MINIMIZING INACCURACY FOR SELF-LOCATING BELIEFS

by Brian Kierland and Bradley Monton
University of Missouri at Columbia and University of Kentucky

Penultimate version of paper forthcoming in *Philosophy and Phenomenological Research*.
Please do not cite this version.

June 16, 2003

Abstract

One's inaccuracy for a proposition is defined as the squared difference between the truth value (1 or 0) of the proposition and the credence (or subjective probability, or degree of belief) assigned to the proposition. One should have the epistemic goal of minimizing the expected inaccuracies of one's credences. We show that the method of minimizing expected inaccuracy can be used to solve certain probability problems involving information loss and self-locating beliefs (where a self-locating belief of a temporal part of an individual is a belief about where or when that temporal part is located). We analyze the Sleeping Beauty problem, the duplication version of the Sleeping Beauty problem, and various related problems.

1. Introduction

According to Bayesianism, an agent represents her opinion via a probability function over propositions, and updates her opinion by conditionalizing on propositions representing new evidence. While Bayesianism is a powerful method for representing the dynamics of partial belief, it does not have the resources to handle cases of information loss. The act of forgetting cannot be modelled as conditionalization on a proposition representing new evidence.[1] This then leads to the question: how should an agent who wants to be a Bayesian but undergoes information loss modify her opinion? This paper will give a partial answer to that question.

There are two types of information loss that could occur: an agent could lose information about which possible world she is in, and an agent could lose information about where she is in the world spatiotemporally. (Throughout this paper, we treat an agent as a temporal part of an individual.) We will focus on the latter sort of information loss. Specifically, we will analyze the Sleeping Beauty problem, the duplication version of the Sleeping Beauty problem, and various related problems. The method of *minimizing inaccuracy* can be used to solve all these probability problems.

2. Minimizing Inaccuracy

Some preliminaries: an uncentered proposition is a set of possible worlds. A centered proposition is a set of possible temporal parts of individuals.[2] A centered world is represented by a

[1]For further discussion of this point see for example Monton 2002.

[2]While we are using the terminology of temporal parts, we do not intend to commit ourselves to a particular metaphysical view about persistence. What we say could be translated into language compatible with an endurance theory of persistence, albeit sometimes awkwardly. For example, instead

truth-value assignment $W(X_i)$ of 1s and 0s to all centered and uncentered propositions X_i.

Let one's credences (or subjective probabilities, or degrees of belief) be represented by the probability function P. If one's credences were completely accurate, then one's probability function would match W. The inaccuracy of one's credence for proposition X can be measured by the following quadratic-loss rule, called the *Brier score* (Brier 1950):

$$S(X) = [W(X) - P(X)]^2$$

For a set of *n* propositions $\mathbf{X} = (X_1, \ldots X_n)$, the Brier score is given by

$$S(\mathbf{X}) = ?_i \, 1/n \, S(X_i)$$

The virtues of minimizing inaccuracy (that is, minimizing one's Brier score) are perhaps intuitively obvious. Consider the framework of full belief, where the three doxastic options are full belief, full disbelief, and suspension of belief. In this framework, the natural epistemic norm connected with accuracy of belief is one which gives positive marks for full belief when the proposition is true and positive marks for full disbelief when the proposition is false. What about suspension of belief? Well, this accuracy norm will also give negative marks for full belief when the proposition is false and negative marks for full disbelief when the proposition is true. Thus, there is an epistemic risk in opting for either full belief or full disbelief. A consequence of the accuracy norm, then, is that suspension of belief is epistemically most appropriate when the evidence does not warrant taking such an epistemic risk.

