THE ROLE OF THE PROTOCOL IN ANTHROPIC REASONING

JOSEPH Y. HALPERN
Cornell University, Computer Science Department

I show how thinking in terms of the protocol used can help clarify problems related to anthropic reasoning and self-location, such as the Doomsday Argument and the Sleeping Beauty Problem.

There is a complex of problems related to anthropic reasoning and self-location, the Doomsday Argument (Leslie 1996) and the Sleeping Beauty Problem (Elga 2000) perhaps chief among them, that have generated a great deal of heat in the philosophy literature. Solutions have been proposed and then disposed of.¹ Here I discuss another solution, which has applicability far beyond the scope of these problems. The point is to make precise exactly what protocol is being followed in all these puzzles. I take a protocol for an agent to be a description of what that agent does at each step, as a function of the agent’s information.² For convenience, as in (Fagin et al. 1995), I assume that nature also follows a protocol. A protocol is essentially what game theorists call a strategy. Bradley (2012) uses the word mechanism in a similar sense.

The importance of thinking about protocols when doing conditioning was already made by Shafer (1985), and is discussed at some length in (Halpern 2003: Chapter 6). Bovens (2012) applies protocols to the analysis of the reports of miracles; Bradley (2012) and Bovens and Ferreira (2010) also apply them in the context of self-location puzzles.³ Before considering protocols in problems of self-location,
I discuss the second-ace puzzle (Freund 1965), since it illustrates the role of the protocol particularly well. The following discussion is largely taken from (Halpern 2003).

Example 1: [The second-ace puzzle] A deck has four cards: the ace and deuce of hearts, and the ace and deuce of spades. After a fair shuffle of the deck, two cards are dealt to Alice. It is easy to see that, at this point, there is a probability of $1/6$ that Alice has both aces, a probability of $5/6$ that Alice has at least one ace, a probability of $1/2$ that Alice has the ace of spades, and a probability of $1/2$ that Alice has the ace of hearts: of the six possible deals of two cards out of four, Alice has both aces in one of them, at least one ace in five of them, the ace of hearts in three of them, and the ace of spades in three of them.

Alice then says, “I have an ace.” Conditioning on this information (by discarding the possibility that Alice was dealt no aces), Bob computes the probability that Alice holds both aces to be $1/5$. This seems reasonable. The probability, according to Bob, of Alice having two aces goes up if he learns that she has an ace. Next, Alice says, “I have the ace of spades.” Conditioning on this new information, Bob now computes the probability that Alice holds both aces to be $1/3$. Of the three deals in which Alice holds the ace of spades, she holds both aces in one of them. As a result of learning not only that Alice holds at least one ace, but that the ace is actually the ace of spades, the conditional probability that Alice holds both aces goes up from $1/5$ to $1/3$. But suppose that Alice had instead said, “I have the ace of hearts.” It seems that a similar argument again shows that the conditional probability that Alice holds both aces is $1/3$.

Is this reasonable? When Bob learns that Alice has an ace, he knows that she must have either the ace of hearts or the ace of spades. And no matter what she says at the second step, his probability that she has both aces goes up to $1/3$. But if the probability goes up from $1/5$ to $1/3$ whichever ace Alice says she has, and Bob knows that she has an ace, then why isn’t it $1/3$ all along?

To analyze this puzzle correctly, we have to specify Alice’s protocol. In this case, a protocol for Alice determines what Alice says at each step. For simplicity, I assume that Alice tells the truth. But even with that restriction, there are many possible protocols that Alice could be following. One protocol proceeds as follows: In the first round, Alice tells Bob whether or not she has an ace. Then, in round 2, Alice tells Bob she has the ace of spades if she has it and otherwise says she hasn’t got it. This protocol is deterministic. With this protocol, there are six possible runs (sequences of events) that could happen, corresponding to the 6 possible pairs of cards that Alice can be dealt. Since the deal is supposed to be fair, each of these

What Bradley calls a selection procedure an be understood as a protocol. See the discussion after the Sleeping Beauty example for further comments on the Bovens and Ferreira paper.
runs has probability $1/6$. With this protocol, the analysis above is perfectly correct. Indeed, after Alice says she has an ace, Bob’s conditional probability that Alice has two aces is indeed $1/5$; and after Alice says she has the ace of spades, Bob’s conditional probability that has both aces as $1/3$. However, with this protocol, the concern as to what happens if Alice tells Bob that she has the ace of hearts does not arise. This cannot happen, according to the protocol. All that Alice can say is whether or not she has the ace of spades.

