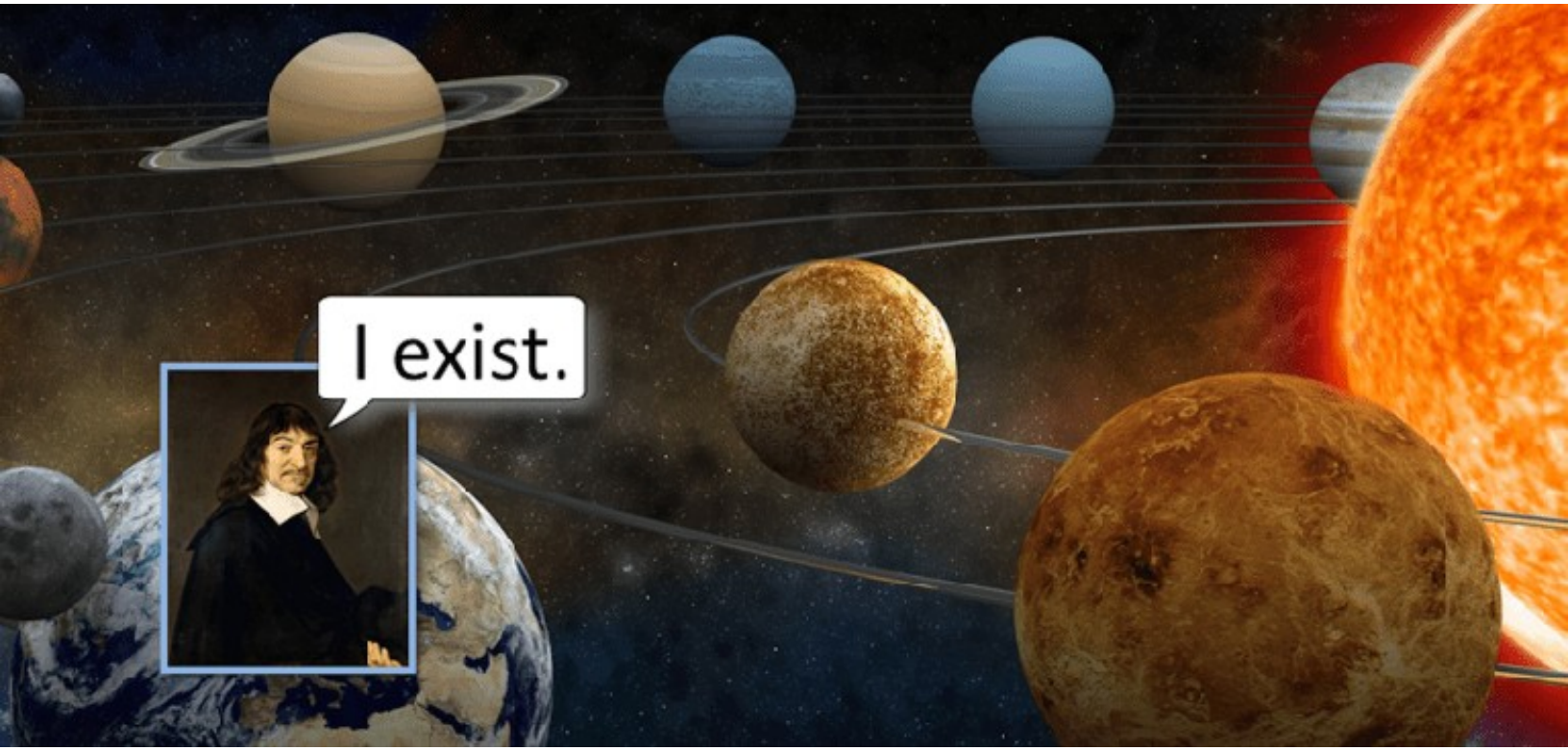

I exist.

Practical Guide to Anthropics

Anthropic decision theory for self-locating beliefs

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is a link post to the "Anthropic decision theory for self-locating beliefs" paper, with the abstract:

> This paper sets out to resolve how agents ought to act in the Sleeping Beauty problem and various related anthropic (self-locating belief) problems, not through the calculation of anthropic probabilities, but through finding the correct decision to make. It creates an anthropic decision theory (ADT) that decides these problems from a small set of principles. By doing so, it demonstrates that the attitude of agents with regards to each other (selfish or altruistic) changes the decisions they reach, and that it is very important to take this into account. To illustrate ADT, it is then applied to two major anthropic problems and paradoxes, the Presumptuous Philosopher and Doomsday problems, thus resolving some issues about the probability of human extinction.

The key points of that paper are also available in this post sequence.

Anthropics: different probabilities, different questions

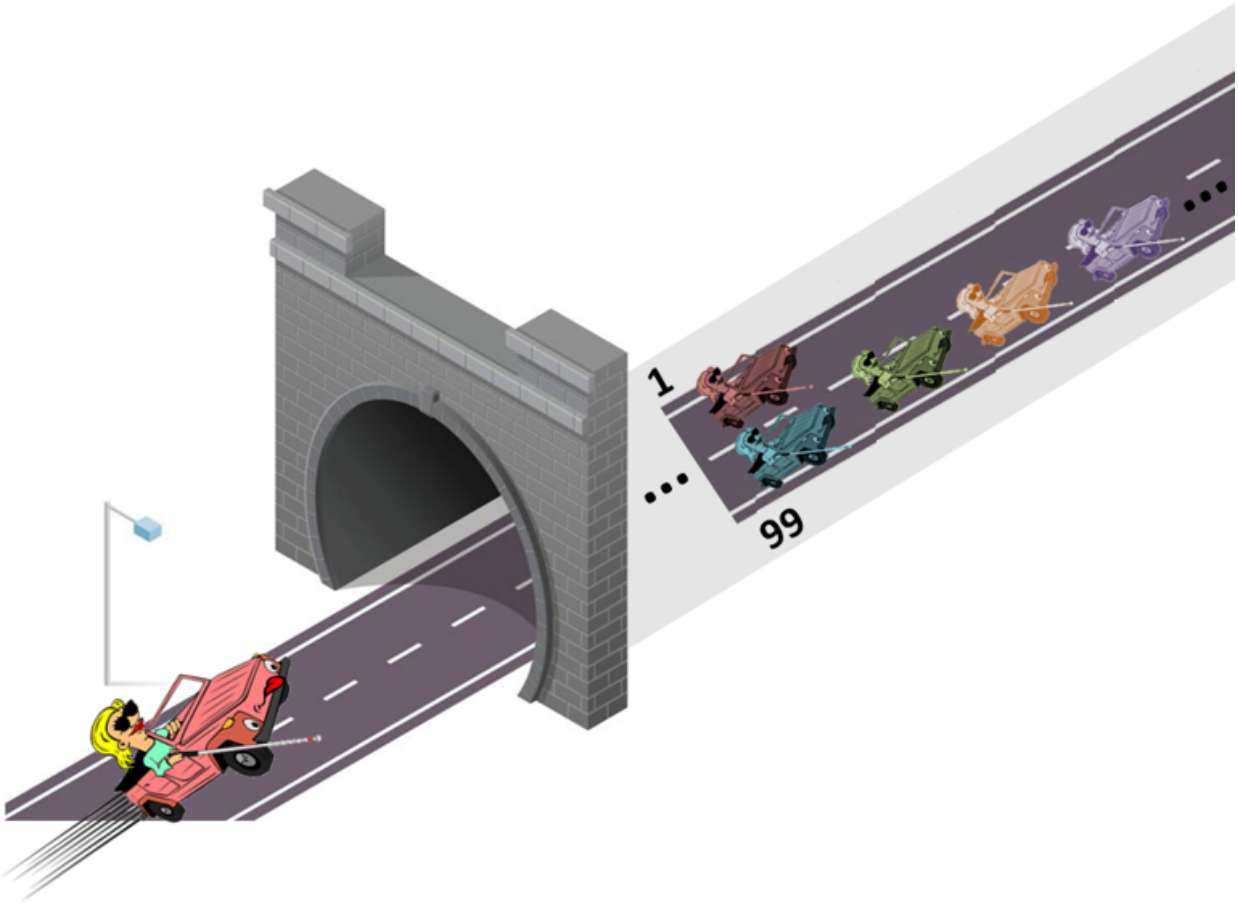
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

I've written before that different theories of anthropic probability are really answers to different questions. In this post I'll try to be as clear as possible on what that means, and explore the implications.

Introduction

One of Nick Bostrom's early anthropic examples involved different numbers of cars in different lanes. Here is a modification of that example:

> You're driving along, when you turn into a dark tunnel and are automatically shunted into the left or the right lane. You can't see whether there are any other cars in your dark lane, but the car radio announces "there are 99 cars in the right lane and 1 in the left lane".



Given that, what is your probability of being in the left lane?

That probability is obviously 1%. More interesting than that answer, is that there are multiple ways of reaching it. And each of these ways corresponds to answering a slightly different question. And this leads to my ultimate answer about anthropic probability:

- Each theory of anthropic probability corresponds to answering a specific, different question about proportions. These questions are equivalent in non-anthropic setting, so each of them feels potentially like a "true" extension of probability to anthropics. Paradoxes and confusion in anthropics results from confusing one question with another.

So if I'm asked "what's the 'real' anthropic probability of X?", my answer is: tell me what you mean by probability, and I'll tell you what the answer is.

0. The questions

If X is a feature that you might or might not have (like being in a left lane), here are several questions that might encode the probability of X:

1. What proportion of potential observers have X?
2. What proportion of potential observers *exactly like you* have X?
3. What is the average proportion of potential observers with X?
4. What is the average proportion of potential observers *exactly like you* with X?

We'll look at each of these questions in turn[1], and see what they say imply in anthropic and non-anthropic situations.

1. Proportion of potential observers: SIA

We're trying to answer "Given *that*, what is your probability of being in the left lane?" The "that" is means being in the tunnel in the above situations, so we're actually looking for a conditional probability, best expressed as:

1. What proportion of the potential observers, who are in the tunnel in the situation above, are also in the left lane?

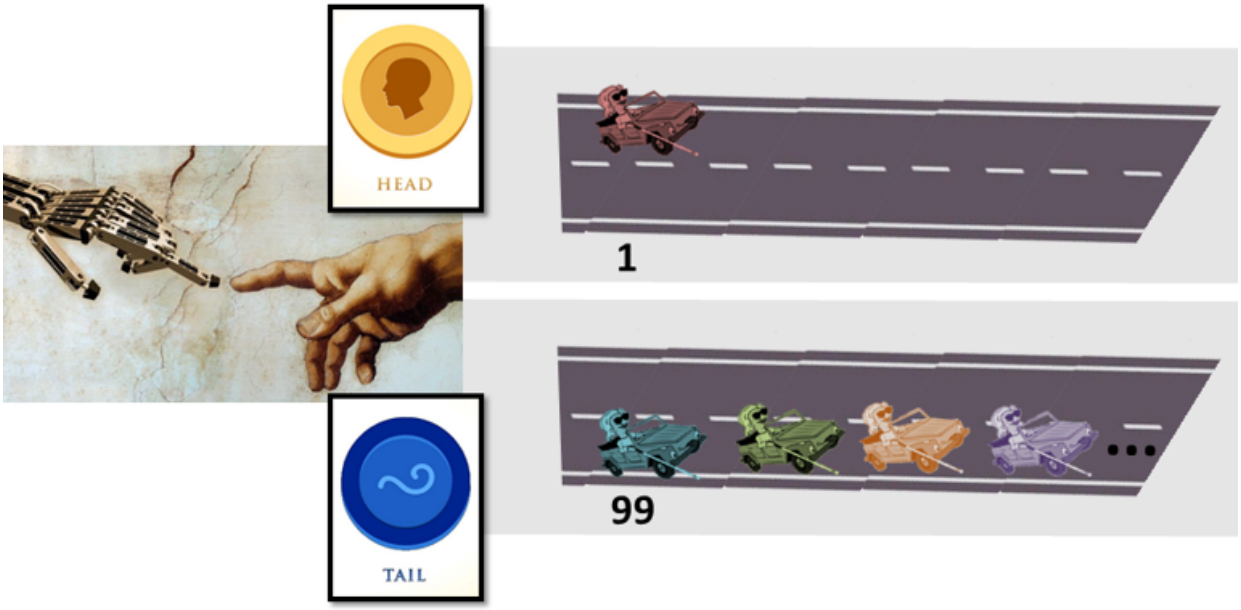
The answer for that is an immediate "one in a hundred", or 1%, since we know there are 100 drivers in the tunnel, and 1 of them is in the left lane. There may be millions of different tunnels, in trillions of different potential universes; but, assuming we don't

need to worry about infinity[2], we can count 100 observers in the tunnel in that situation for each observer in the left lane.

1.1 Anthropic variant

Let's now see how this approach generalises to anthropic problems. Here is an anthropic version of the tunnel problem, based on the incubator version of the Sleeping Beauty problem:

> A godly AI creates a tunnel, then flips a fair coin. If the coin comes out heads, it will create one person in the tunnel; if it was tails, it creates 99 people.
>
> You've just woken up in this tunnel; what is the probability that the coin was heads?



So, we want to answer:

1. What proportion of the potential observers, who are in the tunnel, are also in a world where the coin was heads?

We can't just count off observers within the same universe here, since the 99 and the 1 observers don't exist in the same universe. But we can pair up universes here: for each universe where the coin flip goes heads (1 observer), there is another universe of equal probability where the coin flip goes tails (99 observers).

So the answer to the proportion of potential observers question remains 1%, just as in the non-anthropic situation.

