Future of Humanity Institute
OXFORD UNIVERSITY

FHI TECHNICAL REPORT

Anthropics: why Probability isn’t enough

Stuart Armstrong

Technical Report #2012-2

Cite as:

Armstrong, S. (2012): “Anthropics: why Probability isn’t enough”, *Technical Report* #2012-2, Future of Humanity Institute, Oxford University: pp. 1-10.

The views expressed herein are those of the author(s) and do not necessarily reflect the views of the Future of Humanity Institute.

Anthropics: why probability isn't enough

This paper argues that the current treatment of anthropic and self-locating problems over-emphasises the importance of anthropic probabilities, and ignores other relevant and important factors, such as whether the various copies of the agents in question consider that they are acting in a linked fashion and whether they are mutually altruistic towards each other. These issues, generally irrelevant for non-anthropic problems, come to the forefront in anthropic situations and are at least as important as the anthropic probabilities: indeed they can erase the difference between different theories of anthropic probability, or increase their divergence. These help to reinterpret the decisions, rather than probabilities, as the fundamental objects of interest in anthropic problems.

Acknowledgments

The authors wish to acknowledge the help and support of the Future of Humanity Institute, the James Martin School and the Singularity Institute, as well as the individual advice of Nick Bostrom, Luke Muelhauser, Vincent Mueller, Anders Sandberg, Lisa Makros, Sean O'Heigeartaigh, Daniel Dewey, Eric Drexler and the online community of Less Wrong.

1 Introduction

We cannot have been born on a planet unable to support life. This self-evident truism is an example of anthropic or self-locating reasoning: we cannot see ourselves as 'outside observers' when looking at facts that are connected with our own existence. Less parochial extensions of this look at what anthropic reasoning can say about the fine tuning of constants that support our kind of life in this universe (Carter B. , 1974).

But there are open questions as to what the correct way of resolving anthropic questions is, and how to compute anthropic probabilities. Most of the work in this domain has revolved around various simplified formal problems (see for instance (Bostrom, Anthropic Bias: Observation Selection Effects in Science and Philosophy, 2002), (Aumann, Hart, & Perry, 1996), (Neal, 2006) (Elga, 2000), (Bostrom, The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe, 2001) and (Carter & McCrea, 1983)). The most famous simplified problem being the Sleeping Beauty Problem (Elga, 2000), from which Professor Bostrom (Bostrom, Anthropic Bias: Observation Selection Effects in Science and Philosophy, 2002) has formalised the two main schools of anthropic probability: the Self-Sampling Assumption (SSA) and the Self-Indication Assumption (SIA).

The contention of this paper, however, is that the focus on probability in anthropic problems is misleading, and that other factors are equally important in coming to a decision – indeed, different anthropic probability theories with different estimates of individual impact, for instance, can come to exactly the same decisions. Thus SIA and SSA cannot be seen as representing fundamentally different philosophical positions, only doing so in conjunction with other decision related factors.

But what are these factors? Anthropic problems are not concerned solely about the existence of potentially identical agents, but also of potentially identical agents whose decisions may or may not be linked, and who may or may not care about each other. It is these issues that this paper will explore, through the prism of the Sleeping Beauty problem and a related non-anthropic problem.

2 The Sleeping Beauty Problem

The Sleeping Beauty (Elga, 2000) problem is one of the most fundamental in anthropic probability. It is related to many similar problems, such as the absent-minded driver (Aumann, Hart, & Perry, 1996), the Sailor's Child Problem (Neal, 2006), the incubator and the presumptuous philosopher (Bostrom, Anthropic Bias: Observation Selection Effects in Science and Philosophy, 2002).

