$ | |
| payoffs | | | | $rH_AT_A-\frac{P(H)}{(1-P(H))}\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right)$ |

$r > \frac{P(H)H_A}{P(H)H_A+P(T)T_A}$. Then the cunning bettor places bets against her as in Table 1. Before they are put to sleep, the bettor offers Beauty a bet to win $\frac{1}{(1-P(T))}\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right)$ if the outcome is $T$. Beauty pays her fair price of $\frac{P(T)}{(1-P(T))}\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right)$. Once inside the uncertainty partition, the bettor offers Beauty a bet to win $T_A$ if the outcome is $H$. Now, the fair price for this bet is $rT_A$, yet the bet will actually be placed $H_A$ times if the outcome is $H$ and $T_A$ times if the outcome is $T$. Thus, Beauty's net win from this bet if $H$ will be $(1-r)\,T_AH_A$ and her net loss if tails will be $r\,(T_A)^2$. If we total up the results of these bets, we see that if the outcome is $H$, Beauty's total will be $(1-r)\,T_AH_A - \frac{P(T)}{(1-P(T))}\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right)$, and if the outcome is $T$, her total will be $\left(\frac{P(H)H_AT_A^2}{P(H)H_A+P(T)T_A}\right) - r\,(T_A)^2$. Let $e = r - \frac{P(H)H_A}{P(H)H_A+P(T)T_A}$. Then, some simple manipulation shows that if the outcome is $H$, Beauty loses $eT_AH_A$, and if the outcome is $T$, Beauty loses $eT_A^2$.

Consider now the case where Beauty bets as if her belief is $r < \frac{P(H)H_A}{P(H)H_A+P(T)T_A}$. Then the cunning bettor places these bets with her in accordance with Table 2. Before they are put to sleep, the bettor offers Beauty a bet to win $\frac{1}{(1-P(H))}\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right)$ if the outcome is $H$. Beauty pays her fair price of $\frac{P(H)}{(1-P(H))}\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right)$. Once inside the uncertainty partition, the bettor offers Beauty a bet to win $H_A$ if the outcome is $T$. Now, the fair price for this bet is $(1-r)\,H_A$, yet the bet will actually be placed $H_A$ times if the outcome is $H$ and $T_A$ times if the outcome is $T$. Thus, Beauty's net loss from this bet if $H$ will be $(1-r)\,H_A^2$ and her net win if $T$ will be $rH_AT_A$. If we total up the results of these bets, we see that if the outcome is $H$, Beauty's total will be $\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right) - (1-r)\,H_A^2$, and if the outcome is $T$, her total will be $rH_AT_A - \frac{P(H)}{(1-P(H))}\left(\frac{P(T)H_A^2T_A}{P(H)H_A+P(T)T_A}\right)$. Now, let $e = \frac{P(H)H_A}{P(H)H_A+P(T)T_A} - r$. Then, after some calculations we see, if the outcome is $H$, Beauty loses $eH_A^2$, and if the outcome is $T$, Beauty loses $eH_AT_A$. *QED.*

The first thing to note here is that if we replace all the variables with the relevant values from the original Sleeping Beauty problem ($P\,(H) = P\,(T) = \frac{1}{2}$, $H_A = 1$, and $T_A = 2$), we see that Beauty should bet "as if" the outcome Heads $= \frac{1}{3}$ while within the uncertainty partition. But is it correct to draw the conclusion that Beauty should change her degree of belief in the outcome $H$, or is it more appropriate to say that, in this instance, she should bet *as if* the outcome were $H$? Notice that the appropriate amount for Beauty to bet is a function of *both* her initial probability assignment *and* the number of indistinguishable awakenings given each outcome. Beauty uses her knowledge of the initial probability assignment plus her knowledge of the uncertainy partition to determine the proper betting procedure. Can Beauty always resist a Dutch Book if she bets in this manner? The answer is "no."

If the behavior of the banker is different, then Beauty should not bet in accordance with the setup in *GSB1.* Consider, for example, a different type

of banker, one that behaves more like the off button on the stove. We might describe this as the case of the chalkboard banker.