This is exactly the "self-indication assumption" (SIA) version of probability, which counts observers in other potential universes as if they existed in a larger multiverse of

potential universes[3].

2. Proportion of potential observers exactly like you: SIA again

Let's now look at the second question:

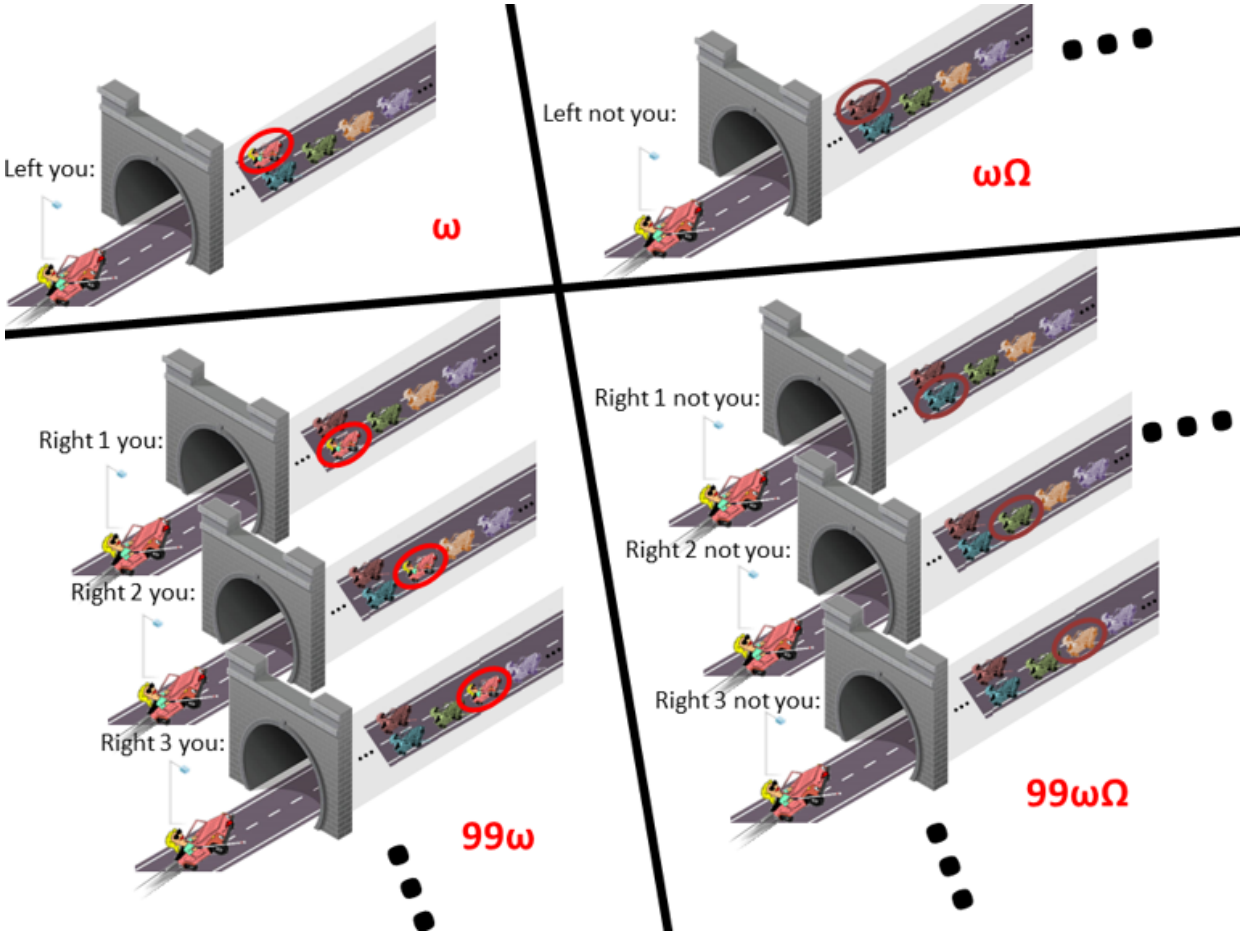
2. What proportion of the potential observers exactly like you, who are in the tunnel in the situation above, are also in the left lane?

The phrase "exactly like you" is underdefined - do you require that the other yous be made of exactly the same material, in the same location, etc... I'll cash out the phrase as meaning "has had the same subjective experience as you". So we can cash out the left-lane probability as:

2. What proportion of the potential observers, with the same subjective experiences as you, who are in the tunnel in the situation above, are also in the left lane?

We can't count off observers within the same universe for this, as the chance of having multiple observers with the same subjective experience in the same universe is very low, unless there are huge numbers of observers.

Instead, assume that one in Ω observers in the tunnel have the same subjective experiences as you. This proportion[4] must be equal for an observer in the left and right lanes. If it weren't, you could deduce information about which lane you were in just from your experiences - so the proportion being equal is the same thing as the lane and your subjective experiences being independent. For any given little ω, this gives the following proportions (where "Right 1 not you" is short for "the same world as 'Right 1 you,' apart from the first person on the right, who is replaced with a non-you observer"):



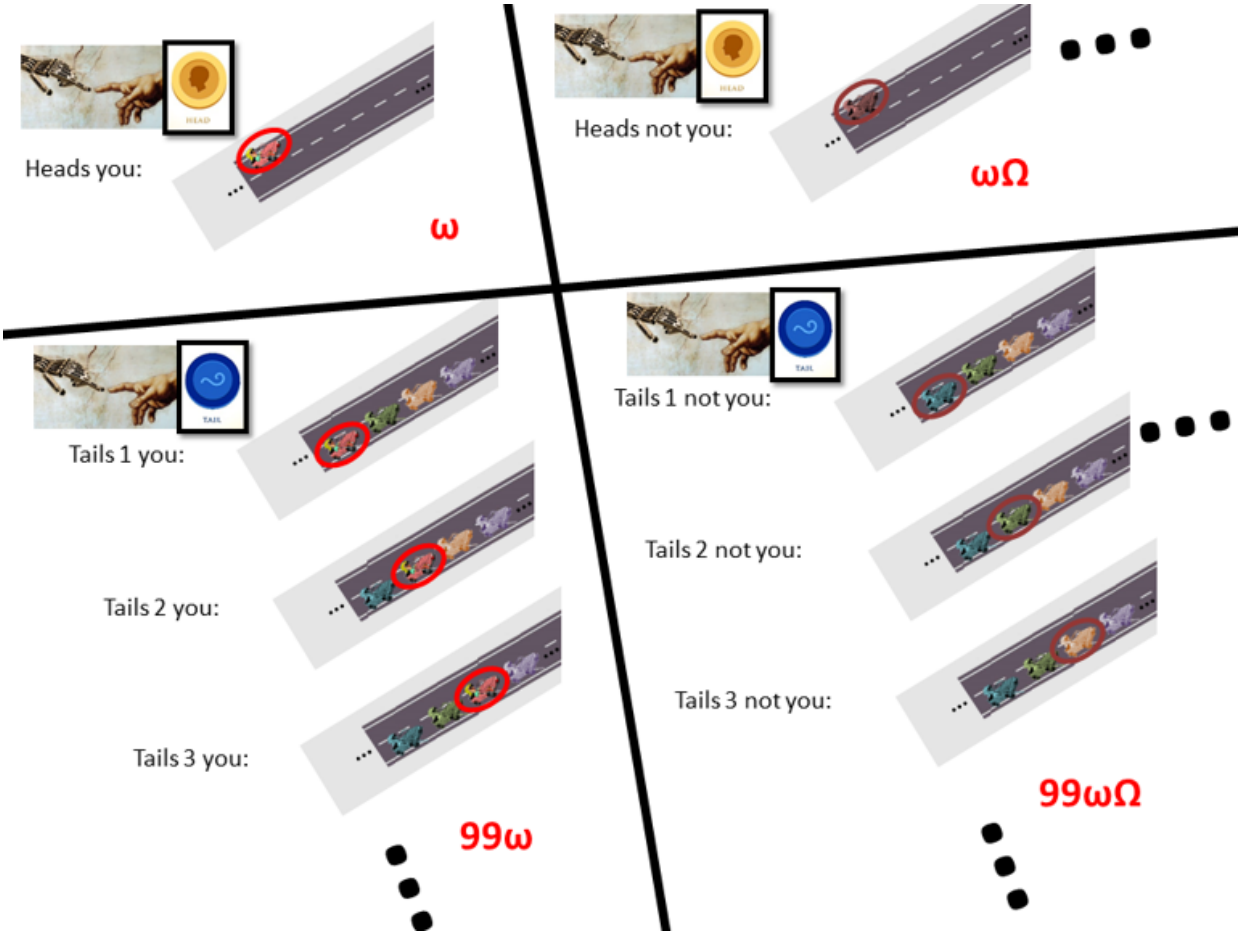
So the proportion of observers in the right/left lane with your subjective experience is 1/Ω the proportion of observers in the right/left lane. When comparing those two proportions, the two 1/Ω cancel out, and we get 1%, as before.

2.1 Anthropic variant

Ask the anthropic version of the question:

2. What proportion of the potential observers who are in the tunnel, with the same subjective experiences as you, are also in a world where the coin was heads?

Then same argument as above shows this is also 1% (where "Tails 1 not you" is short for "the same world as 'Tails 1 you,' apart from the first tails person, who is replaced with a non-you observer"):



This is still SIA, and reflects the fact that, for SIA, the reference class doesn't matter - as long as it include the observers subjectively indistinguishable from you. So questions about you are the same whether we talk about "observers" or "observers with the same subjective experiences as you".

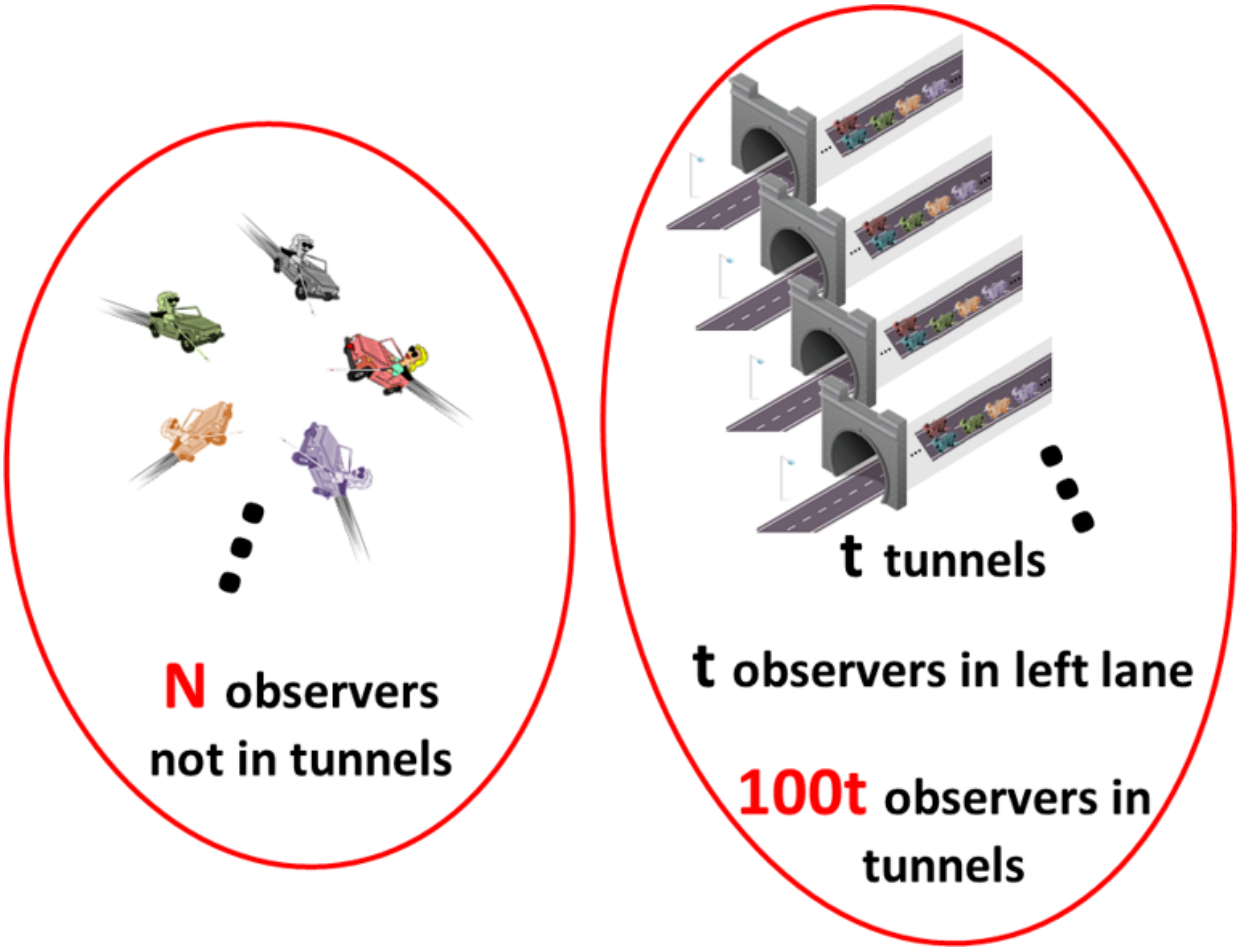
3. Average proportions of observers: SSA

We now turn to the next question:

3. What is the average proportion of potential observers in the left lane, relative to the average proportion of potential observers in the tunnel?

Within a given world, say there are N observers not in the tunnel and t tunnels, so

N + t100 observers in total.