Now consider the framework of partial belief, where the doxastic options are continuum-many, represented by the real numbers between 0 and 1 inclusive. These numbers are one's possible

---

of talking about the opinions of a one-hour temporal part of an agent, we could talk about the opinions of an agent from time *t* to time (*t* + 1 hour).

credences. In this framework, a credence of 1 corresponds to full belief and a credence of 0 corresponds to full disbelief.[3] What about credences in between? To understand them, it is helpful to note the distinction between guesses and estimates (as discussed by for example Jeffrey 1986). When you make a guess, such as how many children someone has, it makes no sense to guess anything other than one of the genuine options. For example, it would make no sense to guess that someone has 3.5 children. But if you are making an estimate, arriving at that value would make sense. This is because, when engaging in estimation, as James Joyce (1998, 587) puts it, "there is no special advantage to being *exactly* right; the goal is to get as *close* as possible to the value of the estimated quantity". Given this distinction, we can say that, in the framework of full belief, one offers guesses as to the truth-value of propositions, while in the framework of partial belief, one offers estimates as to their truth-values.[4]

With this explanation, it should now be clear that the accuracy norm for credences will give higher marks to a credence the closer it is to the actual truth-value of the relevant proposition. Why ever give credences other than 0 or 1, if having a credence of 0 or 1 is the only way to receive the highest marks? The answer parallels the answer as to why someone might suspend belief in the framework of full belief. There is an epistemic risk in opting for a credence of either 0 or 1; while that is

---

[3]What about suspension of belief? There is a controversy about how to represent this doxastic state in a partial belief framework, one which we can safely side-step here. For discussion see Monton 1998, Hájek 1998, and van Fraassen 1998.

[4]Why engage in the practice of estimating the truth-values of propositions? We suspect the answer is largely practical – think of the role credences play in theories of prudential rationality which identify it with maximizing expected utility. We should note, however, that whatever one's reason for engaging in such a practice of estimation, the accuracy norm governing it is still purely epistemic. Thus, we are not claiming that Joyce's (1998) attempted purely epistemic justification of probabilism is not actually purely epistemic.

the only way to get the highest marks, it's also the only way to get the lowest marks. A consequence of the accuracy norm for credences, then, is that less extreme credences are more appropriate when the evidence does not warrant taking the epistemic risk involved in opting for more extreme credences.

Given this kind of accuracy norm for credences, why measure inaccuracy using Brier scores? One commonly noted reason is that the rule that one should minimize one's Brier score is a *proper scoring rule* (Savage 1971, 787-8, 793). A scoring rule is a method for quantitatively establishing how accurate an agent's credences are. A proper scoring rule is a scoring rule that does not give the agent an incentive to change her credences solely in order to get a better score. For example, suppose that a weather forecaster thinks that there is an 80% chance of rain tomorrow. There are some scoring rules such that the forecaster could recognize that his score would be worse if he reported a credence of 0.8 for rain tomorrow than if he reported some other credence (assuming that his score is based on whatever credence he reports). Proper scoring rules do not provide any such perverse incentives.

We believe another reason, partly connected to the previous one, for measuring inaccuracy using Brier scores is that it successfully captures the fact that the accuracy norm for credences places an epistemic risk on opting for extreme credences. By comparison, note that what is called the *linear scoring rule*, which measures the inaccuracy of a credence for X by $|W(X) - P(X)|$, has the result that one should always opt for a credence of 0 or 1.[5] More exactly, following the epistemic goal of minimizing one's expected inaccuracy (and so maximizing one's expected accuracy) when inaccuracy is measured this way leads one to opt for a credence of 0 or 1. (For a proof of this claim see Selten

[5]The only exception is the special case where one begins with a credence of 0.5. In that situation, one will decide that one can minimize one's expected inaccuracy by opting for *any credence*.

1998, 47.) Thus, since the linear scoring rule does not capture an important feature of the accuracy norm for credences, it should be rejected.[6]

