# The Role of the Protocol in Anthropic Reasoning

December 31, 2013

**Abstract**

It is shown how thinking in terms of the protocol used can help clarify problems related to anthropic reasoning, such as the *Doomsday Argument* and the *Sleeping Beauty Problem*.

There is a complex of problems related to anthropic reasoning, 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 (see http://http://www.anthropic-principle.com/?q=resources/preprints for an annotated list of references). Here I discuss another approach to resolving these problems, 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. This point was already made by Shafer [1985], and is discussed at some length in [Halpern 2003][Chapter 6]. Before discussing the Doomsday Argument, I consider 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 0.1:** [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*: a complete description of Alice's rules for generating statements. One protocol that is consistent with the story is that, initially, Alice is dealt two cards. 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 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 still one 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 the run 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 random) 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. I now apply this thinking to problems of anthropic reasoning.

**Example 0.2:** [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 assume that Nature is using.

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

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

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

$$\begin{aligned} & \Pr(\text{the universe ends soon} \mid \text{you are index } i) \\ = \; & \Pr(\text{you are index } i \mid \text{the universe ends soon}) \times \Pr(\text{the universe ends soon}) / \Pr(\text{you are index } i) \\ = \; & \alpha / [n \Pr(\text{you are index } i)] \end{aligned}$$

Moreover,

$$\begin{aligned} & \Pr(\text{you are index } i) \\ = \; & \Pr(\text{you are index } i \mid \text{the universe ends soon}) \Pr(\text{the universe ends soon}) + \\ & \Pr(\text{you are index } i \mid \text{the universe survives a long time}) \Pr(\text{the universe survives a long time}) \\ = \; & \alpha/n + (1-\alpha)/N. \end{aligned}$$

Thus,

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

Since $N \gg n$, this will be close to 1.

But now consider a different model. 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). But if $i$ is between 1 and $n$, then Nature has a choice. By analogy with the first protocol, suppose that Nature chooses the universe will end soon with probability $\alpha$. In this model, it is easy to see that if $i < n$, then

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

Thus, with this protocol, $i$'s prior probability that the universe will end soon is the same as his posterior. Conditioning on his index has no effect. ∎

We can debate which of Nature's protocols is more appropriate here. To me, the second protocol seems more appropriate here (since your index is chosen before you consider the age of the universe), but there is clearly room for debate. 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.

I next consider two examples discussed by Bostrom [2012]. The analysis is very similar.

**Example 0.3:** [The incubator] In an otherwise empty world, a machine called "the incubator" works as follows. It tosses a fair coin. If the coin lands tails, 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. Initially the world is dark. When the lights are switched on, you discover that you have a black beard. What should be your credence that the coin landed tails?

Bostrom [2012] introduces what he calls the *Self-Indication Assumption (SIA)*, which 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." In particular, based on this assumption, he argues that it should be twice as likely that there are two observers than one.

I would argue that this assumption implicitly assumes a particular protocol for the incubator. To understand this issue, suppose that there are $M$ people in the world, with $M = 1$ if

the coin lands tails, and $M = 2$ if the coin lands heads. We can take a possible world to be a pair $(i, M)$, where $M$ is either 1 or 2, and $i$ is a natural number between 1 and $M$. Note that the world determines both $M$ (the number of observers) and $i$ (who you are). I assume that if $i = 1$ then you have a black beard, and if $i = 2$, then you have a white beard. If we assume that Nature's protocol is to choose a world at random, then indeed it is the case that SIA holds. But this seems rather strange, since it means that the prior probability of heads (which corresponds to the world $(1, n)$) is $1/3$, contradicting the assumption that the incubator starts by tossing a fair coin. Nevertheless, it is easy to see that, in this model, $\Pr(\text{tails} \mid \text{black beard}) = 1/2$.

But now consider a model where the incubator tosses a coin at random to determine $M$, and then if $M = 2$, tosses another coin at random to determine $i$. In this model, the prior probability of heads is $1/2$, but $\Pr(\text{tails} \mid \text{black beard}) = 2/3$. This second model seems more consistent with the assumption that the incubator starts by tossing a fair coin. ∎

**Example 0.4:** [Observer-relative chances] [Bostrom 2012][p. 131]. Suppose the following takes place in an otherwise empty world. A fair coin is flipped by an automaton and if it falls heads, ten humans are created; if it falls tails, one human is 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 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 probabilistic model a process for choosing who "you" are. Here is one model: A fair coin is tossed, and either 2 or 11 people are created depending on whether it lands tails or heads; the bookie is chosen randomly from among those created. In that case: $\Pr(\text{heads} \mid \text{you are the bookie})$ is high: $(1/4)/((1/4) + (1/22))$. This seems to be the model that Bostrom [2012] is implicitly using.

But now consider the following model: 11 "virtual" people are created, and the bookie is chosen among them uniformly at random. Then a fair coin is chosen to decide whether all 11 virtual people are to be actualized, or only two. If the coin lands heads, all 11 virtual people are actualized; if it lands tails, the bookie is actualized, as well as a second person chosen uniformly at random from the 10 non-bookies. Now $\Pr(\text{heads} \mid \text{you are the bookie}) = 1/2$.

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 0.5:** [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 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 strategy that you believe Nature is using.

There are two plausible choices. The first assumes that the coin is tossed first, then the day that Beauty is woken up is chosen. Of course, there is no choice to make if the coin lands heads; Beauty is just woken up in that case. If the coin lands heads, another coin is tossed to determine "now." Independent of the bias of the second coin, the probability of heads given "now" is $1/2$. The second assumes that you first choose whether "now" should be Monday or Tuesday (with equal likelihood), and then toss the coin. If Tuesday is chosen and the coin lands heads, this outcome is ignored, and the experiment is repeated. With this model, the probability of heads given "now" is $1/3$. ∎

The point of these examples should be clear: when dealing with problems with subtle problems involving probability, it is important to clarify how and when *all* probabilistic decisions are made, and this involves clarifying what protocols are being used by the participants and by Nature. In problems involving anthropic reasoning, the key decision is often in deciding how "you" are chosen. The process for choosing "you" has not been explicitly discussed in other papers. As I have tried to demonstrate, how it is done makes a big difference.

# References


Bostrom, N. (2012). *Anthropic Bias*. New York and London: Routledge.

Elga, A. (2000). Self-locating belief and the Sleeping Beauty problem. *Analysis 60*(2), 143–147.

Freund, J. E. (1965). Puzzle or paradox? *American Statistician 19*(4), 29–44.

Halpern, J. Y. (2003). *Reasoning About Uncertainty*. Cambridge, Mass.: MIT Press.

Halpern, J. Y. (2005). Sleeping beauty reconsidered: Conditioning and reflection in asynchronous systems. In T. S. Gendler and J. Hawthorne (Eds.), *Oxford Oxford Studies in Epistemology*, Volume 1, pp. 111–142.

Leslie, J. (1996). *The End of the World*. New York: Routledge.

Shafer, G. (1985). Conditional probability. *International Statistical Review 53*(3), 261–277.