The proportion of observers in the left lane is t/(N + t100) while the proportion of observers in the tunnel is 100t/(N + t100). The ratios of the these proportions in 1 : 100 .

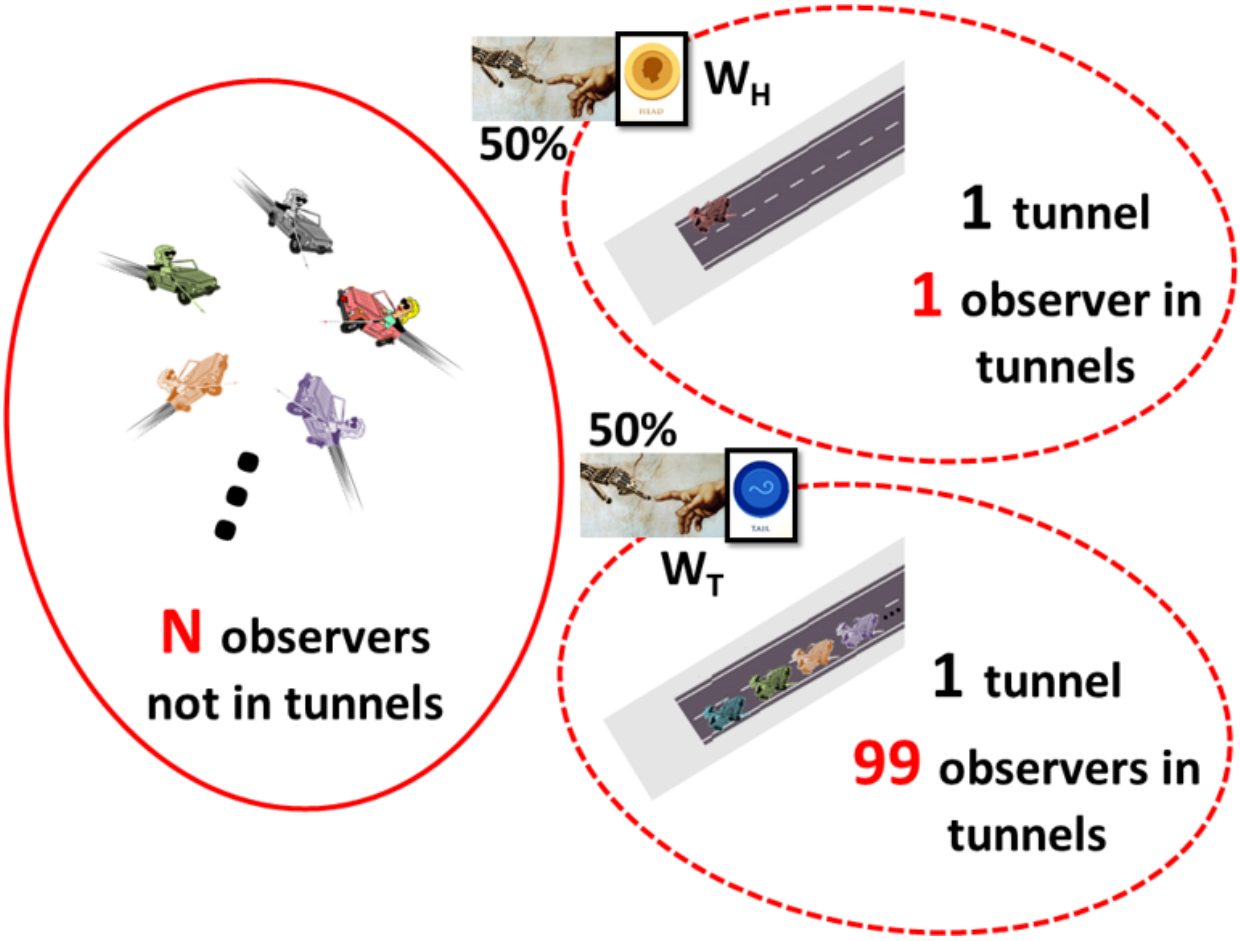
Then notice that if a and b are in a 1 : 100 proportion in every possible world, the averages of a and b are in a 1 : 100 proportion as well[5], giving the standard probability of 1%.

3.1 Anthropic variant

The anthropic variant of the question is then:

3. What is the average proportion of potential observers in a world where the coin was heads, relative to the average proportion of potential observers in the tunnel?

Within a given world, ignoring the coin, say there are N observers not in the tunnel, and t tunnels. Let's focus on the case with one tunnel, t = 1. Then the coin toss splits this world into two equally probable worlds, the heads world, W_H, with N + 1 observers, and the tails world, W_T with N + 99 observers:



The proportion of observers in tunnels in W_H is $\frac{1}{N+1}$. The proportion of observers in tunnels in W_T is $\frac{99}{N+99}$. Hence, across these two worlds, the average proportion of observers in tunnels is the average of these two, specifically

$$\frac{1}{2}\left(\frac{1}{N+1} + \frac{99}{N+99}\right) = \frac{50N}{(N+1)(N+99)}.$$

If N is zero, this is 99/99 = 1; this is intuitive, since N = 0 means that all observers are in tunnels, so the average proportion of observers in tunnels is 1.

What about the proportion of observers in the tunnels in the heads worlds? Well, this is $\frac{1}{N+1}$ is the heads world, and 0 is the tails world, so the average proportion is:

$$\frac{1}{2}\left(\frac{1}{N+1} + 0\right) = \frac{1}{2(N+1)}.$$

If N is zero, this is 1/2 -- the average between 1, the heads world proportion for N = 0 in W_H (all observers are heads world observers in tunnels) and 0, the proportion of heads world observers in the tails world W_T.

Taking the ratio (1/2)/1 = 1/2, the answer to that question is 1/2. This is the answer given by the "self-sampling assumption" (SSA), with gives the 1/2 response in the sleeping beauty problem (of which this is a variant).

In general, the ratio would be:

$$\frac{1}{2(N+1)} \div \frac{50N + 99}{(N+1)(N+99)} = \frac{N+99}{100N+198}$$

If N is very large, this is approximately 1/100, i.e. the same answer as SIA would give.

This shows the fact that, for SSA, the reference class of observers is important. The N, the number of observers that are not in tunnel, define the probability estimate. So how we define observers will determine our probability[6].

So, for a given pair of worlds equally likely worlds, W_H and W_T, the ratio of question 3. varies between 1/2 and 1/100. This holds true for multiple tunnels as well. And it's not hard to see that this implies that, averaging across all worlds, we also get a ratio between 1/2 (all observers in the reference class are in tunnels) and 1/100 (almost no observers in the reference class are in tunnels).

4. Average proportions of observers exactly like you: FNC

Almost there! We have a last question to ask:

4. What is the average proportion of potential observers in the left lane, with the same subjective experiences as you, relative to the average proportion of potential observers in the tunnel, with the same subjective experiences as you?

I'll spare you the proof that this gives 1% again, and turn directly to the anthropic variant:

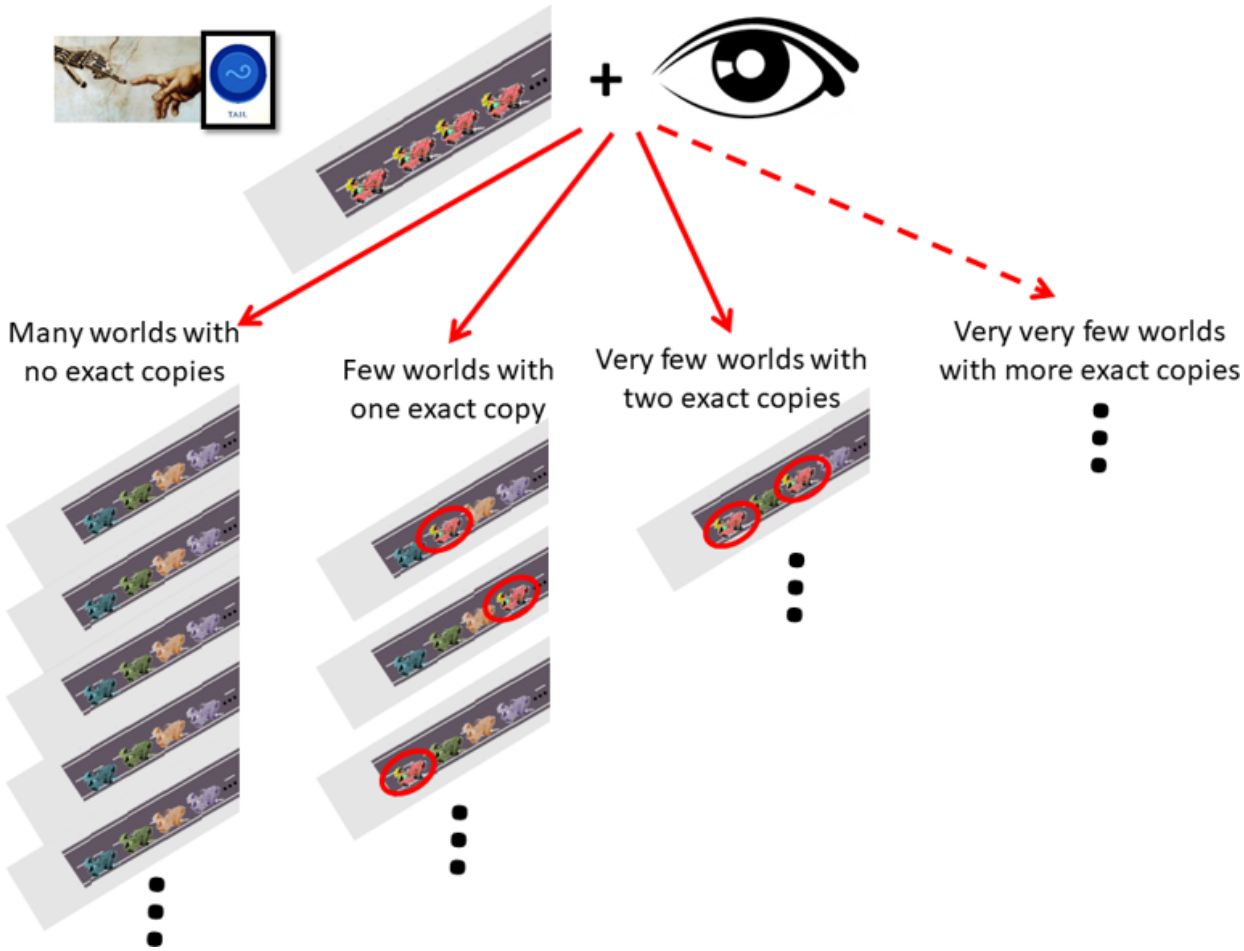
4. What is the average proportion of potential observers in a world where the coin was heads, with the same subjective experiences as you, relative to the average proportion of potential observers in the tunnel, with the same subjective experiences as you?

By the previous section, this is the SSA probability with the reference class of "observers with the same subjective experiences as you". This turns out to be FNC, full non-indexical conditioning (FNC), which involves conditioning on any possible observation you've made, no matter how irrelevant. It's known that if all the observers have made the same observations, this reproduces SSA, but that as the number of unique observations increases, this tends to SIA.

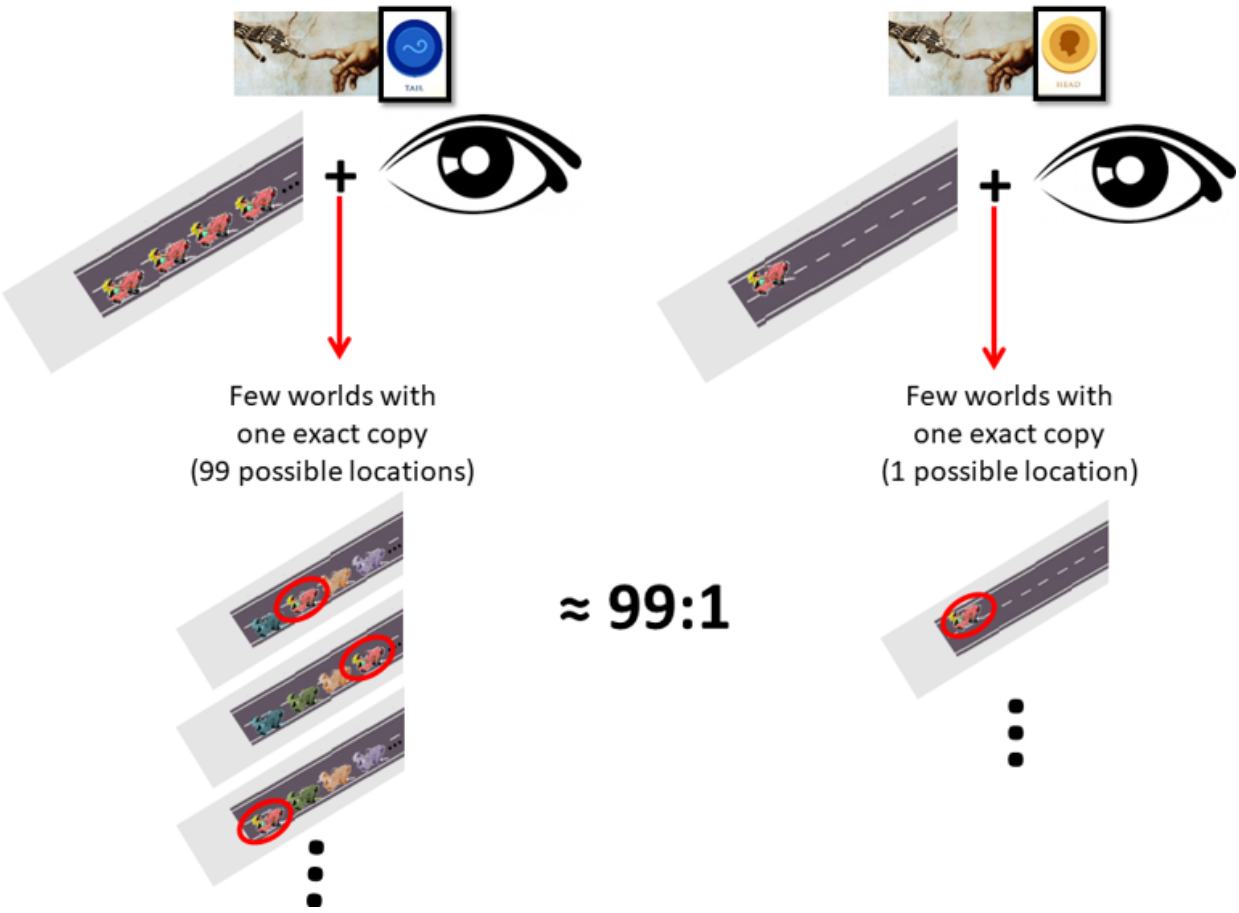
That's because FNC is inconsistent - the odds of heads to tails change based on irrelevant observations which change your subjective experience. Here we can see what's going on: FNC is SSA with the reference class of observers with the same subjective experiences as you. But this reference class is variable: as you observe more, the size of the reference class changes, decreasing[7] because others in the reference class will observe something different to what you do.