Now consider a different protocol (although one still consistent with the story). Again, in round 1, Alice tells Bob whether or not she has an ace. However, now, in round 2, Alice tells Bob which ace she has if she has an ace (and says nothing if she has no ace). This still does not completely specify the protocol. What does Alice tell Bob in round 2 if she has both aces? One possible response is for her to say “I have the ace of hearts” and “I have the ace of spades” with equal probability. This protocol is almost deterministic. The only probabilistic choice occurs if Alice has both aces. With this protocol there are seven runs. Each of the six possible pairs of cards that Alice could have been dealt determines a unique run with the exception of the case where Alice is dealt two aces, for which there are two possible runs (depending on which ace Alice tells Bob she has). Each run has probability $1/6$ except for the two runs where Alice was dealt two aces, which each have probability $1/12$.

It is still the case that after Alice says that she has an ace, Bob’s conditional probability that Alice has two aces is $1/5$. What is the situation in round 2, after Alice says she has the ace of spades? In this case Bob considers three runs possible, the two where Alice has the ace of spades and a deuce, and the one where Alice has both aces and tells Bob she has the ace of spades. Notice, however, that after conditioning, the probability of the point on the run where Alice has both aces is $1/5$, while the probability of each of the other two points is $2/5$! This is because the probability of the run where Alice holds both aces and tells Bob she has the ace of spades is $1/12$, half the probability of the runs where Alice holds only one ace. Thus, Bob’s probability that Alice holds both aces in round 2 is $1/5$, not $1/3$, if this is the protocol. The fact that Alice says she has the ace of spades does not change Bob’s assessment of the probability that she has two aces. Similarly, if Alice says that she has the ace of hearts in round 2, the probability that she has two aces remains at $1/5$.

Now suppose that instead of randomizing in round 2 if she has both aces, Alice says “ace of spades.” In that case, if Alice does say “I have the ace of spades” in round 2, the probability according to Bob that she has both aces is back to $1/3$, but if Alice says “I have the ace of hearts,” the probability according to Bob that she has both aces is 0 (since she would never say “I have the ace of hearts” if she has both aces).

One last protocol: Again in round 1, Alice tells Bob whether she has an ace. Then in round 2, she chooses one of the two cards in her hand (uniformly at ran-
and tells Bob which it is. Now there are 12 possible runs, two for each of the possible pairs of cards that Alice could have. With this protocol, after Alice says that she has the ace, it is again the case that Bob’s conditional probability that she has both aces is 1/5. And after Alice says “I have the ace of spades”, the probability goes up to 1/3; it also goes up to 1/3 after she says “I have the ace of hearts”. But now there is no paradox. The probability is not 1/3 no matter what Alice says. For example, Alice could say “I have the two of spades” (an option that was implicitly excluded at the beginning), in which case Bob’s probability that she has both aces is 0.

This example illustrates how the choice of protocol determines the conditional probabilities, and how thinking in terms of protocols illuminates what is going on. Since the story does not describe Alice’s protocol, one important conclusion that we can draw here is that there is no “right” answer as to what the probability that Alice has both aces is. The answer depends on the protocol being used. I now apply this approach to problems of anthropic reasoning and self-location.