We will now discuss this notion of *expected* inaccuracy. Unless one knows W(X), one cannot know an agent's inaccuracy for X. One can, however, calculate the agent's expected inaccuracy, if one knows the expected value of W(X) – that is, if one knows what the chances are that X is true.[7] For example, suppose Alice assigns probability 1/2 to the proposition H that a particular coin lands Heads.[8] Suppose that the coin is actually biased in favor of Heads, such that it lands Heads 2/3 of the time. (In other words, 2/3 of the time W(H) = 1, and 1/3 of the time W(H) = 0.) Alice's expected inaccuracy for H is

$$S_E(H) = 2/3\ (1 - 1/2)^2 + 1/3\ (0 - 1/2)^2$$

$$= 1/4.$$

If Alice were to assign probability 2/3 to Heads, her expected inaccuracy would be lower:

$$S_E(H) = 2/3\ (1 - 2/3)^2 + 1/3\ (0 - 2/3)^2$$

$$= 2/9.$$

---

[6]Patrick Maher (2002) defends the linear scoring rule to some extent, in the course of arguing against Joyce (1998). But Maher does not seem to realize that the linear scoring rule has important drawbacks, such as its not being a proper scoring rule.

[7]Here the reader can substitute her preferred account of chances which aren't solely subjective, such as frequencies, propensities, or objective chances. A concept of this sort is needed, for example to distinguish between coins which are actually fair and coins which are actually biased (regardless of anyone's subjective probabilities regarding these coins). Subjective probability will still play a role: for example, when an agent fully believes that a coin is fair, she assigns subjective probability 1 to the proposition that the frequency/propensity/objective chance of Heads is 1/2.

[8]Here and elsewhere, we are assuming that an agent's credences are synchronically coherent – that is, at any particular time, the agent's credences obey the probability axioms.

One can easily verify that the 2/3 answer minimizes expected inaccuracy.

Suppose that there are two temporal parts (of some specified duration) of Alice which both have an opinion about the coin flip. There are two ways to calculate Alice's expected inaccuracy in this situation: we can calculate her expected total inaccuracy or her expected average inaccuracy. For example, if she assigns probability 1/2 to H during both time intervals, then her expected total inaccuracy is

$$S_{ET}(H) = 1/4 + 1/4 = 1/2,$$

while her expected average inaccuracy is

$$S_{EA}(H) = 1/2\ (1/4 + 1/4) = 1/4.$$

On either way of calculating expected inaccuracy, one gets the result that Alice minimizes her expected inaccuracy when she assigns 2/3 to H both times. One may suspect that this holds generally: the epistemic goals of minimizing expected average inaccuracy and minimizing expected total inaccuracy always give the same result regarding which credence minimizes expected inaccuracy. In fact, when self-locating beliefs are involved, this is not always the case.

3. Sleeping Beauty

On Sunday Sleeping Beauty is put to sleep, and she knows that on Monday researchers will wake her up, and then put her to sleep with a memory-erasing drug that causes her to forget that waking-up. She also knows that 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) and many others argue that the answer is 1/3, while David Lewis (2001) and some others argue that the answer is 1/2. We will argue that both these answers are epistemically permissible: 1/3 is obtained by following the goal of minimizing expected total inaccuracy, while 1/2 is obtained by following the goal of minimizing expected average inaccuracy.

In the Heads possible world, there is one temporal part of Beauty which has an opinion about Heads, where the length of the temporal part is taken to be the amount of time Beauty is kept awake. In the Tails possible world, however, there are two temporal parts of Beauty which have an opinion about Heads. When one calculates expected total inaccuracy, one sums the inaccuracy for each temporal part, while when one calculates expected average inaccuracy, one averages the inaccuracy for each temporal part.

Consider first the goal of minimizing expected total inaccuracy. Supposing that Beauty gives the 1/3 answer,

$$S_{ET}(H) = 1/2\ (1 - 1/3)^2 + 1/2\ [(0 - 1/3)^2 + (0 - 1/3)^2]$$

$$= 1/3.$$

The 1/2 factors are there because the coin is fair: half the time W(H) = 1, and half the time W(H) = 0. One can easily verify, here and in the cases below, that the expected inaccuracy for T is the same as that for H.

Suppose now that Beauty gives the 1/2 answer:

$$S_{ET}(H) = 1/2\ (1 - 1/2)^2 + 1/2\ [(0 - 1/2)^2 + (0 - 1/2)^2]$$

$= 3/8.$

Beauty's expected total inaccuracy for H is higher if she gives the 1/2 answer. In fact, one can easily verify that the 1/3 answer minimizes the expected total inaccuracy for H.