But SSA is not consistent across reference class changes! So FNC is not stable across new observations, even if those observations are irrelevant to the probability being estimated.

For example, imagine that we started, in the tails world, with all 99 copies exactly identical to you, and then you make a complex observation. Then that world will split in many worlds where there are no exact copies of you (since none of them made exactly the same observation as you), a few worlds where there is one copy of you (that made the same observation as you), and many fewer worlds where there are more than one copy of you:



In the heads world, we only have no exact copies and one exact copy. We can ignore the worlds without observers exactly like us, and concentrate one the worlds with a single observer like us (this represents the vast majority of the probability mass). Then, since there are 99 possible locations in the tails world and 1 in the heads world, we get a ratio of roughly 99 : 1 for tails over heads:



This give a ratio of roughly 100 : 1 for "any coin result" over heads, and shows why FNC converges to SIA.

5. What decision to make: ADT

There's a fifth question you could ask:

5. What is the best action I can take, given what I know about the observers, our decision algorithms, and my utility function?

This transforms transforms the probability question into a decision-theoretic question. I've posted at length on Anthropic Decision Theory, which is the answer to that question. Since I've done a lot of work on that already, I won't be repeating that work here. I'll just point out that "what's the best decision" is something that can be computed independently of the various versions of "what's the probability".

5.1 How right do you want to be?

An alternate characterisation of the SIA and SSA questions could be to ask, "If I said 'I have X', would I want most of my copies to be correct (SIA) or my copies to be correct in most universes (SSA)?"

These can be seen as having two different utility functions (one linear in copies that are correct, one that gives rewards in universes where my copies are correct), and acting to maximise them. See the post here for more details.

6. Some "paradoxes" of anthropic reasoning

Given the above, let's look again at some of the paradoxes of anthropic reasoning. I'll choose three: the Doomsday argument, the presumptuous philosopher, and Robin Hanson's take on grabby aliens.

6.1 Doomsday argument

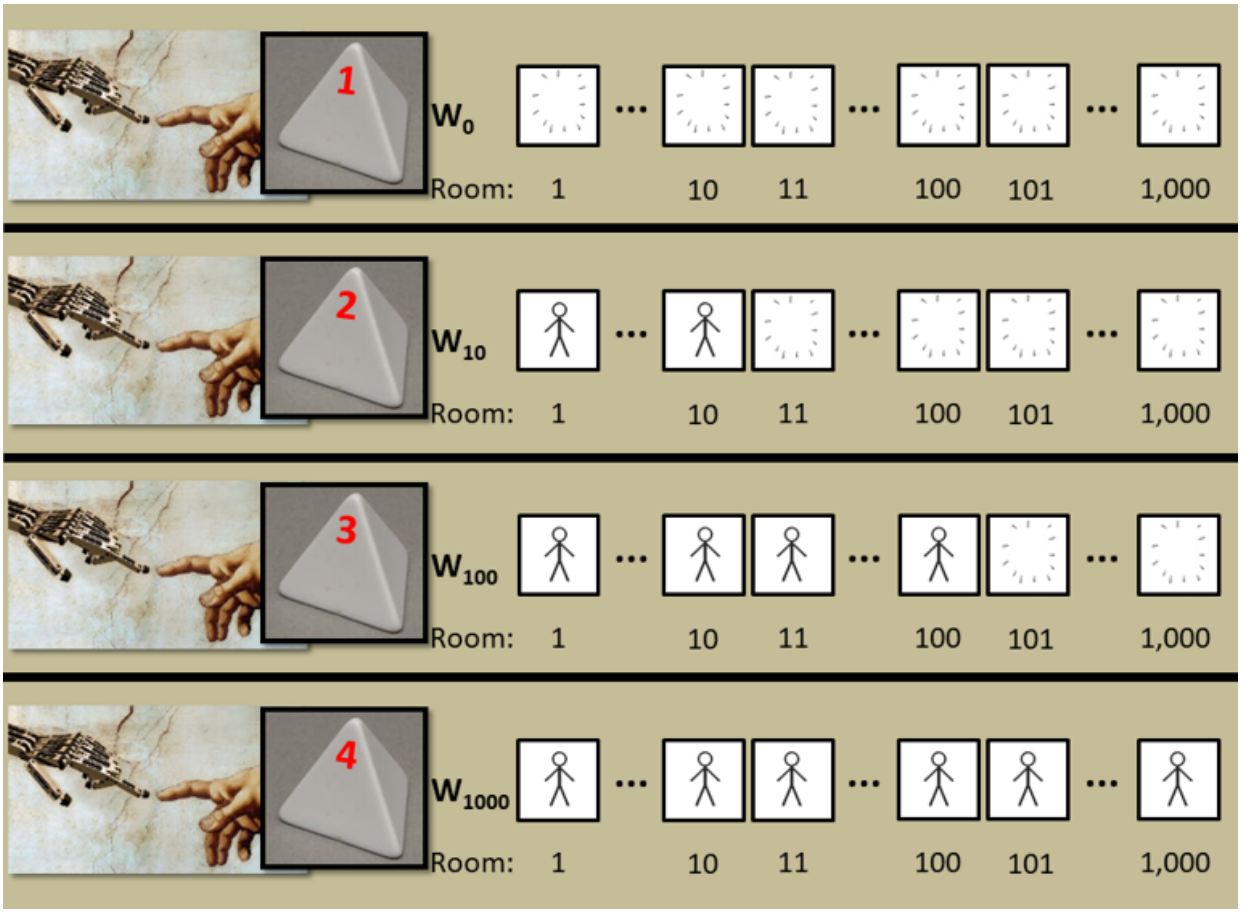
The Doomsday argument claims that the end of humanity is likely to be at hand - or at least more likely than we might think.

To see how the argument goes, we could ask "what proportion of humans will be in the last 90% of all humans who have ever lived in their universe?" The answer to that is, tautologically[8], 90%.

The simplest Doomsday argument would then reason from that, saying "with 90% probability, we are in the last 90% of humans in our universe, so, with 90% probability, humanity will end in this universe before it reaches 100 times the human population to date."

What went wrong there? The use of the term "probability", without qualifiers. The sentence slipped from using probability in terms of ratios within universes (the SSA version) to ratios of which universes we find ourselves in (the SIA version).

As an illustration, imagine that the godly AI creates either world W_0 (with 0 humans), W_{10} (with 10 humans), W_{100} (with 100 humans), or $W_{1,000}$ (with $1,000$ humans). Each option is with probability 1/4. These human are created in numbered room, in order, starting at room 1.



Then we might ask:

- A. What proportion of humans are in the last 90% of all humans created in their universe?

That proportion is undefined for W_0. But for the other worlds, the proportion is 90% (e.g. humans 2 through 10 for W_{10}, humans 11 through 100 for W_{100} etc...). Ignoring the undefined world, the average proportion is also 90%.

Now suppose we are created in one of those rooms, and we notice that it is room number 100. This rules out worlds W_0 and W_{10}; but the average proportion remains 90%.

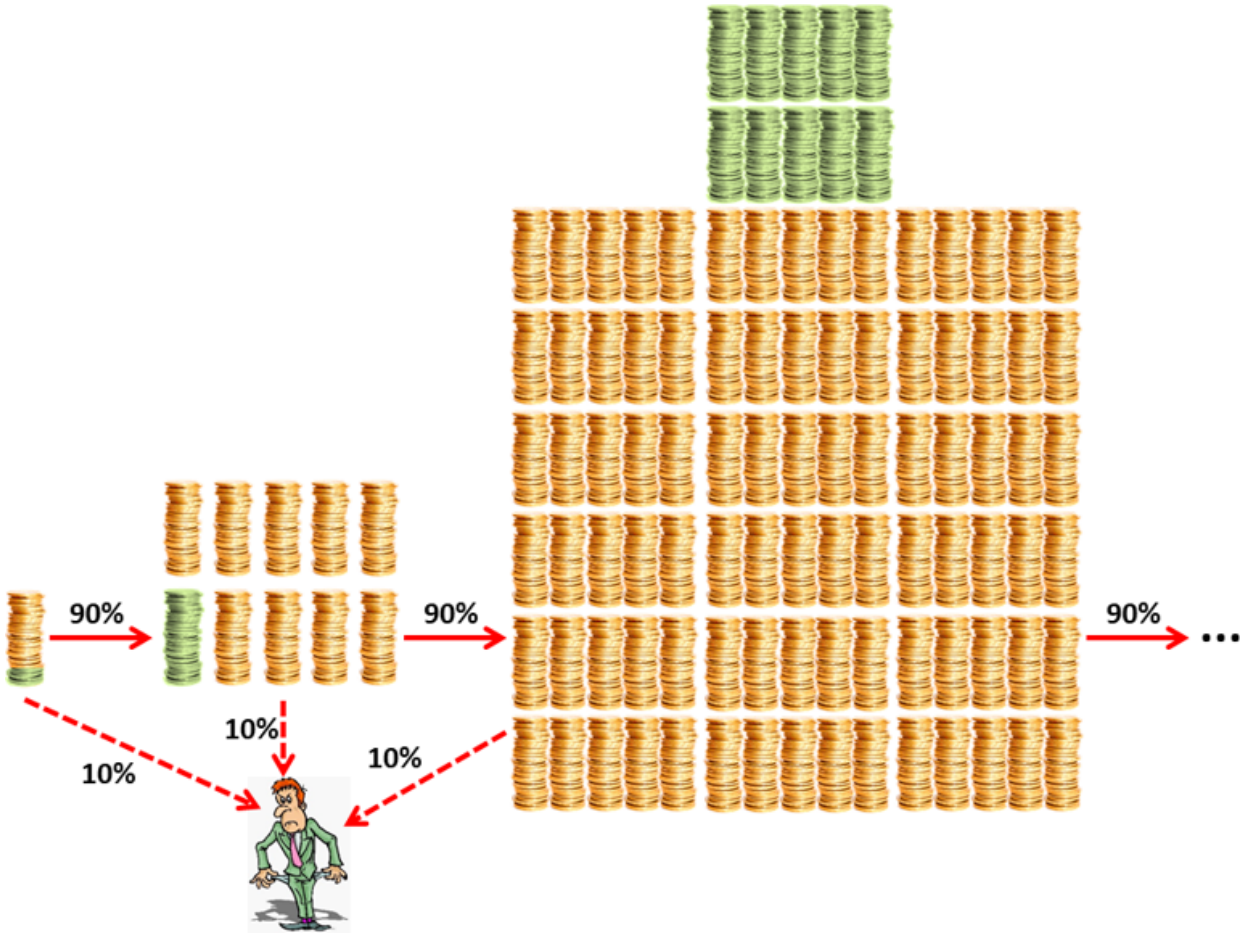
But we might ask instead:

- B. What proportion of humans in room 100 are in the last 90% of all humans created in their universe?

As before, humans being in room 100 eliminates worlds W_0 and W_{10}. The worlds W_{100} and $W_{1,000}$ are equally likely, and each have one human in room 100. In W_{100}, we are in the last 90% of humans; in $W_{1,000}$, we are not. So the answer to question B is 50%.

Thus the answer to A is 90%, the answer to B is 50%, and there is no contradiction between these.

Another way of thinking of this: suppose you play a game where you invest a certain amount of coins. With probability 0.9, your money is multiplied by ten; with probability 0.1, you lost everything. You continue re-investing the money until you lose. This is illustrated by the following diagram, (with the initial investment indicated by green coins):



Then it is simultaneously true that:

1. 90% of all the coins you earnt were lost the very first time you invested them, and
2. You have only 10% chance of losing any given investment.

So being more precise about what is meant by "probability" dissolves the Doomsday argument.