In the standard setup, Sleeping Beauty is put to sleep on Sunday, and awoken again Monday morning, without being told what day it is. She is put to sleep again at the end of the day. A fair coin was tossed before the experiment began. If that coin showed heads, she is never reawakened. If the coin showed tails, she is fed a one-day amnesia potion (so that she does not remember being awake on Monday) and is reawakened on Tuesday, again without being told what day it is. At the end of Tuesday, she is left to sleep again (see (Elga, 2000)). For the purpose of this paper, she will be subsequently awoken on the next Sunday, to get on with her life and enjoy any gains she may have accrued during the experiment.

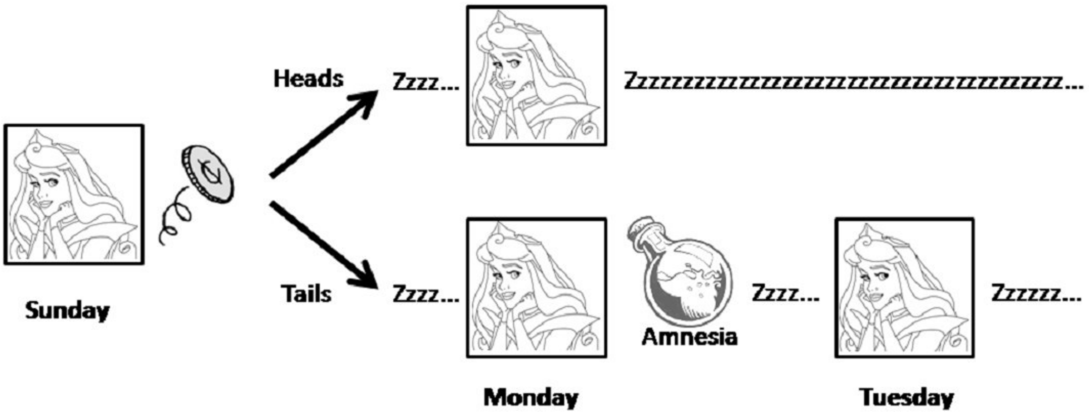

Figure 1: The Sleeping Beauty Problem

The implausibility of the setup need not concern us much; there are more realistic variants, such as the Sailor's Child problem, but the Sleeping Beauty formulation best at focusing on the core issues.

The question then is what probability a recently awoken or created Sleeping Beauty should give to the coin falling heads or tails and it being Monday or Tuesday when she is awakened (alternately, she could be asked for her betting odds on this issue – a more meaningful question, as we shall see). For the probability calculation, there are two main schools of thought, making use of either the Self-Sampling Assumption, or the Self-Indication Assumption (Bostrom, Anthropic Bias: Observation Selection Effects in Science and Philosophy, 2002).

2.1 The Self-Sampling Assumption

The self-sampling assumption (SSA) relies on the insight that Sleeping Beauty, before being put to sleep on Sunday, expects that she will be awakened in future. Thus her awakening grants her no extra information, and she should continue to give the same credence to the coin flip being heads as she did before, namely 1/2.

In the case where the coin is tails, there will be two copies of Sleeping Beauty, one on Monday and one on Tuesday, and she will not be able to tell, upon awakening, which copy she is. She should assume that both are equally likely. This leads to SSA:

> Self-Sampling Assumption:
> All other things being equal, an observer should reason as if they are randomly selected from the set of all *actually existent* observers (past, present and future) in their reference class.

There are some issues with the concept of 'reference class' (Bostrom, Anthropic Bias: Observation Selection Effects in Science and Philosophy, 2002), but here it is enough to set the reference class to be the set of all other Sleeping Beauties woken up in the experiment.