Now consider the goal of minimizing expected average inaccuracy. Supposing that Beauty gives the 1/3 answer,

$$S_{EA}(H) = 1/2\ (1 - 1/3)^2 + 1/2\ \{1/2\ [(0 - 1/3)^2 + (0 - 1/3)^2]\}$$

$$= 5/18.$$

Supposing that Beauty gives the 1/2 answer,

$$S_{EA}(H) = 1/2\ (1 - 1/2)^2 + 1/2\ \{1/2\ [(0 - 1/2)^2 + (0 - 1/2)^2]\}$$

$$= 1/4.$$

One can easily verify that the 1/2 answer minimizes the expected average inaccuracy for H. We see that the 1/3 answer is obtained by following the goal of minimizing expected total inaccuracy, while the 1/2 answer is obtained by following the goal of minimizing expected average inaccuracy.[9]

What reasons could one give in favor of one of these epistemic goals over the other? The proponent of minimizing expected total inaccuracy can be expected to reason as follows. Beauty should care about being inaccurate for each temporal part of her that has an opinion. Since in the Tails world there are two temporal parts of her that have an opinion about H, then the inaccuracy for each temporal part should matter to her. Beauty wants more than just her opinion at a particular time to be

[9]It is worth noting a parallel between the Sleeping Beauty problem and Newcomb's problem: just as the one-box and two-box answers to Newcomb's problem are the results of two competing types of decision theories, so are the 1/3 and 1/2 answers to the Sleeping Beauty problem.

minimally inaccurate; she wants her sum total of opinions to be minimally inaccurate.

The proponent of minimizing expected average inaccuracy, on the other hand, can be expected to reason as follows. Beauty, qua epistemic agent, should only care about being inaccurate regarding the opinion she currently has; she should not worry about other opinions she may have at other times. From an epistemic standpoint, when deciding what her current opinion should be, she should only consider the expected inaccuracy for her current opinion. Thus, her goal should be to minimize the expected inaccuracy for the opinion of her current temporal part. But since she does not know which temporal part she is, her best guess for the expected inaccuracy of the opinion of her current temporal part is the expected average inaccuracy for the various possible temporal parts of her. It follows that she should minimize her expected average inaccuracy.

We do not see any conclusive arguments in favor of one epistemic goal over the other, but perhaps this lack of a definitive answer is to be commended. After careful consideration of a relatively simple probability puzzle, smart people continue to disagree. It could be that one side or the other is simply wrong, but we prefer the conclusion that neither side is incorrect; it's just that they have different epistemic goals, each of which is epistemically permissible.

There is a variation of the Sleeping Beauty problem which is worth considering. Suppose that, some specified interval of time after Beauty wakes up, she is told what day it is. What probability should she assign to H on Monday after she is told that it is Monday? When Beauty knows that it is Monday, there is no difference between the method of minimizing expected average inaccuracy and minimizing expected total inaccuracy; both methods give the answer that her probability for H should be 1/2:

$$S_E(H) = 1/2\ (1 - 1/2)^2 + 1/2\ (0 - 1/2)^2$$

$$= 1/4.$$

One can easily verify that this answer minimizes expected inaccuracy. Elga also gives the 1/2 answer for this scenario, but Lewis gives the answer of 2/3. Without giving a conclusive argument against the 2/3 answer, we will simply report that we have the (widely shared) opinion that that answer is implausible. Given that one sticks with the 1/2 answer after having woken up on Monday, we believe that one should continue to assign probability 1/2 after being told that it is Monday. Lewis thinks otherwise because Lewis treats learning that it is Monday via Bayesian conditionalization, but we maintain that conditionalization is sometimes inappropriate for situations involving self-locating beliefs and information loss.