6.2 Presumptuous philosopher

Nick Bostrom introduced the presumptuous philosopher thought experiment to illustrate a paradox of SIA:

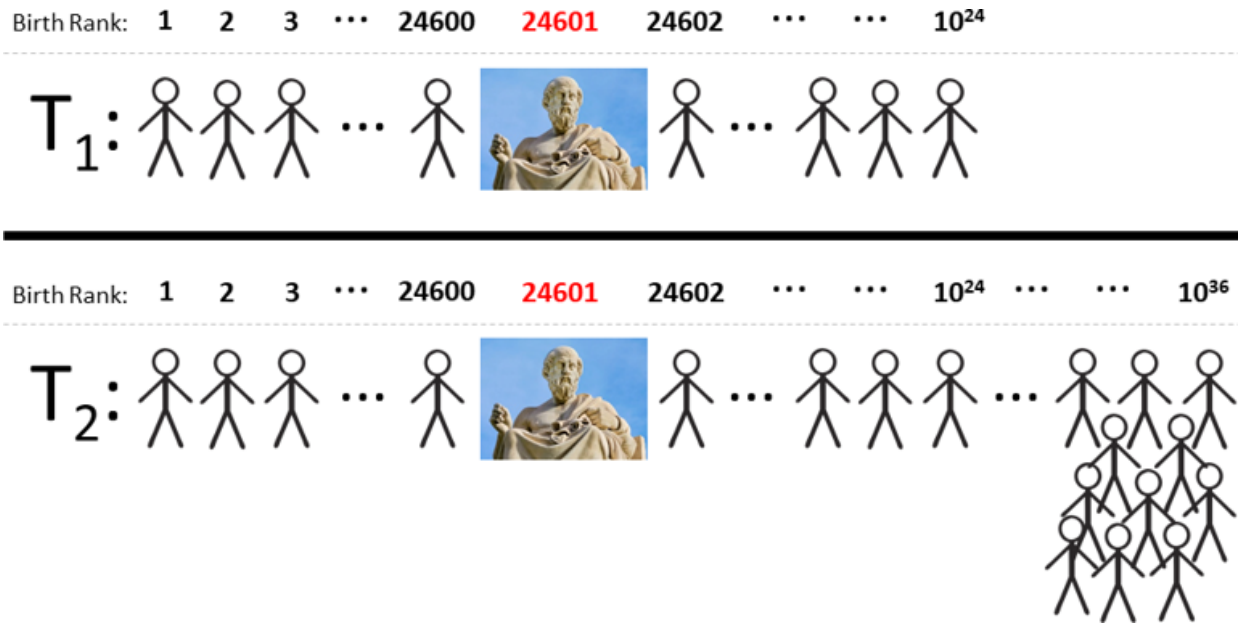
> It is the year 2100 and physicists have narrowed down the search for a theory of everything to only two remaining plausible candidate theories: T1 and T2 (using considerations from super-duper symmetry). According to T1 the world is very, very big but finite and there are a total of a trillion trillion observers in the cosmos. According to T2, the world is very, very, very big but finite and there are a trillion trillion *trillion* observers. The super-duper symmetry considerations are indifferent between these two theories. Physicists are preparing a simple experiment that will falsify one of the theories. Enter the presumptuous philosopher: “Hey guys, it is completely unnecessary for you to do the experiment, because I can already show you that T2 is about a trillion times more likely to be true than T1!”

The first thing to note is that the presumptuous philosopher (PP) may not even be right under SIA. We could ask:

- A. What proportion of the observers exactly like the PP are in the T_1 universes relative to the T_2 universes?

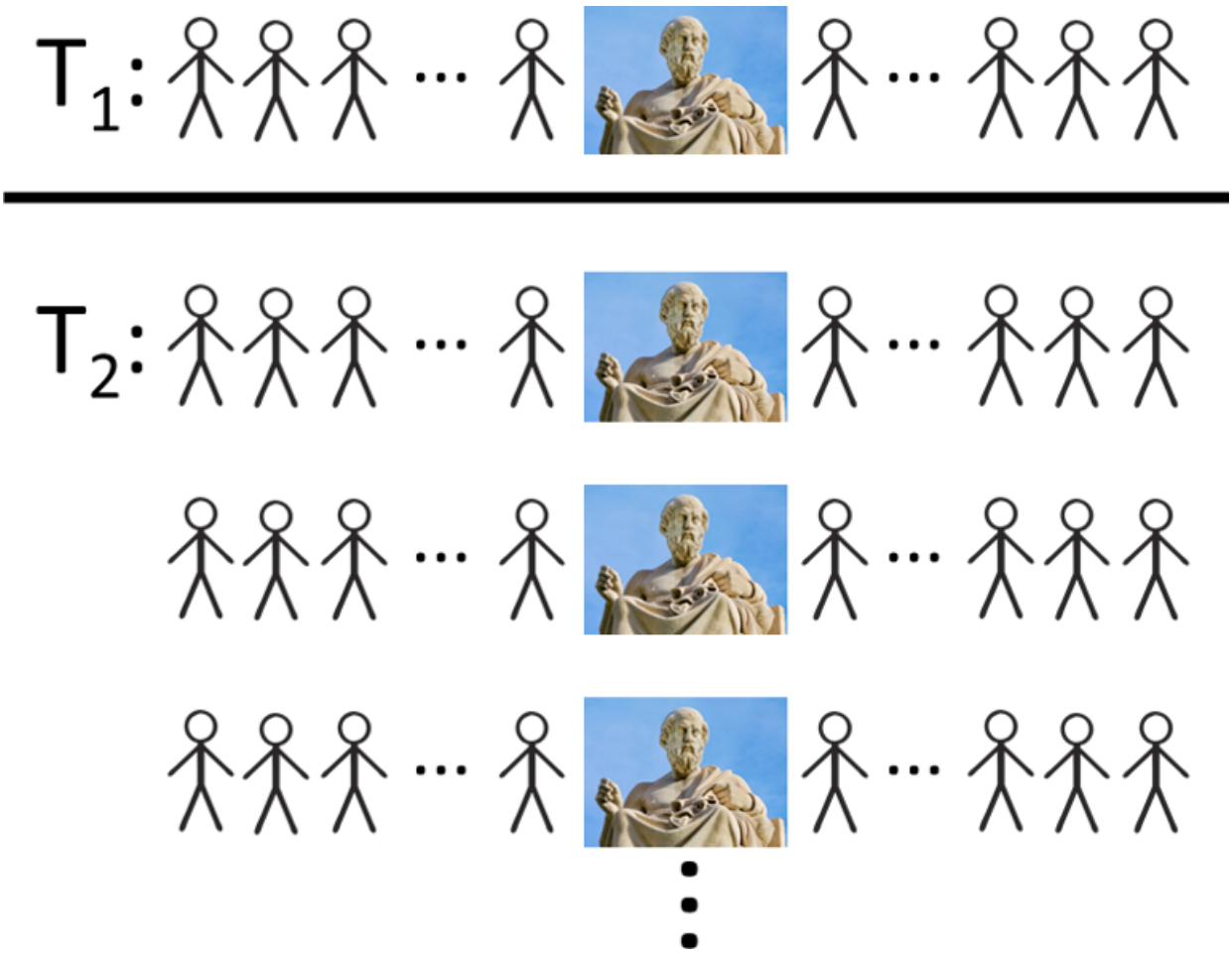
Recall that SIA is independent of reference class, so adding "exactly like the PP" doesn't change this. So, what is the answer to A.?

Now, T_2 universes have a trillion times more observers than the T_1 universes, but that doesn't necessarily mean that the PP are more likely in them. Suppose that everyone in these universes knows their rank of birth; for the PP it's the number 24601:



Then since all universes have more that 24601 inhabitants, the PP exists equally likely in T_1 universes as T_2 universes; the proportion is therefore 50% (interpreting "the super-duper symmetry considerations are indifferent between these two theories" as meaning "the two theories are equally likely").

Suppose however, the the PP does not know their rank, and the T_2 universes are akin to a trillion independent copies of the T_1 universes, each of which has an independent chance of generating an exact copy of PP:



Then SIA would indeed shift the odds by a factor of a trillion, giving a proportion of $1/(10^{12} + 1)$. But this is not so much a paradox, as the PP is correctly thinking "if all the exact copies of me in the multiverse of possibilities were to guess we were in T_2 universes, only one in a trillion of them would be wrong".

But if instead we were to ask:

-

2. What is the average proportion of PPs among other observers, in T_1 versus T_2 universes?

Then we would get the SSA answer. If the PPs know their birth rank, this is a proportion of 10^{12} : 1 *in favour of* T_1 universes. That's because there is just one PP in each universe, and a trillion times more people in the T_2 universes, which dilutes the proportion.

If the PP doesn't know their birth rank, then this proportion is the same[9] in the T_1 and T_2 universes. In probability terms, this would mean a "probability" of 50% for T_1 and T_2.

6.3 Anthropics and grabby aliens

The other paradoxes of anthropic reasoning can be treated similarly to the above. Now let's look at a more recent use of anthropics, due to Robin Hanson, Daniel Martin, Calvin McCarter, and Jonathan Paulson.

The basic scenario is one in which a certain number of alien species are "grabby": they will expand across the universe, at almost the speed of light, and prevent any other species of intelligent life from evolving independently within their expanding zone of influence[10].

Humanity has not noticed any grabby aliens in the cosmos; so we are not within their zone of influence. If they had started nearby and some time ago - say within the Milky Way and half a million years ago - then they would be here by now.

What if grabby aliens recently evolved a few billion light years away? Well, we wouldn't see them until a few billion years have passed. So we're fine. But if humans had instead evolved several billion years in the future, then we wouldn't be fine: the grabby aliens would have reached this location before then, and prevented us evolving, or at least would have affected us.

Robin Hanson sees this as an anthropic solution to a puzzle: why did humanity evolve early, i.e. only 13.8 billion years after the Big Bang? We didn't evolve as early as we possibly could - the Earth is a latecomer among Earth-like planets. But the smaller stars will last for trillions of years. Most habitable epochs in the history of the galaxy will be on planets around these small stars, way into the future.

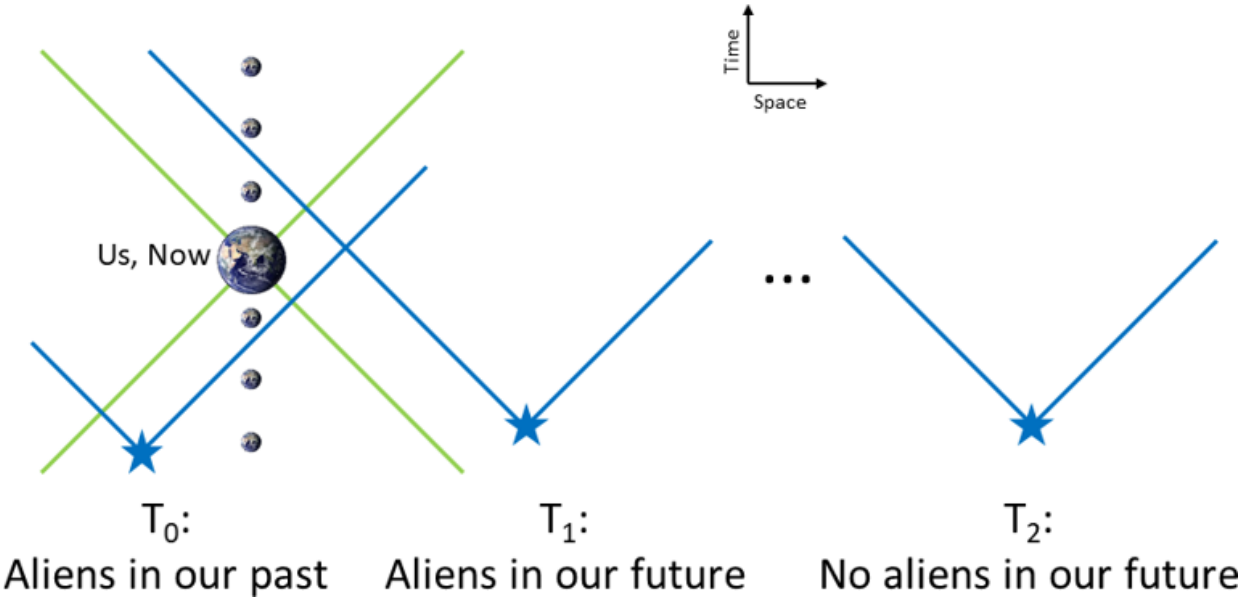
One possible solution to this puzzle is grabby aliens. If grabby aliens are likely (but not too likely), then we could only have evolved in this brief window before they reached us. I mentioned that SIA doesn't work for this (for the same reason that it doesn't care about the Doomsday argument). Robin Hanson then responded:

> If your theory of the universe says that what actually happened is way out in the tails of the distribution of what could happen, you should be especially eager to find alternate theories in which what happened is not so far into the tails. And more willing to believe those alternate theories because of that fact.

That is essentially Bayesian reasoning. If you have two theories, T_1 and T_2, and your observations are very unlikely given T_1 but more likely given T_2, then this gives extra weight to T_2.

Here we could have three theories:

0. T_0: "There are grabby aliens nearby"
1. T_1: "There are grabby aliens a moderate distance away"
2. T_2: "Any grabby aliens are very far away"



The T_0 can be ruled out by the fact that we exist. Theory T_1 posits that humans could not have evolved much later than we did (or else the grabby aliens would have stopped us). Theory T_2 allows for the possibility that humans evolved much later than we did. So, from T_2's perspective, it is "surprising" that we evolved so early; from T_1's perspective, it isn't, as this is the only possible window.

But by "theory of the universe", Robin Hanson meant not only the theory of how the physical universe was, but the anthropic probability theory. The main candidates are SIA and SSA. SIA is indifferent between T_1 and T_2. But SSA prefers T_1 (after updating on the time of our evolution). So we are more surprised under SIA than under SSA, which, in Bayesian/Robin reasoning, means that SSA is more likely to be correct.