**Example 2:** [The doomsday argument] Suppose that we are uncertain as to when the universe will end. For simplicity, we consider only two hypotheses: the universe will end in the near future, after only $n$ humans have lived, or it will end in the distant future, after $N$ humans have lived, where $N \gg n$. You are one of the first $n$ people. What can you conclude about the relative likelihood of the two hypotheses? That depends in part on your prior beliefs about these two hypotheses. But it also depends, as I show now, on the protocol that we take Nature to be using. In particular, for you to decide the relative likelihood of these two hypotheses, you need to decide how Nature determined the ending time of the universe and how Nature chose you.4

Just about all papers on the subject implicitly assume that Nature is using the following protocol. Nature first chooses the ending time of the universe, and then chooses who you are uniformly at random among the people who live in the universe. The latter point is typically modeled by saying that Nature chooses an index $i$ for you, where $i$ is viewed as your birth order (you are the $i$th person to be born) and either $1 \leq i \leq n$ if the universe ends soon, or $1 \leq i \leq N$ if the universe survives for a long time. You are assumed to know your index, so you can condition on that information. We are interested in

$$\Pr(\text{the universe ends soon} \mid \text{you are index } i).$$

---

4. You might say that there is no Nature. But somehow the ending date of the universe must be chosen and who you are must be chosen. I am taking “Nature” to be a representation of—or a “metaphor for”, in the words of a reviewer of this paper—these processes (which I am assuming are probabilistic processes). Similar comments apply to all other appearances of Nature in this paper.
Suppose that your prior that the universe ends soon is \( \alpha \) (i.e., you believe that Nature chose the universe to end soon with probability \( \alpha \)). Then, by Bayes’ rule, and assuming that you are chosen uniformly at random among the individuals in the universe (this is the anthropic principle), we get that

\[
\Pr(\text{the universe ends soon} \mid \text{you are index } i) = \frac{\Pr(\text{you are index } i \mid \text{the universe ends soon}) \times \Pr(\text{the universe ends soon})}{\Pr(\text{you are index } i)} = \frac{\alpha}{n \Pr(\text{you are index } i)}
\]

Moreover,

\[
\Pr(\text{you are index } i) = \Pr(\text{you are index } i \mid \text{the universe ends soon}) \times \Pr(\text{the universe ends soon}) + \Pr(\text{you are index } i \mid \text{the universe survives a long time}) \times \Pr(\text{the universe survives a long time}) = \frac{\alpha}{n} + \frac{1 - \alpha}{N}.
\]

Thus,

\[
\Pr(\text{the universe ends soon} \mid \text{you are index } i) = \frac{\alpha}{n} / \left[ \frac{(\alpha}{n} + \frac{1 - \alpha}{N} \right] = \frac{\alpha}{n + (1 - \alpha) / N}.
\]

Since \( N \gg n \), this will be close to \( \alpha \).

But now consider a different protocol for Nature. First, Nature chooses who you are (i.e., chooses an index \( i \) between 1 and \( N \)), uniformly at random, and then chooses when the universe ends. Now if \( i > n \), Nature’s choice is determined (this is analogous to Alice’s choice being determined in the second protocol if she gets something other than a pair of aces); the universe must survive a long time. But if \( i \) is between 1 and \( n \), then Nature has a choice. By analogy with the first protocol, suppose that if \( i \) is between 1 and \( n \), then Nature decides that the universe will end soon with probability \( \alpha \). With this protocol, it is almost immediate to see that if \( i < n \), then

\[
\Pr(\text{the universe ends soon} \mid \text{you are index } i) = \alpha
\]

Thus, with this protocol, your posterior probability that the universe will end soon \( (\Pr(\text{the universe ends soon} \mid \text{you are index } i)) \) is the same as the probability that nature chose an early ending date given that nature had a choice to make \( (\Pr(\text{the universe ends soon} \mid i < n)) \). Conditioning on the actual index did not affect this probability.
I do not mean to suggest that these are the only two protocols for Nature, although these seem to me the most obvious ones, and the ones closest to the spirit of the story. There are certainly more intricate protocols where, for example, the ending time of the universe depends on the index in a more detailed way (e.g., if \( i < n \) and \( i \) is odd, then the universe will end in the distant future; if \( i < n \) and \( i \) is even, then the universe will end in the near future). Given that the conditional probability depends so much on Nature’s protocol, an obvious question is how we can determine Nature’s protocol. In general, we cannot. To me, the second protocol seems more appropriate here—it seems more reasonable to me that your index is chosen before you consider the age of the universe. My argument for reasonableness is admittedly rather weak: to me the primary choice is who you are. Once you exist, how long the universe will survive is only one of many questions that you could have asked: whether the stock market will go up tomorrow, whether the person you like is interested in going out with you, whether it will rain tomorrow, and so on. From your perspective, these questions are meaningless until you exist. So it seems to me that you have to be chosen first. But there is clearly room for debate here. For example, we can debate whether the first protocol is really a feasible protocol for nature. Specifically, since each choice that Nature makes regarding the ending time arguably defines a different universe, we can debate whether Nature can choose an arbitrary person to be the same “you” in these different universes (see, for example, Stalnaker 2012). What I would argue should not be open to debate is the need to make clear what the protocol is. In this case, that means making clear whether the index is chosen before or after the ending time of the universe is chosen.