Some have said that the Sleeping Beauty problem is a case of experience duplication: Beauty has the same experiences on Monday as on Tuesday, and that's why she doesn't know what day it is. But in fact experience duplication is an inessential part of the problem. We can imagine another variant of the Sleeping Beauty problem, where Beauty knows that either on Monday she wears blue pajamas and on Tuesday she wears red pajamas, or on Monday she wears red pajamas and on Tuesday she wears blue pajamas, but she doesn't know which, and is indifferent between the two possibilities. When Beauty wakes and finds herself wearing blue pajamas, say, she knows that if the coin lands Tails her other waking temporal part wears red pajamas. Nevertheless, she doesn't have any more information about whether it's Monday or Tuesday than she does in the original Sleeping Beauty problem. In this variant, then, she and her other waking temporal part have different experiences, while all essential aspects of the problem are unchanged.

4. Duplicating Beauty

Consider now the case of *Duplicating Beauty*.[10] On Sunday God tells her that at midnight he will flip a fair coin; if the coin lands Heads God will do nothing, but if the coin lands Tails God will create a qualitatively identical duplicate of Beauty (Beauty*), on a qualitatively identical planet (Earth*). 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?

Currently, people's opinions on this problem are not as settled as they are for the Sleeping Beauty problem. Informal discussion suggests that many of the people who give the 1/3 answer for the case of Sleeping Beauty feel compelled to give that answer for the case of Duplicating Beauty, since the only difference is that in the Sleeping Beauty problem one is dealing with two qualitatively identical temporal parts of one person, existing at two different times, while in the Duplicating Beauty problem one is dealing with two qualitatively identical temporal parts of two people, existing at the same time.

The difficulty with the 1/3 answer (as some of its proponents recognize) is that the reasoning that leads to that answer has counterintuitive consequences in other scenarios. Consider first a variant of the Duplicating Beauty problem, where instead of God creating one duplicate when the coin lands Tails, God creates 999 duplicates. By the reasoning that leads to the 1/3 answer in Duplicating Beauty problem, Beauty should assign credence 1/1000 to Heads in the variant problem. Now consider a further variation, where the coin is somewhat biased in favor of heads, and Beauty knows this. Beauty will still end up assigning a low credence to Heads, since her initial bias in favor of Heads will get

[10]This is a combination of the Sleeping Beauty scenario with the duplication scenario described by Elga (2003).

swamped by the large probability shift in favor of Tails.

Now consider the hypothesis that the world is continuously splitting into a large number of duplicate worlds. After one second, for example, the world splits into one billion duplicate worlds, and after one more second, each of these worlds splits into one billion duplicate worlds, and so on. Our credence for the proposition *S* that a hypothesis of this sort is true is very low, but non-zero. (The splitting worlds version of many-worlds quantum mechanics is a hypothesis of this sort, and our credence for this version of quantum mechanics is also low but non-zero.) By the reasoning that leads to the 1/3 answer for the Duplicating Beauty problem, our credence for *S* should keep going up. Eventually, we would end up assigning a credence for *S* that is close to 1.

Frank Arntzenius has considered this sort of hypothesis in this context. Arntzenius (2003) is a proponent of the 1/3 answer to the Duplicating Beauty problem, and he (via personal communication) says that his prior probability for *S* is 0. Assuming that one is a Bayesian, this ensures that one's probability for *S* stays at 0, since conditionalization can never raise a probability assignment of 0. We believe that assigning probability 0 to *S* is unreasonable, though. We want to recognize the possibility of a scenario where God comes to earth and tells everyone that *S* is true; we would want to increase our probability for *S* on the basis of such testimonial evidence.

According to the epistemic goals discussed in the previous section, what probability should Duplicating Beauty assign to H on Monday? The mathematical reasoning from the previous section holds for this problem, by replacing talk about the current temporal part and another possible temporal part of Sleeping Beauty with talk about the actual Duplicating Beauty and a possible duplicate of her. If Duplicating Beauty's goal is to minimize expected total inaccuracy, for her and a possible duplicate of

her, then she should give the 1/3 answer, while if her goal is to minimize expected average inaccuracy, again for her and a possible duplicate of her, then she should give the 1/2 answer.