But let's not talk about anthropic probability theories; let's instead see what questions are being answered. SIA is equivalent with asking the question:

1. What proportions of universes with human exactly like us, have moderately close grabby aliens (T_1) versus very distant grabby aliens (T_2)?

Or, perhaps more relevant to our future:

1. In what proportions of universes with human exactly like us, would those humans, upon expanding in the universe, encounter grabby aliens (T_1) or not encounter them (T_2)?

In contrast, the question SSA is asking is:

2. What is the average proportion of humans among all observers, in universes where there are nearby grabby aliens (T_1) versus very distant grabby aliens (T_2)?

If we were launching an interstellar exploration mission, and were asking ourselves what "the probability" of encountering grabby alien life was, then question 1. seems a closer phrasing of that than question 2. is.

And question 2. has the usual reference class problems. I said "observers", but I could have defined this narrowly as "human observers"; in which case it would have given a more SIA-like answer. Or I could have defined it expansively as "all observers, including those that might have been created by grabby aliens"; in that case SSA ceases to prioritise T_1 theories and may prioritise T_2 ones instead. In that case, humans are indeed "way out in the tails", given T_2: we are the very rare observers that have not seen or been created by grabby aliens.

In fact, the same reasoning that prefers SSA in the first place would have preferences over the reference class. The narrowest reference classes are the least surprising - given that we are humans in the 21st century with this history, how surprising is it that we are humans in the 21st century with this history? - so they would be "preferred" by this argument.

But the real response is that Robin is making a category error. If we substitute "question" for "theory", we can transform his point into:

> If your question about the universe gets a very surprising answer, you should be especially eager to ask alternate questions with less surprising answers. And more willing to believe those alternate questions.

---

1. We could ask some variants of questions 3. and 4., by maybe counting causally disconnected segments of universes as different universes (this doesn't change questions 1. and 2.). We'll ignore this possibility in this post. ↩
2. And also assuming that the radio's description of the situation is correct! ↩
3. Notice here that I've counted off observers with other observers that have exactly the same probability of existing. To be technical, the question which gives SIA probabilities should be "what proportion of potential observers, weighted by their probability of existing, have X?" ↩

4. More accurately: probability-weighted proportion. ↩

5. Let W be a set of worlds, p a probability distribution over W. Then the expectation of a is

   $E(a) = \sum_{W \in W} p(W)a_W = \sum_{W \in W} p(W)b_W/100 = (1/100)\sum_{W \in W} p(W)b_W = (1/100)E(b)$,

   which is 1/100 times the expectation of b. ↩

6. If we replace "observers" with "observer moments", then this question is equivalent with the probability generated by the *Strong Self-Sampling Assumption* (SSSA). ↩

7. If you forget some observations, your reference class can *increase*, as previously different copies become indistinguishable. ↩

8. Assuming the population is divisible by 10. ↩

9. As usual with SSA and this kind of question, this depends on how you define the reference class of "other observers", and who counts as a PP. ↩

10. This doesn't mean they will sterilise planets or kill other species; just that any being evolving within their control will be affected by them and know that they're around. Hence grabby aliens are, by definition, not hidden from view. ↩

SIA is basically just Bayesian updating on existence

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

I love the sleeping Beauty problem, but I think it leads people astray, by making them think that anthropic reasoning is about complicated situations with multiple duplicates. That leads to people saying erroneous things, like "SIA implies there are a lot of observers".

But anthropic reasoning isn't that complicated; SIA, especially, is mostly just Bayesian updating.

Specifically, if there are **no exact duplicates of yourself** in your universe, SIA is just Bayesian updating on the fact that you exist; this is the same update that an outside observer would make, if informed of your existence. So, if theory T_i has prior probability p_i and gives you a q_i probability of existing, then SIA updates T_i's probability to $q_i p_i$ (and then renormalises everything):

$$P_{SIA}(T_i) = P(T_i|\text{existence}) = \frac{P(T_i)\,P(\text{existence}|T_i)}{\sum_j P(T_j)\,P(\text{existence}|T_j)}.$$

This result is easy to see - since SIA is independent of reference class, just restrict the reference class to exact copies of you. If there is only on one such copy in the universe, then the update rule follows.

Even if there are multiple exact copies of you, you can still mostly see SIA as Bayesian updating over your *future* observations. See this footnote[1] for more details.

Indirect effect on population size

So, what does this mean for the number of observers in the universe? Well, SIA can have an indirect effect on population size. If, for instance, theory T_0 posits that life is likely to happen, then our existence is more likely, so T_0 gets a relative boost by SIA compared with most other theories.

So, SIA's boosting of other observers' existence is only an indirect effect of it boosting our existence. The more independent our existence is of them, or the more independent we suspect it might be, the less impact SIA has on them.

---

1. Suppose that there are N exact copies of you, and that they are going to make n independently random observations. Then as soon as n is much bigger than

$\log_2(N)$, you can expect that each copy will make a different observation; so, ultimately, you expect there to be only one exact future copy of you.

So if you Bayesianly update for each possible future copy (weighted by the probability of that future observation), you will get SIA. This is the trick that full non-indexical conditioning uses.

This can be seen as a partial solution to the Boltzmann brain problem: Boltzmann brains won't diverge, because they won't have future experiences. Personally, I prefer to address the issue by mixing in a bit of decision theory; my decisions are only relevant if I'm not a Boltzmann brain, so I'll start with "I exist and am not a Boltzmann brain" as an initial assumption. ↩

Non-poisonous cake: anthropic updates are normal

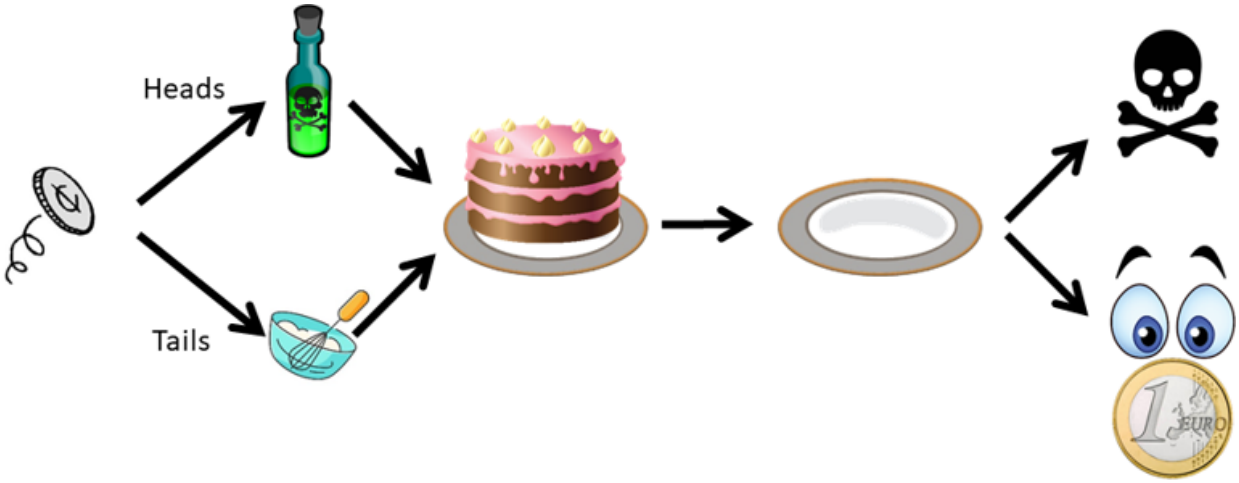
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

I am on a quest to show that anthropics probability are normal, at least in the absence of exact duplicates.

So consider this simple example: a coin is tossed. This coin is either fair, is 3/4 biased to heads, or 3/4 biased to tails; the three options are equally likely. After being tossed, the coin is covered, and you eat a cake. Then you uncover the coin, and see that it was tails.

You can now update your probabilities on what type of coin it was. It goes to a posterior of 1/6 on the coin being heads-biased, 1/3 on it being fair, and 1/2 on it being tails-biased[1]. Your estimated probability of it being tails on the next toss is (1/6)(1/4) + (1/3)(1/2) + (1/2)(3/4) = 7/12.

Now you are told that, had the coin come up heads, there would have been poison in the cake and you would have died before seeing the coin.



This fact makes the problem into an anthropic problem: you would never have been alive to see the coin, had it come up heads. But I can't see how that would have changed your probability update. If we got ethics board approval, we could actually run this experiment. And for the survivors in the tail worlds, we could toss the coin a second time (without cake or poison), just to see what it came up as. In the long run, we would indeed get roughly 7/12 tails frequency. So the update was correct, and the poison makes no difference.

Again, it seems that, if we ignore identical copies, anthropics is just normal probability theory. Now, if we knew about the poison, then we could deduce that the coin was tails from our survival. But that information gives us exactly the same update as seeing the coin was actually tails. So "I survived the cake" is exactly the same type of information as "the coin was tails".

Incubators

If we had more power in this hypothetical thought experiment, we could flip the coin and create you if it comes up tails. Then, after getting over your surprise, you could bet on the next flip of the coin - and the odds on that will be the same as in the poison cake and in the non-anthropic-case. Thus updates are the same if:

1. Standard coin toss, you see tails.
2. Poison cake situation, you survive the cake.
3. You're created on tails flip and notice you exist.

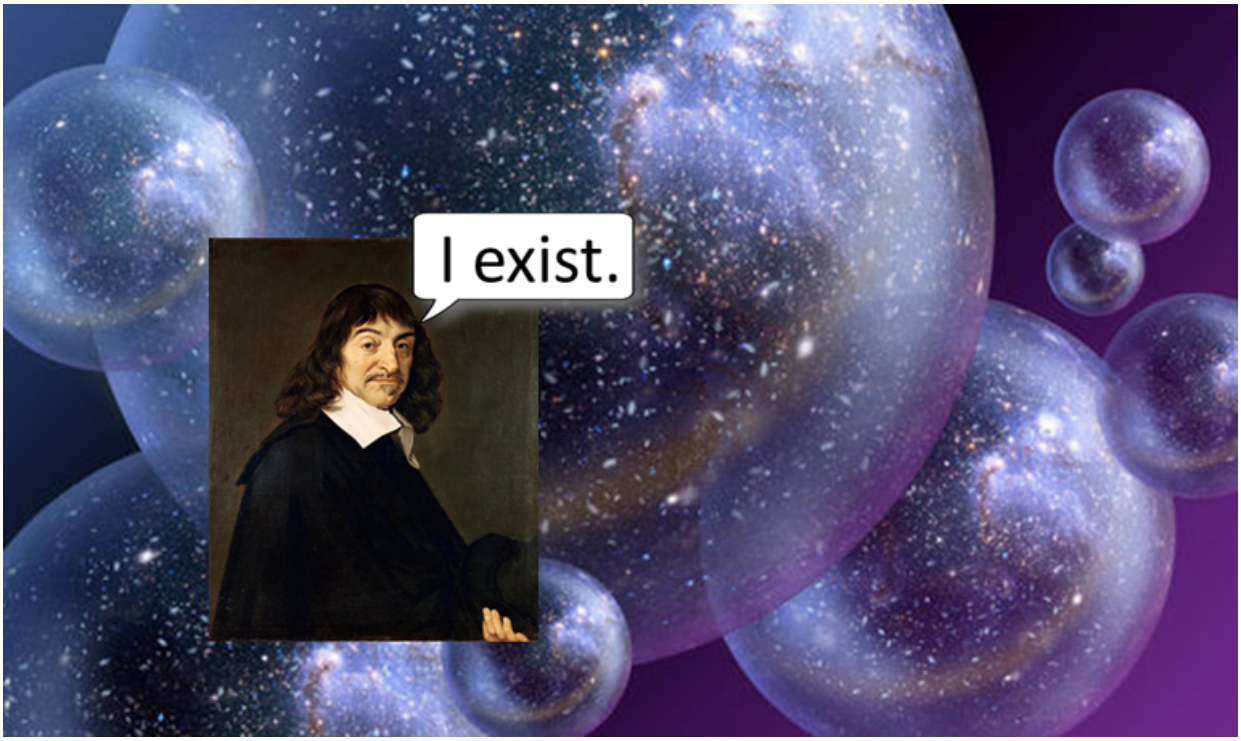
---

1. The probability of tails given the heads-biased coin is 1/4; given the fair coin it is 1/2 = 2/4, and given tails-biased it is 3/4. So the odds are 1 : 2 : 3; multiplying these by the (equal) prior probabilities doesn't change these odds. To get probabilities, divide the odds by 6, the sum of the odds, and get 1/6, 2/6 = 1/3 and 3/6 = 12. ↩

Anthropics in infinite universes

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

When talking about anthropics, people often say things like "assume the universe is finite; weird things happen in infinite universes". I've myself argued that SSA breaks down when we encounter infinities; SIA breaks down sooner, when we encounter expected infinities.



You can formalise this informally[1] with the thought that:

1. In an infinite universe, anything can happen, no matter how unlikely: life must exist somewhere. So our existence doesn't tell us anything about life; its probability could be anything at all.

A superficially convincing argument; but not one you'd use for anything else. For instance, consider the following:

2. In an infinite universe, anything can happen, no matter how unlikely: if gravity didn't exist, somewhere it must seem to exist by shear chance. So our observation of gravity doesn't tell us anything about gravity; its probability could be anything at all.

I've argued before that anthropic questions are pretty normal. Why would we accept the reasoning in question 1, but reject it in question 2?

We shouldn't. We can deal with questions like 2 by talking about limits of probabilities in larger and larger spaces, or by discounting distant observations (similar to sections

2.3 and 3.1 in infinite ethica). So we might define conditional probabilities like $P(X \mid Y)$ in an infinite universe in the following way:

- Let $P_{rl}(X \mid Y)$ be the ratio of observers, within a large hypersphere of radius r centered on location l, that observe X and Y, relative to the proportion that observes Y. If this tends to a limit as $r \to \infty$, independently of l, then define that limit to be $P(X \mid Y)$.