Given this, the probability calculations become straightforward:

$$\begin{gathered} P_{SSA}\ Heads\ = 1/2 \\ P_{SSA}\ Tails\ = 1/2 \\ P_{SSA}\ Monday|Heads\ = 1 \\ P_{SSA}\ Tuesday|Heads\ = 0 \\ P_{SSA}\ Monday|Tails\ = 1/2 \\ P_{SSA}\ Tuesday|Tails\ = 1/2 \end{gathered}$$

By Bayes' theorem, these imply that:

$$\begin{gathered} P_{SSA}\ Monday\ = 3/4 \\ P_{SSA}\ Tuesday\ = 1/4 \end{gathered}$$

2.2 The Self-Indication Assumption

There is another common way of doing anthropic probability, namely to use the self-indication assumption (SIA). This derives from the insight that being woken up on Monday after a heads, being woken up on Monday after a tails, and being woken up on Tuesday are all subjectively indistinguishable events, which each have an objective probability 1/2 of happening, therefore we should consider them equally probable. This is formalised as:

> Self-Indication Assumption:
> All other things being equal, an observer should reason as if they are randomly selected from the set of all *possible* observers.

Note that this definition of SIA is slightly different from that used in Bostrom (Bostrom, Anthropic Bias: Observation Selection Effects in Science and Philosophy, 2002); what we would call SIA he designated as the combined SIA+SSA. We shall stick with the definition above, however, as it is coming into general use. Note that there is no mention of reference classes, as one of the great advantages of SIA is that any reference class will do, as long as it contains the observers in question.

Given SIA, the three following observer situations are equiprobable (each has an 'objective' probability 1/2 of happening), and hence SIA gives them equal probabilities 1/3:

$$P_{SIA}\ Monday\ \&\ Heads\ \ = 1/3$$
$$P_{SIA}\ Monday\ \&\ Tails\ \ = 1/3$$
$$P_{SIA}\ Tuesday\ \&\ Tails\ \ = 1/3$$

This allows us to compute the probabilities:

$$P_{SIA}\ Monday\ \ = 2/3$$
$$P_{SIA}\ Tuesday\ \ = 1/3$$
$$P_{SIA}\ Heads\ \ = 1/3$$
$$P_{SIA}\ Tails\ \ = 2/3$$

SIA and SSA are sometimes referred to as the thirder and halfer positions respectively, referring to the probability they give for Heads.

3 Beyond probability

The debate between SIA and SSA can be intense, with many formal and informal arguments given in favour of each position. There is a quite common 'practical' argument for SIA, which can be distilled as:

> Sleeping Beauty should bet that tails came up at any odds better than 2:1, because if *she* does so, *she* will end up winning twice on tails and once on heads, thus *she* comes out ahead on expectation. If asked ahead of time, that's the course of action*she* would have decided upon.

There are many interesting things about this argument, all of them of relevance to this paper. First of all, note that the argument is not given in terms of anthropic probability, but in terms of behaviour: Sleeping Beauty *should* bet this way, because she will benefit from doing so. But who exactly is this "she"? The term is used four times in the argument; the first time it refers to an awakened Sleeping Beauty being given a choice of what odds to bet on. The second time refers to all the awakened versions in her time-line. The third one seems to refer to Sleeping Beauty after her final re-awakening, when she is free to enjoy the benefits of all the bets. And the fourth one refers to Sleeping Beauty before the whole experiment is run. Crucially, all of these are assumed to have the same interests.

The conditional 'if *she* does so, *she* will end up winning twice on tails...' similarly deserves scrutiny: the first *she* is the currently reasoning agent, whereas the second concerns both her and the other version of her: so the decision of one versionis taken to influence that of the other. Finally, the 'if asked ahead of time' clause brings up the issue of precommitments. This paper will look at all these issues in turn.

3.1 Probabilities or decisions

The purpose of expected utility maximisation (Neumann & Morgenstern, 1944) is to arrive at the correct decision in a particular situation: the probability, utility function and similar components are simply tools for this purpose. Every decision has three components: the probabilities, the utility function, and the impact of each possible option that the agents can choose.

In general, however, the probability is taken to be the most important variable, with the decision as a simple direct consequence of it. This is valid in most non-anthropic situations, as the utility functions and the impact of each option are taken to be fixed. Uncertainty about utility functions is a subtle issue, and is often taken to be irrelevant to the situation at hand. In contrast, uncertainty about the impact of each option is either taken to not exist, or submersed into the expected utility value for each option. Hence the probability seems to be the only variable of interest. Thus talking about probabilities in terms of either "degree of belief" (Finetti, 1970) or "betting odds" becomes interchangeable.