Note that it would not help to imagine God somehow performing this experiment repeatedly so that we could check apply a frequentist interpretation and check what the “true” conditional probability. God cannot perform the experiment without specifying the protocol! The “true” conditional probability depends on the protocol.

I next consider two examples discussed by Bostrom (2012). The analysis is very similar.

Example 3: [The incubator] In an otherwise empty world, a machine called “the incubator” works as follows. It tosses a fair coin. If the coin lands heads, then it creates one room and a man with a black beard. If the coin lands tails, then it creates two rooms, one with a black-bearded man and one with a white-bearded man.\(^5\) Initially the world is dark. What should be our credence that the coin landed tails? When the lights are switched on, you discover that you have a black beard. Now what should be your credence that the coin landed tails?

---

\(^{5}\) Bostrom (2012) takes there to be one room if the coin lands tails and two if the coin lands heads; I have switched it here to make the presentation consistent with the Sleeping Beauty problem (Example 5).
Again, I consider two protocols for Nature. In the first, the incubator tosses a coin at random to determine $M$, the number of rooms, and then if $M = 2$, Nature chooses your beard color at random and places you in one of the rooms. With this protocol, the prior probability of tails is $1/2$. The probability of tails given that you are in a dark room is still $1/2$. Finally, $\Pr(\text{tails} \mid \text{black beard}) = 1/3$.

In the second protocol, Nature first chooses your beard color; you are equally likely to have a black beard and a white beard. Then the incubator tosses a fair coin to determine $M$. If $M = 1$ and you have a black beard, you are placed in the one room that is created; if you have been chosen to have a white beard, then you are not placed in a room; the other person is placed in the room and has a black beard. If $M = 2$, then you and the other person are placed in different rooms, and the other person has a beard whose color is different from yours. With this protocol, the prior probability of tails is again $1/2$, but the probability of tails given that you are in a dark room is $2/3$ (since you would not have placed in a room if the coin had landed heads and you had a white beard), while $\Pr(\text{tails} \mid \text{black beard}) = 1/2$.

Bostrom (2012) analyzed this example using two assumptions that he called the Self-Sampling Assumption (SSA) and the Self-Indication Assumption (SIA). The former says “One should reason as if one were a random sample from the set of all observers in one’s reference class”, while the latter says “Given that you exist, you should (other things being equal) favor hypotheses according to which many observers exist over hypotheses under which few observers exist.” The first protocol gives the same answers as Bostrom’s SSA analysis, while the second gives the answer that Bostrom gets using what he takes to be a combination of SSA and SIA. I would argue that it is clearer to replace assumptions like SIA and SSA by assumptions on Nature’s protocol; we do not need to appeal to SSA and SIA (which seem to me somewhat fuzzy statements at best) to analyze the problem. (Alternatively, we could try to restate SIA and SSA as protocols. If it could be done, this would have the benefit of making them more precise.)

**Example 4:** [Observer-relative chances] (Bostrom 2012: 131). Suppose the following takes place in an otherwise empty world. A fair coin is flipped by an automaton and if it falls heads, one human is created; if it falls tails, ten humans are created. In addition to these people, one other human is created independently of how the coin falls. The latter human we call the bookie. The people created as a result of the coin toss we call the group. Everybody knows these facts. Furthermore, the bookie knows that she is the bookie, and the people in the group know that they are in

---

6. I omit the details of Bostrom’s arguments here, nor do I try to formalize SSA and SIA. The details are not necessary for the discussion that follows.
the group. The question is, what would be the fair odds if the people in the group want to bet against the bookie on how the coin fell?