Note that this definition works just as well for Y = "we observe the force of gravity to be blah" as with Y = "we exist".

Now, that definition might not be ideal (in particular, "radius" is not defined for relativistic space-time). No problem: different definitions of probability are asking different questions, and can lead to different anthropic probabilities, just as in the finite case.

I'll call these class of questions "SIA-limit questions", since they are phrased as ratios of observers, and dependent on how we use limits to define probability in infinite universes. They each lead to various "SIA-limit anthropic probability theories"; in most standard situations, these should reach the same answers as each other.

---

1. Yes, it's perfectly possible to formalise informally, and I encourage people to do it more often. ↩

The SIA population update can be surprisingly small

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

With many thanks to Damon Binder, and the spirited conversations that lead to this post, and to Anders Sandberg.

People often think that the self-indication assumption (SIA) implies a huge number of alien species, millions of times more than otherwise. Thought experiments like the presumptuous philosopher seem to suggest this.

But here I'll show that, in many cases, updating on SIA doesn't change the expected number of alien species much. It all depends on the prior, and there are many reasonable priors for which the SIA update does nothing more than double the probability of life in the universe[1].

This can be the case even if the prior says that life is very unlikely! We can have a situation where we are astounded, flabbergasted, and disbelieving about our own existence - "how could we exist, how can this beeeeee?!?!?!?" - and still not update much - "well, life is still pretty unlikely elsewhere, I suppose".

In the one situation where we have an empirical distribution, the "Dissolving the Fermi Paradox" paper, the effect of the SIA anthropics update is to multiply the expected civilization per planet by **seven**. Not seven orders of magnitude - just seven.

The formula

Let $\rho \in [0,1]$ be the probability of advanced space-faring life evolving on a given planet; for the moment, ignore issues of life expanding to other planets from their one point of origin. Let f be the prior distribution of ρ, with mean μ and variance σ^2. This means that, if we visit another planet, our probability of finding life is μ.

On this planet, we exist[2]. Then if we update on our existence we get a new distribution f'; this distribution will have mean μ':

$$\mu' = \mu\left(1 + \frac{\sigma^2}{\mu^2}\right).$$

To see a proof of this result, look at this footnote[3].

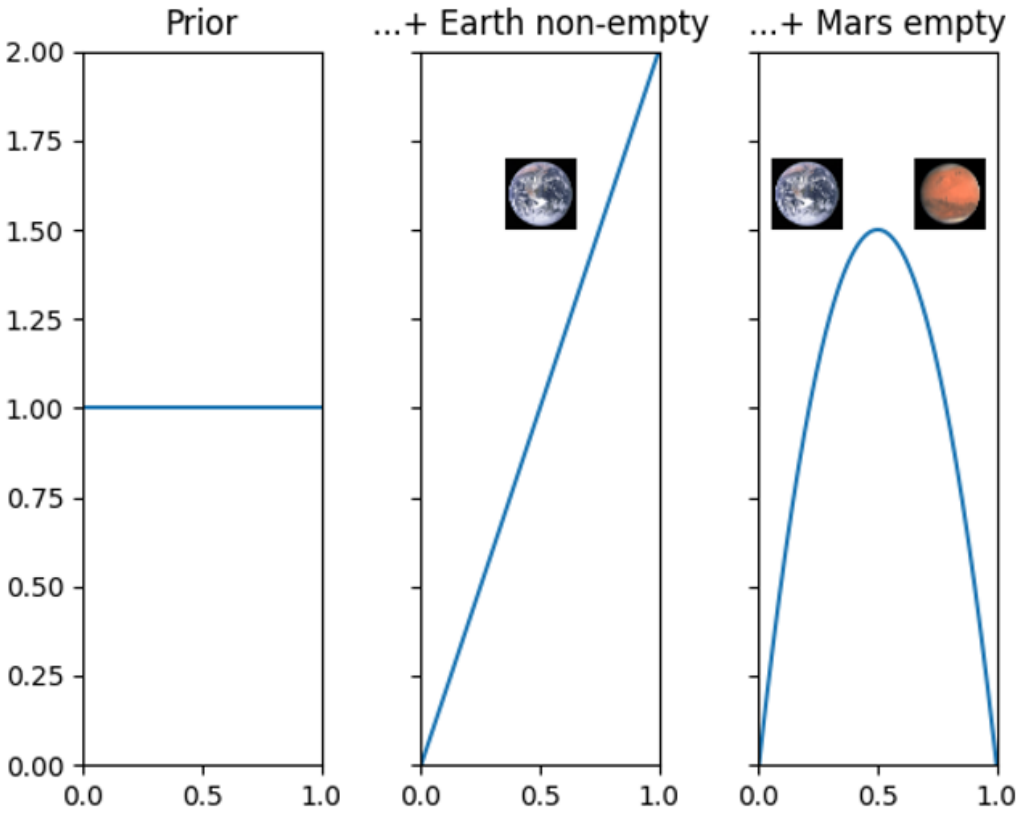
Define $M_{\mu,\sigma^2} = 1 + \sigma^2/\mu^2$ to be this multiplicative factor between μ and μ'; we'll show that there are many reasonable situations where M_{μ,σ^2} is surprisingly low: think 2 to 100, rather than in the millions or billions.

Beta distributions I

Let's start with the most uninformative prior of all: a uniform prior over $[0, 1]$. The expectation of ρ is $\int_0^1 \rho d\rho = 1/2$, so, without any other information, we expect a planet to have life with 50% probability. The variance is $\sigma^2 = 1/12$.

Thus if we update on our existence on Earth, we get the posterior $f'(\rho) = 2\rho$; the mean of this is 2/3 (either direct calculation or using $M_{1/2,1/12} = 1 + 4/12 = 4/3$).

Even though this change in expectation is multiplicatively small, it does seem that the uniform prior and the $f'(\rho)$ are very different, with $f'(\rho)$ heavily skewed to the right. But now consider what happens if we look at Mars and notice that it hasn't got life. The probability of no life, given ρ, is $1 - \rho$. Updating on this and renormalising gives a posterior $6\rho(1 - \rho)$:



The expectation of $6\rho(1-\rho)$, symmetric around 1/2, is of course 1/2. Thus one extra observation (that Mars is dead) has undone, in expectation, all the anthropic impact of our own existence.

This is an example of a beta distribution for $\alpha = 2$ and $\beta = 2$ (yes, beta distributions have a parameter called β and another one that's α; just deal with it). Indeed, the uniform prior is also a beta distribution (with $\alpha = \beta = 1$) as is the anthropic updated version 2ρ (which has $\alpha = 2, \beta = 1$).

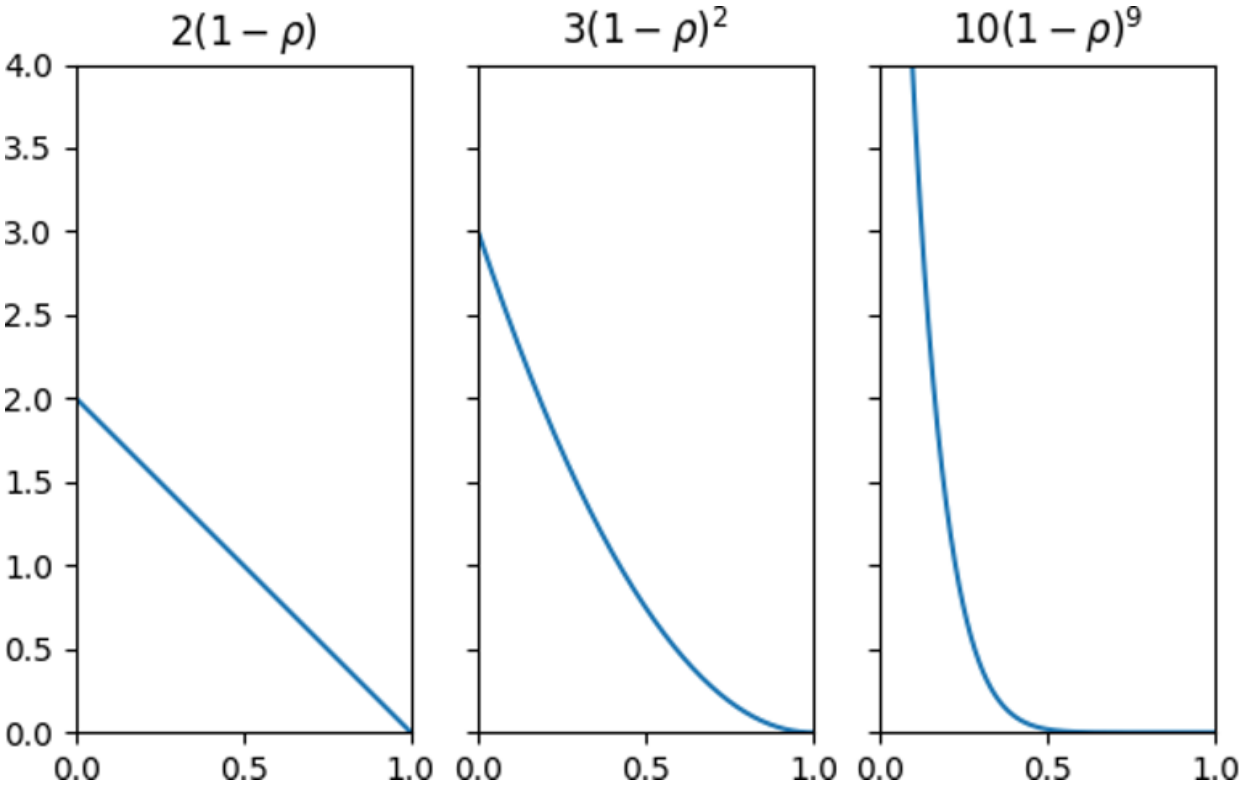
The update rule for beta distributions is that a positive observation (ie life) increases α by 1, and a negative observation (a dead planet) increases β by 1. The mean of an updated beta distribution is a generalised version of Laplace's law of succession: if our prior is a beta distribution with parameters α and β, and we've had m positive observations and n negative ones, then the mean of the posterior is:

$$\frac{\alpha + m}{\alpha + \beta + m + n}.$$

Suppose now that we have observed n dead planets, but no life, and that we haven't done an anthropic update yet, then we have a probability of life of $\alpha/(\alpha + \beta + n)$. Upon adding the anthropic update, this shifts to $(\alpha + 1)/(\alpha + \beta + n + 1)$, meaning that the multiplicative factor is at most $(\alpha + 1)/\alpha$. If we started with the uniform prior with its $\alpha = 1$, this multiplies the probability of life by at most 2. In a later section, we'll look at $\alpha < 1$.

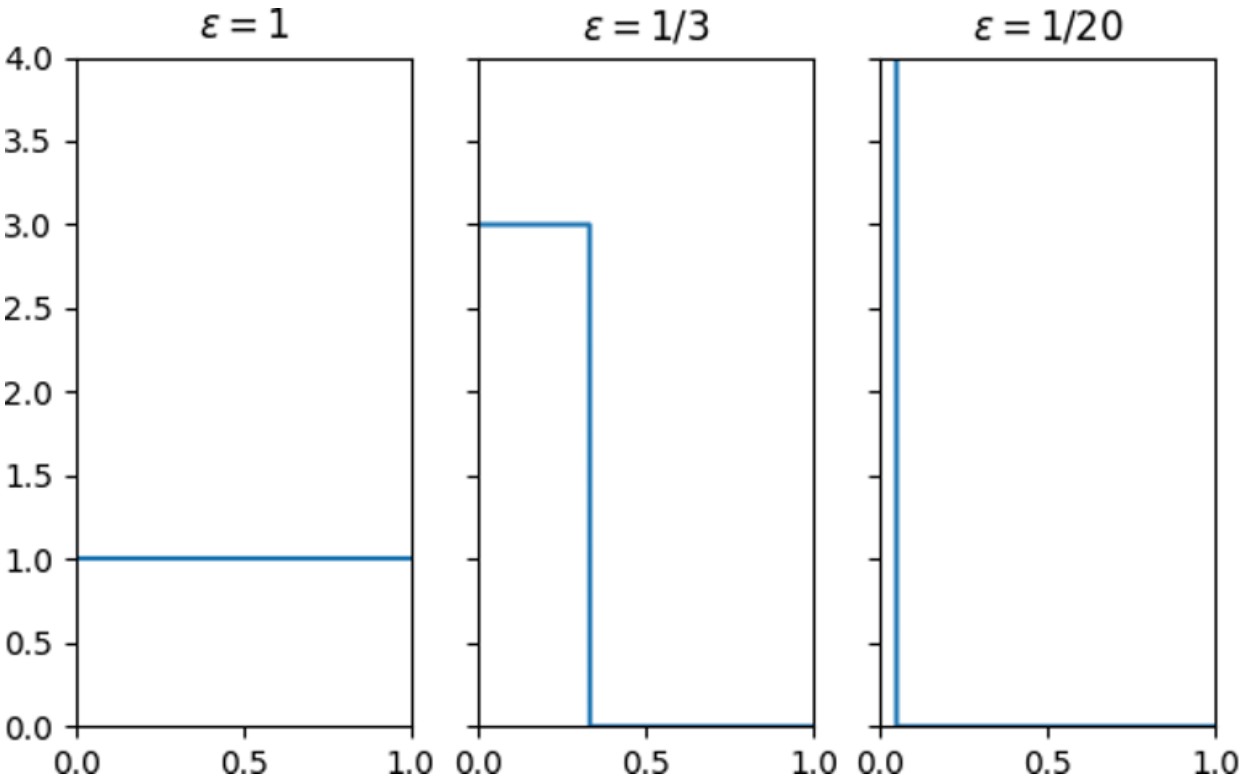
High prior probability is not required for weak anthropic update

The uniform prior has $\alpha = \beta = 1$ and starts at expectation 1/2. But we can set $\alpha = 1$ and a much higher β, which skews the distribution to the left; for example, for $\beta = 2$, 3, and 10:



Even though these priors are skewed to the left, and have lower prior probabilities of life (1/3, 1/4, and 1/11), the anthropic update has a factor M_{μ,σ^2} that is less than 2.