Anthropic situations are quite different, however, and probability no longer co-varies so strictly with the decision. Anthropic questions are not only about similaragents who may or may not exist; they are about agents who may or may not be responsible for each other's decisions, and who may or may not care about each other. These factors correspond to probability, impact of taking an option, and utility; in most anthropic problems, the first is overemphasised, the second has a philosophical uncertainty that cannot be reduced to an expectation value, and the third is underspecified.

3.2 Individual or linked decisions

Whether an agent is responsible for actions they do not directly cause is the conundrum explored in the Newcomb problem (Nozick, 1969), on which causal (Lewis, 1981) and evidential (Gibbard & Harper, 1981) decision theory diverge.

The issue can be seen more easily in Psy-Kosh's non-anthropic problem (Psy-Kosh, 2009), adapted here as a variant of the Sleeping Beauty problem. There are two agents sitting inseparate rooms when a coin is flipped. If it comes out heads, one agent is asked to vote on whether the coin was heads or tails. If it comes out tails, both agents are asked to vote. In the heads world, if the single decider voted 'heads', £10 would be paid to a charity both agents approve of; if he voted 'tails', £7 would be paid to it. In the heads world, if they both vote 'heads', the charity would get £8, while if they both voted 'tails' it would get £10. If they voted differently, the charity would get nothing. The payoffs are summarised in this table:

Coin flip	Vote	Payoff
Heads	heads	£10
	tails	£7
Tails	both heads	£8
	both tails	£10
	agents vote differently	£0

Table 1: A non-anthropic version of the sleeping beauty problem

It is now no longer obvious what the impact of saying 'heads' versus 'tails' is. While clear for the single agent in the heads world, the impact in the tails world depends upon the vote of the other agent. To simplify the conundrum, we can assume the agents use similar methods of decision, and so their votes are linked: they will both vote the same way (and they are both aware of this fact).

There are at least three possible stances that could be taken here. A strict causal decision theorist would claim that there is no way in which one agent's decision will cause the other's, so this linking is irrelevant. They will attempt to estimate what the other agent would decide using independent means, and only then will they make their decision in consequence – a process complicated by the fact the other agent is trying to do the same thing.

Or they could admit the linking, and reason as if they were controlling both agents simultaneously, an approach closer to evidential decision theory. They could take either total or a divided responsibility for the outcome in this case. For total responsibility, each agent reasons as if their vote makes the whole of the difference – after all, the world in which they vote heads is a world in which all agents vote heads, and vice versa. For divided responsibility, they divide the outcome by the number of voting agents, arguing that it makes no sense that both agents claim total responsibility for the same outcome.

We can use these last two 'impact theories' to get the decisions, assuming utility linear in money. The probability of being asked to vote is 0.5 in the heads world and 1 in the tails world. For total responsibility, saying 'heads' produces(0.5(£10)+1(£8))/1.5=£26/3 of expected cash, while saying 'tails' produces (0.5(£7)+1(£10))/1.5=£27/3 expected cash. For divided responsibility, 'heads' and 'tails' produce (0.5(£10)+1(£8/2))/1.5=£18/3 and (0.5(£7)+1(£10/2))/1.5=£17/3 respectively. So under total responsibility the agent(s) will vote 'tails' and under divided responsibility the agent(s) will vote 'heads'.

These would be the same decisions given by SIA agents in the equivalent anthropic situation.SSA agents would put equal probabilities on heads and tails, and so will have expected utilities of 0.5(£10)+0.5(£8)=£18/2 for heads and 0.5(£7)+0.5(£10)=£17/2 for tails under total responsibility, and 0.5(£10)+0.5(£8/2)=£7 for heads and 0.5(£7)+0.5(£10/2)=£6 for tails under divided responsibility.