Again, I would argue that to answer this question, we need to incorporate in the protocol a process for choosing who “you” are. Here is one protocol: A fair coin is tossed, and either 2 or 11 people are created depending on whether it lands heads or tails. The bookie and “you” are chosen from among those created, independently, with uniform probability. In that case: \( \text{Pr} (\text{heads} | \text{you are the bookie}) \) is high: \( \frac{1}{4} / (\frac{1}{4} + \frac{1}{22}) = 11/13 \). This seems to be the protocol that Bostrom (2012) is implicitly using.

But now consider the following protocol: 11 “virtual” people are created, and the bookie and you are chosen among them, again, independently, with uniform probability. Then a fair coin is chosen to decide whether only two virtual people are to be actualized, or all 11. If the coin lands tails, all 11 virtual people are actualized, and if it lands heads, the bookie is actualized, as well as a second person chosen uniformly at random from the 10 non-bookies. Now the probability that you are the bookie is clearly \( 1/11 \). The probability that the coin lands heads and you are the bookie is \( 1/22 \), so \( \text{Pr} (\text{heads} | \text{you are the bookie}) = 1/2 \).

Again, the key question involves deciding when “you” are chosen, relative to the other decisions that have to be made. The protocol makes this clear. Making this decision before the other decisions are made (in this case, before the size of the world is chosen) seems just as consistent with Bostrom’s story as making it after the other decisions are made.

I conclude by considering the Sleeping Beauty problem.

Example 5: [Sleeping Beauty] Here is the description of the problem, taken from Elga (2000).

Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (heads: once; tails: twice). After each waking, they will put you back to sleep with a drug that makes you forget that waking. When you are first awakened, to what degree ought you believe that the outcome of the coin toss is heads?

The discussion of Sleeping Beauty here is relatively brief, so the main points are not so different from those made in the other examples. To simplify the discussion, let me suppose that the first wakening happens on Monday and the second (if there is one) happens on Tuesday. The two standard answers to this question are \( 1/2 \) (it was \( 1/2 \) before you were put to sleep and you knew all along that you would be woken up, so it should still be \( 1/2 \) when you are actually woken up) and \( 1/3 \) (on
the grounds that you should consider each of the following three events equally likely when you are woken up: it is now Monday and the coin landed heads; it is now Monday and the coin landed tails; it is now Tuesday and the coin landed tails). Not surprisingly, I now argue that the answer depends on the protocol.

It may seem at first as there is no room for Nature here. After all, the only probabilistic step is the coin toss, and the probabilities of the outcomes are clearly specified; it is a fair coin. However, from your point of view, there is uncertainty regarding when “now” is. Is it Monday or Tuesday? How is this uncertainty to be resolved? To understand this issue, consider a simpler question. Suppose that there is no coin toss. Instead, the researchers simply wake you up on Monday, give you the amnesia drug, and then wake you up again on Tuesday. Each time you are woken up, you are asked how likely it is to be Monday. I cannot think of any meaningful response other than that obtained by applying the principle of indifference and taking this probability to be \(1/2\). Similarly, I assume that you consider it equally likely to be Monday and Tuesday conditional on the coin landing tails.7

Nevertheless, there is still room for Nature here. There are two reasonable protocols for generating a probability on the set \{\text{(Monday, heads)}, \text{(Tuesday, heads)}, \text{(Tuesday, tails)}\} of worlds that you need to use to determine the probability of heads, depending on whether the coin is tossed before or after “now” is determined. According to the first protocol, the coin is tossed first, and then “now” is chosen. Of course, there is no choice to make if the coin lands heads: you are just woken up once, so the probability that it is Monday given that the coin landed heads is 1. As I suggested above, the probability that it is Monday given that the coin landed tails is \(1/2\). With this protocol, the probability of both \text{(Monday, tails)} (“the coin lands tails and it is Monday”) and \text{(Tuesday, tails)} is \(1/4\). Thus, the probability that you should ascribe to heads is \(1/2\).