> *Generalized Sleeping Beauty 2*: A random procedure generates one of two outcomes, call them $H$ and $T$. Beauty assigns the probability $P(T)$ to $T$ and $P(H) = (1 - P(T))$ to $H$. If the outcome is $H$, Beauty will be awoken $H_A$ indistinguishable times. If the outcome is $T$, Beauty will be awoken $T_A$ indistinguishable times. A bettor may place bets with Beauty both before she is put to sleep the first time, and within the uncertainty partition. The bettor is also unable to distinguish states within the uncertainty partition, and does not know the outcome of the coin toss. A banker will store the bets for Beauty and the bettor; money does not change hands unless the banker has stored the bets. The banker does not enter the uncertainty partition with Beauty and the bettor, however. Thus, although he can store their first bet immediately (*i.e.* before Beauty and the bettor are put to sleep), he can only store the outcome of whatever bets are made inside the uncertainty partition once Beauty and the bettor emerge. Furthermore, within the partition, a single tiny blackboard is available for storing information about bets. When Beauty and the bettor are put to sleep, the details of their initial bet remain on the blackboard, so its blankness would give no indication that that instance of awakening were the first. Furthermore, since both parties are aware of the behavior of the banker, Beauty is free to insist the blackboard be erased prior to the placing of any bet.
>
> *Claim*: in order to avoid a Dutch Book, once Beauty is inside the uncertainty partition, she must bet *as if* her belief in $H$ is $P(H)$ (and always insist that the blackboard be erased before placing any bet).
>
> *Proof*: obvious.[3]

In *GSB2*, Beauty must still calculate correct betting procedure from her knolwedge of the initial probabilities plus her knowledge about the structure of the uncertainty parition and the behavior of the banker. Since she knows the banker behaves in a different manner, however, her calculation of how to bet has produced a different answer. This is analogous to the case of deciding how to act for actions of different types. *The question of whether to drop a coin in the*

[3] To see this, simply note that only a single bet will survive the behavior of Beauty and the bettor within the uncertainty partition. Thus, any standard Dutch Book argument against arbitrarily changing probabilities would apply. Note also that if we interpret $H_A$ and $T_A$ not as number of awakenings, but as number of times a bet is placed, we can simply apply our rule from *GSB1*: $\frac{P(H)H_A}{P(H)H_A+P(T)T_A} = \frac{P(H)1}{P(H)1+P(T)1} = \frac{P(H)}{1} = P(H)$. *QED.*

*piggy bank, and that of whether to push the button on the stove may have quite different answers in the same epistemic setup.*

If one continues to reject the talk of betting "as if," and to insist that these are simply two different problems, with two different rules about how Beauty should change her beliefs, one should consider this third form of generalized Sleeping Beauty, combining the betting procedures of *GSB1* and *GSB2*:

> *Generalized Sleeping Beauty 3*: A random procedure generates one of two outcomes, call them $H$ and $T$. Beauty assigns the probability $P(T)$ to $T$ and $P(H) = (1 - P(T))$ to $H$. If the outcome is $H$, Beauty will be awoken $H_A$ indistinguishable times. If the outcome is $T$, Beauty will be awoken $T_A$ indistinguishable times. Two bettors A and B, with corresponding bankers A' and B', will place bets with Beauty before they enter the uncertainty partition and within the partition. All three, Beauty, A, and B, will traverse the same epistemic states. Bettor A and his banker A' will place and record bets with Beauty in accordance with the Hitchcock setup (as in *GSB1*). Bettor B and his banker B' will place and record bets with Beauty in accordance with the chalkboard method (as in *GSB2*).
>
> *Claim*: in order to avoid a Dutch Book, once Beauty is inside the uncertainty partition, she must bet with A *as if* her belief in $H$ is $\frac{P(H)H_A}{P(H)H_A+P(T)T_A}$, and she must bet with B *as if* her belief in $H$ is $P(H)$ (and always insist that the blackboard be erased before placing any bet with B).
>
> *Proof*: obvious.

In *GSB3*, we see that maintaining that the Dutch Book arguments of *GSB1* and *GSB2* characterize rational *belief change* for Beauty lands us in an inconsistency, for we seem forced to say both that Beauty has changed her belief and she has not. Alternately, we may avoid inconsistency, but court absurdity, by insisting that Beauty's belief in $H$ should change to $\frac{P(H)H_A}{P(H)H_A+P(T)T_A}$, but she should bet *as if it had not* against bettor B.