It should be noted that SIA with divided responsibility give the same estimated expected utility as SSA with total responsibility, up to an unimportant 2/3 scaling factor. This isn't a coincidence – the extra probability weight SIA gives to universes with many agents is equivalent to the extra impact weight total responsibility does. This means that SSA agents with total responsibility would always make the same decision as SIA agents with divided responsibility. Which means that proclaiming 'SIA' or 'SSA' is not enough to determine behaviour; conversely, it is not possible to deduce the probability system used from utility functions and behaviour. "Degree of belief" and "betting odds" come apart.

Since anthropic problems are often full of identical or linked agents, determining the impact of their choices would be just as important as determining the right anthropic probability system to use.

3.3 Selfish or Selfless utilities

Using total or divided responsibility or causal decision theory is an important philosophical distinction that can drastically change the situation, so it is important to specify this in any decision problem. In theory, there is no similar problem with the utility function: any sufficiently specified utility function is enough to compute the decision. The problem is instead practical: in anthropic problems, the utility function is rarely sufficiently specified.

With identical agents in many anthropic problems, some of whom are the same agent at different moment, the generic 'selfish' assumption for each individual preference is no longer a natural default. And when agents are making linked decisions, the difference can no longer be ignored: if one agent's decision is linked with others, then they will compute different utilities depending on whether they care about their other versions or not. Thus not specifying whether agents are selfish or selfless (or some mix in between) is akin to only half specifying the utility function.

The difference can be illustrated by a slight modification to the non-anthropic problem of Table 1. The rewards are now given directly to the voters, not to the charity, and in the tails world, if the pair votes 'tails' they each get £5, while voting 'heads' getsthem each £4. Their decisions are still linked, as before. For selfless agents, this setup is exactly the same as before: their personal monetary gain is halved in the tails world, but this is compensated by the fact that the other agent's monetary gain counts in their own utility.

If the agents are selfish, on the other hand, all their utility gains are indeed halved in the tails world, and both total and divided responsibility will now vote 'heads'. Similarly to how 'SIA' + 'divided responsibility' was seen to cause the same choices as 'SSA' + 'total responsibility', we get another isomorphism between 'selfless' + 'divided responsibility' and 'selfish' + 'total responsibility'. With large populations, the difference between selfish and selfless can utterly dominate other considerations: see for instance (Armstrong, 2012) on the Presumptuous Philosopher problem (Bostrom, Anthropic Bias: Observation Selection Effects in Science and Philosophy, 2002). Hence the issue cannot and should not be ignored when anthropic problems are posed.

3.4 Precommitments

Another important problem in decision theory is Parfit's hitchhiker (Parfit, 1984) and similar situations. Here, a starving hitchhiker meets a driver who could drive him to the nearest town. The hitchhiker promises to pay the driver a reward upon arrival. Unfortunately, both the driver and hitchhiker are selfish and perfectly rational, and realise the hitchhiker will have no incentive to pay once he is driven to the town. So the driver motors away, leaving him to die. There is a lot of literature on this subject, which will not be summarised here; the main point being that it is advantageous for an agent to be able to make binding precommitments.

Looking at the non-anthropic problem in Table 1 from the outside, we could argue that the right course of action is for the agents to say 'heads' – if all the agents did that, the charity's expected gain is higher. So it might seem that 'divided responsibility' is a more correct approach. But this neglects one key player – the agent in the heads world who was not selected to vote. If all agents were allowed to make binding precommitments before any of them is selected, they would choose to precommit for 'heads', whether or not they were using divided or total responsibility. This is because 'selecting for voting' changes the size of the 'population' under consideration (going down from two relevant agents to one in the heads world), while an early precommitment 'locks in' the original 'population size'. And for fixed population size, divided and total responsibility are the same, matching up in the same way as total and average utilitarianism do.