According to the second protocol, “now” is chosen first—it is either Monday or Tuesday (with equal likelihood)—and then the coin is tossed. If Tuesday is chosen and the coin lands heads, then the experiment has already ended, and you are not asked anything. (This is somewhat similar to the second protocol in the analysis of the incubator puzzle: if you have a white beard and only one room is created, you are not placed in a room.) With this protocol, \(\Pr(\text{Monday, heads}) = \Pr(\text{Monday, tails})\). We must also have \(\Pr(\text{Monday, tails}) = \Pr(\text{Tuesday, tails})\). Thus, all three possibilities are equally likely, so must all have probability \(1/3\). The probability that you should ascribe to heads when you are woken up is \(1/3\). To me, the

---

7. I remark that, in analogous settings, Halpern and Tuttle (1993) declare this question to be meaningless. They take the event “it is Monday and the coin landed tails” to be nonmeasurable, and do not ascribe it a probability. Thus, the conditional probability of Monday given that the coin landed tails is also undefined. While I continue to be sympathetic to the idea of using nonmeasurability here, to better compare the approach here to others in the literature, I take Monday and Tuesday to be equally likely, conditional on the coin landing tails.
first protocol seems more reasonable—it seems more consistent with the presentation of the story to think of the coin as being tossed first. But again, reasonable people can disagree.8

The point of these examples should be clear: when dealing with subtle problems involving self-location and probability, it is important to clarify how and when all probabilistic decisions are made; this involves clarifying what protocols are being used by the participants and by Nature. All the protocols that I have discussed here have the same structure. Two decisions have to be made; one involves picking a person, while the other involves deciding on some other feature of the problem (the day that Sleeping Beauty will be woken up; the color of the beard; the duration of the universe). The question is in which order these decisions should be made. While we might have intuitions about what the order ought to be, there seems to be no compelling argument that one of them is “right”.

Although the protocols discussed here all have this structure, I do not believe that all problems of philosophical interest will involve protocols like this (or even that all self-location problems will necessarily involve protocols like this). To take just one example, Grove and Halpern (1997) have done an analysis of van Fraassen’s (1981) Judy Benjamin problem by providing an explicit updating protocol for Judy; Bovens and Ferreira (2010) suggest a different protocol (which leads to the same conclusion, but for quite different reasons). Both protocols are quite different from those considered here.

An obvious question is what constitutes a valid protocol for a particular problem. Unfortunately, I don’t have a definitive answer to this question. It is largely subjective: a modeler will need to decide if a protocol is true to the spirit of the description of a situation. Some protocols seem clearly inappropriate (e.g., a protocol for the Doomsday problem where the length of time that the universe survives depends on whether the index i chosen is odd or even). In some cases, the protocol generating the probabilities in the problem is quite clear; in others, there can be reasonable disagreement about it.

8. In (Halpern 2005), two different approaches are considered for ascribing probability in what computer scientists call asynchronous systems—ones where there is uncertainty regarding when “now” is. These ideas are then applied to the Sleeping Beauty problem. The two approaches can be viewed as arising from the two different protocols sketched above. Bovens and Ferreira (2010) also discuss the Sleeping Beauty problem in terms of protocols, and stress the importance of thinking in terms of protocols. Not surprisingly, they also provide two protocols for Sleeping Beauty; one leads to the answer 1/3, and the other to 1/2. But their protocol that gives 1/2 involves telling you when you are woken up what day it is not. When you are told that it is not Tuesday, then you give probability 1/2 to heads. This seems different in spirit from Elga’s presentation of the problem, as they themselves point out. In Elga’s presentation, you are not conditioning on extra information.
In any case, the discipline of doing a formal analysis in terms of protocols forces a modeler to be careful about exactly what the sample space is (see Grove and Halpern 1997 and Halpern 2003 for further discussion of this point), which helps further clarify issues. As I hope I have made clear, thinking in terms of protocols should prove useful for many problems of philosophical interest.

Acknowledgements

Work supported in part by NSF grants IIS-0812045, IIS-0911036, and CCF-1214844, by AFOSR grants FA9550-08-1-0438, FA9550-09-1-0266, and FA9550-12-1-0040, and by ARO grant W911NF-09-1-0281. Thanks to Robert Rand and the anonymous reviewers of the paper for useful comments.

References