By moving to the general instance of Sleeping Beauty, we open up the space for a more detailed consideration of the relationship between *types of actions* and the structure of Beauty's uncertainty partition. I believe much of the structure here is obscured by the symmetries in the original Sleeping Beauty problem. The stipulations that i) the initial random event be a fair coin toss; ii) the number of Tails awakenings be twice the number of Heads awakenings, and iii) Beauty awakes on Monday no matter what the outcome of the toss have distracted some philosophers from the essence of the problem, namely the interaction between number of indistinguishable events within the partition and types of actions Beauty may perform. Of course, there is some space here for disagreement, but those who remain concerned with the question of degrees of belief in the absence of action would do well to revisit Ramsey's discussion of the difficulties in assigning numerical values to such ([Ra26], section 3, esp. 169-72). Once we introduce actions into the problem, we can see clearly that it is not Beauty's

beliefs which change, but rather her method for calculating correct behavior. These methods must change for actions which are sensitive to the number of times they are performed in different ways. We have suggested above both a piggy-bank style bettor and a stove button style bettor, but we may easily imagine devices for storing bets of the light switch variety, or which depend upon the number of occurrences in an even more complex fashion. In such cases, then, a Dutch Book argument becomes a testing ground for calculating correct action, rather than correct belief.

It might be worth noting that a confusion analogous to that in the philosophical literature was already present in the Game Theory literature in 1997. Piccione and Rubinstein identify two methods for assigning beliefs in a problem analogous to Sleeping Beauty (example 5 of [Pi97]). One is analogous to our betting procedure in *GSB1* (which they call "consistency"), the other analogous to the betting procedure described in *GSB2* (which they call "$Z$-consistency"). They note that "if the decision maker adopts $Z$-consistency then he is exposed to a sort of 'money pump"' ([Pi97], 14); this "money pump" is just a Hitchcock-style Dutch Book. An important distinction here, however, is that Piccione and Rubinstein are already discussing consistency of beliefs with actions. Thus, although they come down on the side of "belief change," that decision, in their case, is indexed by the relevant actions. Thus, when they note that "the paradoxical flavor of the absentminded driver example is unaffected by the type of consistency we adopt" ([Pi97], 14), they are speaking of a distinct paradox from that in the philosophical literature. In Elga (2000), the apparent paradox is that one's beliefs should change when one has received no new information. In Piccione and Rubinstein (1997) the paradox is that one's plan of how to act should change although one has entered a state one was certain one would pass through. Aumann, *et al.* respond quite forcefully that "Absent-mindedness and imperfect recall, while interesting, entail no time inconsistency or paradox" ([Au97], 120). The force of Aumann's argument is that it is perfectly consistent for one to plan before reaching that state to act in a manner different from how one would act at present. This is analogous to our "bet as if" language. From her knowledge of the uncertainty partition she knows she will enter, the agent can calculate the correct action to perform *once she gets there*. Beauty may calculate that betting *as if* the coin came up heads with probability $\frac{1}{3}$ is the appropriate action once she wakes inside her uncertainty partition, but this is not in tension with her belief that the coin toss is fair. Rather, Beauty has used her total knowledge of the situation to compute correct action, a computation performed just as easily outside the partition as within it.

# References

[Au97] Aumann, Robert, Sergiu Hart, and Motty Perry (1997) "The Forgetful Passenger," in *Games and Economic Behavior* 20; 117-120.

[Di07] Dieks, Dennis (2007) "Reasoning about the Future: Doom and Beauty," in *Synthese* 156; 427-439.

[El00] Elga, Adam (2000) "Self-locating Belief and the Sleeping Beauty Problem," in *Analysis* 60.2; 143-7.

[Hi04] Hitchcock, Christopher (2004) "Beauty and the Bets," in *Synthese* 139; 405-20.

[Pi97] Piccione, Michele and Ariel Rubinstein (1997) "On the Interpretation of Decision Problems with Imperfect Recall," in *Games and Economic Behavior* 20; 3-24.

[Ra26] Ramsey, Frank (1926) "Truth and Probability," in Ramsey, 1931, *The Foundations of Mathematics and other Logical Essays*, Ch. VII, ed. by R. B. Braithwaite. New York: Harcourt, Brace and Company. 156-198. [*1999 electronic edition*]

[Ru98] Rubinstein, Ariel (1998) *Modeling Bounded Rationality*, MIT Press.

[Whi06] White, Roger (2006) "The Generalized Sleeping Beauty Problem: A Challenge for Thirders," in *Analysis* 66:2; 114-9.