Very early precommitments can be seen as a rawlsian veil of ignorance' situation (Rawls, 1971), making even selfish agents want to choose globally beneficent outcomes, since they don't know what their position will be. So the potential for binding precommitment also affects the difference between selfish and selfless utility functions. Consider for instance the selfish variant of the non-anthropic problem in the previous section. Precommitments before voters are chosen have no effect in the tails world, but do have an impact in the heads world: there, one of the agents has a 50% chance of not being given a vote. So when computing the difference between 'precommit to saying heads' and 'precommit to saying tails', the agents must take into consideration the fact they may never get to vote. This halves the expected gain in the heads world, and, since the gains were already halved in the tails world because of the agents' selfishness, this returns the choices to those in the original selfless problem.

4 Conclusion

Instead of one contentious issue in anthropic problems – whether to use SIA or SSA – it now seems that there are many, with the issues of total or divided responsibility, selfishness or selflessness towards other copies, and the existence of possible precommitments adding on to the tally. But this explosion of complexity is partially illusionary, as there are less degrees of freedom than it might seem: 'SIA' + 'divided responsibility' results in the same decisions as 'SSA' + 'total responsibility', as do 'selfless' + 'divided responsibility' and 'selfish' + 'total responsibility'. And precommitments can further remove some of these divergences.

Thus the space of possible decision algorithms that would result in actually different choices is smaller than it would seem. And the anthropic probabilities can no longer be considered as encoding the genuine uncertainty in the problem: fixing the utility function and (especially) choosing the way of assigning responsibility to the impact of linked decision encodes other critical information about how to deal with the decision problem.

Finally, it should be noted that a lot of anthropic decision problems can be solved without needing to work out the anthropic probabilities and impact responsibility at all (see for instance the approach in (Armstrong, 2012)). The 'practical' argument given in section 3 works: *if* there is a future copy to benefit from all their gains, *if* all the Sleeping Beauties are altruistic towards this future copy, and *if* they can make binding precommitments, then betting on better than 2:1 odds on tails is the correct decision.

Bibliography

Armstrong. (2012). Anthropic decision theory for self-locating beliefs. *submitted for publication*.

Aumann, R. J., Hart, S., & Perry, M. (1996). Absent-minded driver. *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*.

Bostrom, N. (2001). The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe. *Synthese, 127*(3).

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy.* New York: Routledge.

Carter, B. (1974). Large Number Coincidences and the Anthropic Principle in Cosmology. *IAU Symposium 63: Confrontation of Cosmological Theories with Observational Data* (pp. 291–298). Dordrecht: Reidel.

Carter, B., & McCrea, W. H. (1983). The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society of London, 310*(1512), 347-363.

Dieks, D. (2007). Reasoning about the future: Doom and Beauty. *Synthese, 156*(3), 427-439.

Elga, A. (2000). Self-locating Belief and the Sleeping Beauty Problem. *Analysis, 60*, 143-147.

Finetti, B. d. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica, 34*, 129–145.

Gibbard, A., & Harper, W. L. (1981). Counterfactuals and two kinds of expected utility. In *Ifs: Conditionals, Beliefs, Decision, Chance, and Time* (pp. 153-190).

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy, 59*(1), 5-30.

Neal, R. M. (2006). Puzzles of anthropic reasoning resolved using full non-indexical conditioning. *Arxiv preprint math/0608592*.

Neumann, J. v., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior.* Princeton, NJ, Princeton University Press.

Nozick, R. (1969). Newcomb's Problem and Two principles of Choice. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (pp. 114-115). Dordrecht: Reidel.

Parfit, D. (1984). *Reasons and Persons.* Oxford: Clarendon Press.

Psy-Kosh. (2009). *Psy-Kosh comments on Outlawing Anthropics: An Updateless Dilemma.* Retrieved from Less Wrong: http://lesswrong.com/lw/17c/outlawing_anthropics_an_updateless_dilemma/13e1

Rawls, J. (1971). *A Theory of Justice.* Cambridge, Massachusetts: Belknap Press.