We believe that the correct answer is 1/2; Duplicating Beauty should follow the epistemic goal of minimizing expected average inaccuracy. Epistemically, Beauty should not care about the inaccuracies of other people. The goal of an epistemic agent should be to get *her own* beliefs in line with reality; she should not sacrifice accuracy in her own beliefs for the sake of reducing inaccuracy in other people's beliefs. If her goal is to minimize her own expected inaccuracy, then she should minimize the expected average inaccuracy for her and the possible duplicate of her, since she does not know which of those possible individuals she is.

In sum, the difference between the Sleeping Beauty case and the Duplicating Beauty case is an epistemically relevant difference, and the method of minimizing inaccuracy can be used to show this. It matters that, in the Sleeping Beauty case, the two temporal parts in the Tails world are temporal parts of the same person; it is open to Beauty to take this into account when reasoning about what credence she should assign to her temporal part. It matters that, in the Duplicating Beauty case, the two temporal parts in the Tails world are temporal parts of different people; from an epistemic standpoint Beauty should not let the inaccuracies of other people influence how she assigns her own credences.

It's worth noting that there may be non-epistemic goals that favor the 1/3 answer to the Duplicating Beauty problem. Suppose that on the Monday afternoon after the duplication, Beauty and (if she exists) Beauty* will be subjected to an amount of pain equal to their inaccuracy for H. Suppose that Beauty is a utilitarian, so her goal is to minimize the total amount of pain the world. On Sunday, when Beauty is trying to decide what credence to assign to H on Monday, she has an ethical reason to

decide to assign credence 1/3 to H on Monday, since that credence will minimize the expected total inaccuracy on Monday, and hence will minimize the expected total amount of pain received on Monday. We find this interesting but not problematic; it is to be expected that one's epistemic goals and one's ethical goals can sometimes conflict.

5. Conclusion

Minimizing expected inaccuracy is a general epistemic goal for agents. In situations that do not involve self-locating beliefs or information loss, following that epistemic goal is compatible with Bayesianism. In situations that involve self-locating beliefs and information loss, Bayesianism does not always apply, but the method of minimizing expected inaccuracy still does. The method is not univocal though: one could minimize expected average inaccuracy or expected total inaccuracy. Which one should do depends on the specifics of the problem in question, and sometimes there is no right answer.[11]

[11]Thanks to Frank Arntzenius, Adam Elga, Branden Fitelson, and Sam Ruhmkorff for helpful discussion.

References

Arntzenius, Frank (2003), "Self-locating beliefs, reflection, conditionalization, and Dutch books", *Journal of Philosophy*, forthcoming.

Brier, Glenn (1950), "Verification of forecasts expressed in terms of probability", *Monthly Weather Review* 78: 1-3.

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

Elga, Adam (2003), "Defeating Dr. Evil with self-locating belief", *Philosophy and Phenomenological Research*, forthcoming.

Hájek, Alan (1998), "Agnosticism meets Bayesianism", *Analysis* 58: 199-206.

Lewis, David (2001), "Sleeping Beauty: reply to Elga", *Analysis* 61: 171-76.

Jeffrey, Richard (1986), "Probabilism and Induction", *Topoi* 5: 51-58.

Joyce, James (1998), "A nonpragmatic vindication of probabilism", *Philosophy of Science* 65: 575-603.

Maher, Patrick (2002), "Joyce's argument for probabilism", *Philosophy of Science* 69: 73-81.

Monton, Bradley (1998), "Bayesian agnosticism and constructive empiricism", *Analysis* 58: 207-12.

Monton, Bradley (2002), "Sleeping Beauty and the forgetful Bayesian", *Analysis* 62: 47-53.

Savage, Leonard (1971), "Elicitation of personal probabilities and expectations", *Journal of the American Statistical Association* 66: 783-801.

Selten, Richard (1998), "Axiomatic characterization of the quadratic scoring rule", *Experimental Economics* 1: 43-62.

van Fraassen, Bas (1998), "The agnostic subtly probabilified", *Analysis* 58: 212-220.