Also note that if we scale the prior f by a small ϵ, so replace f(ρ) on the range $[0, 1]$ with f(ρ/ϵ)/ϵ on the range $[0, \epsilon]$, then μ is multiplied by ϵ and σ^2 is multiplied by ϵ^2. Thus $M_{\mu,\epsilon}$ is unchanged. Here, for example, is the uniform distribution, scaled down by $\epsilon = 1$, $\epsilon = 1/3$, and $\epsilon = 1/20$:



All of these will have the same M_{μ,σ^2} (which is 4/3, just as for the uniform distribution). And, of course, doing the same scaling with the various beta distributions we've seen up until now will also keep M_{μ,σ^2} constant.

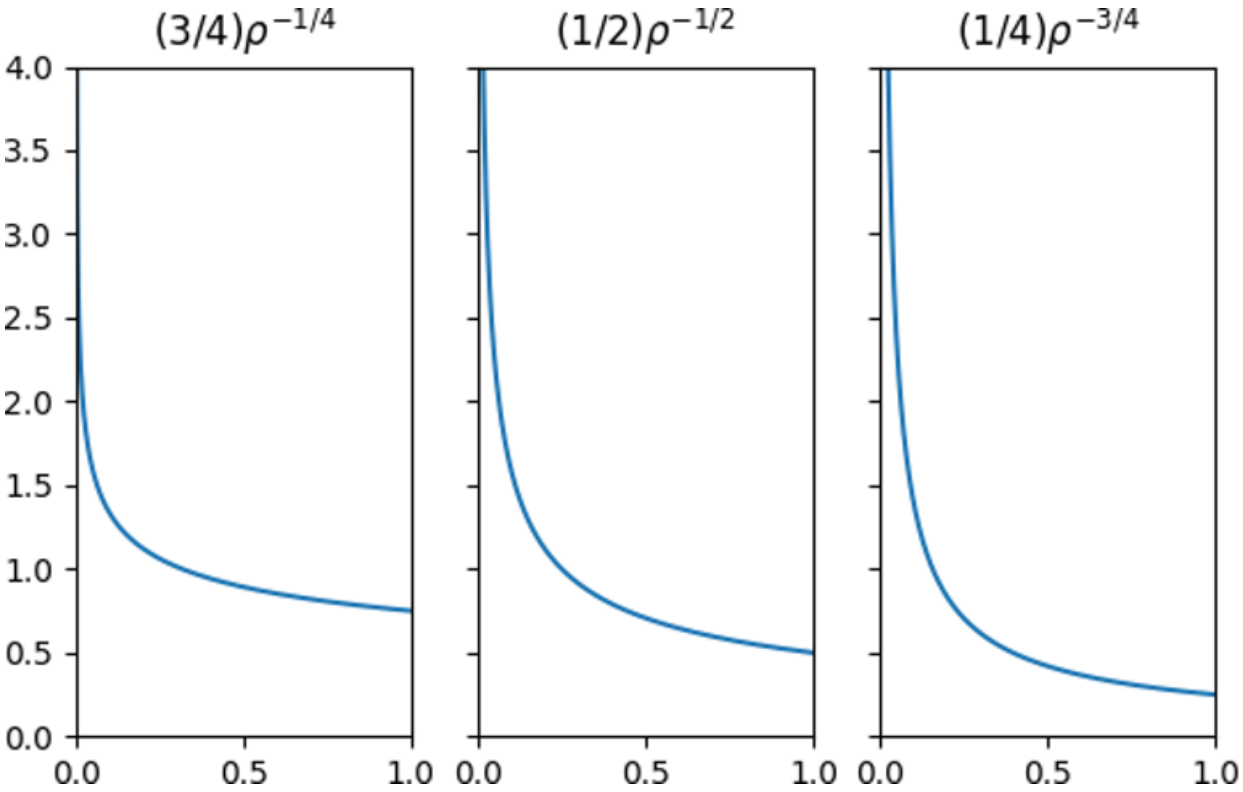
Thus there are a lot of distributions with very low μ (ie very low prior probability of life) but an M_{μ,σ^2} that's less than 2 (ie the anthropic update is less than a doubling of the probability of life).

Beta distributions II and log-normals

The best-case scenario for M_{μ,σ^2} is if f assigns probability 1 to $\rho = \mu$. In that case, $\sigma^2 = 0$ and $M = 1$: the anthropic update changes nothing.

Conversely, the worse-case scenario for M_{μ,σ^2} is if f only allows $\rho = 0$ and $\rho = 1$. In that case, f assigns probability μ to 1 and $1 - \mu$ to 0, for a mean of μ and a variance of $\sigma^2 = \mu - \mu^2$, and a multiplicative factor of $M_{\mu,\sigma^2} = 1/\mu$. In this case, after anthropic update, f' assigns certainty to $\rho = 1$ (since any life at all, given this f, means life on all planets).

But there are also more reasonable priors with large M_{μ,σ^2}. We've already seen some, implicitly, above: the beta distributions with $\alpha < 1$. In that case, M_{μ,σ^2} is bounded by $(\alpha + 1)/\alpha$. If $\alpha = 3/4$ and $\beta = 1$, for instance, this corresponds to the (unbounded) distribution $f(\rho) = (3/4)\rho^{-1/4}$; the multiplicative factor is below 7/3, which is slightly above 2. But as α declines, the multiplicative factor can go up surprisingly fast; at $\alpha = 1/2$ it is 3, at $\alpha = 1/4$ it is 5:



In general, for $\alpha = 1/n$, the multiplicative factor is bounded by $n + 1$. This gets arbitrarily large as $\alpha \to 0$. Though $\alpha = 0$ itself corresponds to the improper prior $f(\rho) = 1/\rho$, whose integral diverges. On a log scale, this corresponds to the log-uniform distribution, which is roughly what you get if you assume "we need N steps, each of probability p, to get life; let's put a uniform prior over the possible Ns".

It's not clear why one might want to choose $\alpha = 1/10^{20}$ for a prior, but there is a class of prior that is much more natural: the log-normal distributions. These are random variables X such that log(X) is normally distributed.

If we choose log(X) to have a mean that is highly negative (and a variance that isn't too large), then we can mostly ignore the fact that X takes values above 1, and treat it as a prior distribution for ρ. The mean and variance of the log-normal distributions can be explicitly defined, thus giving the multiplications factor as:

$$M_{\mu,\sigma^2} = \exp \bar{\sigma}^2.$$

Here, $\bar{\sigma}^2$ is the variance of the normal distribution log(X). This $\bar{\sigma}^2$ might be large, as it denotes (roughly) "we need N steps, each of probability p, to get life; let's put a uniform-ish prior over a range of possible Ns". Unlike $1/\rho$, this is a proper prior, and a plausible one; therefore there are plausible priors with very large M_{μ,σ^2}. The log normal is quite likely to appear, as it is the approximate limit of multiplying together a host of different independent parameters.

Multiplication law

Do you know what's more likely to be useful than "the approximate limit of multiplying together a host of different independent parameters"? Actually multiplying together independent parameters.

The famous Drake equation is:

$$R_* \cdot f_p \cdot n_e \cdot f_l \cdot f_i \cdot f_c \cdot L.$$

Here R^* is the number of stars in our galaxy, f_p the fraction of those with planets, n_e the number of planets that can support life per star that has planets, f_l the fraction of

those that develop life, f_l the fraction of those that develop intelligent life, f_c the fraction of those that release detectable signs of their existence, and L is the length of time those civilizations endure as detectable.

Then the proportion of advanced civilizations per planet is qf_lf_i, where q is the proportion of life-supporting planets among all planets. To compute the M of this distribution, we have the highly useful result (the proof is in this footnote[4]):

- Let X_i be independent random variables with multiplicative factors M_i, and let M be the multiplicative factor of $X = X_1 \cdot X_2 \cdot \ldots \cdot X_n$. Then $M = \prod_i M_i$ - the total M is the product of the individual M_i.

The paper "dissolving the Fermi paradox" gives estimated distributions for all the terms in the Drake equation. The q, which doesn't appear in that paper, is a constant, so has $M_q = 1$. The f_i has a log-uniform distribution from 0.001 to 1; the M can be computed from the mean and variance of such distributions, so $M_{f_i} = \log(1/0.001)\frac{1-0.001^2}{2(1-0.001)^2} \approx 3.5$.

The f_l term is more complicated; it is distributed like $g(X) = 1 - e^{-e^{X \cdot 50 \log(10)}}$ where X is a standard normal distribution. Fortunately, we can estimate its mean and variance without having to figure out its distribution, by numerical integration of g(x) and $g(x^2)$ on the normal distribution. This gives $\mu \approx 0.5$, $\sigma^2 \approx 0.25$ and $M \approx 2$. The overall the multiplicative effect of anthropic update is:

$$M_{planet} \approx 7.$$

What if we considered the proportion of advanced civilization per star, rather than per planet? Then we can drop the q term and add in f_p and n_e. Those are both estimated to be distributed as log-uniform on $[0.1, 1]$; for a total M of

$$M_{star} \approx 14.$$

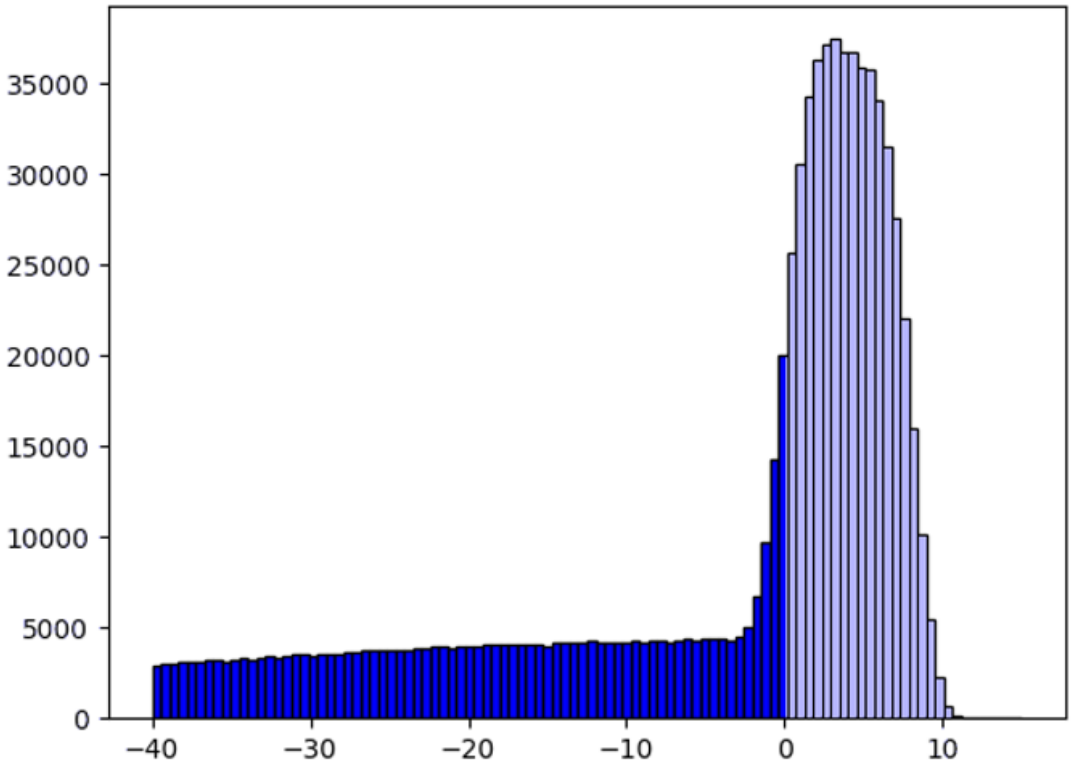
Why is the M higher for civilizations per star than civilizations per planet? That's because when we update on our existence, we increase the proportion of civilizations per planet, but we also update the proportion of planets per star - both of these can make life more likely. The M_{star} incorporates both effects, so is strictly higher than M_{planet}.

We can do the same by considering the number of civilizations per galaxy; then we have to incorporate R_* as well. This is log-uniform on $[1, 100]$, giving:

$$M_{galaxy} \approx 32.$$

What about if we include the Fermi observation (the fact that we don't see anything in our galaxy)? The "dissolving the Fermi paradox" paper shows there are multiple different ways of including this update, depending on how we parse out "not seeing anything" and how easy it is for civilizations to expand.

I did a crude estimate here by taking the Fermi observation to mean "the proportion of civilizations per galaxy must be less than one". Then I did a Monte-Carlo simulation, ignoring all results above 0 on the log scale:



From this, I got an estimated mean of 0.027, variance of 0.014, and a total multiplier of:

$$M_{galaxy,\ Fermi} \approx 21.$$

With the Fermi observation and the anthropic update combined, we expect 0.56 civilizations per galaxy.

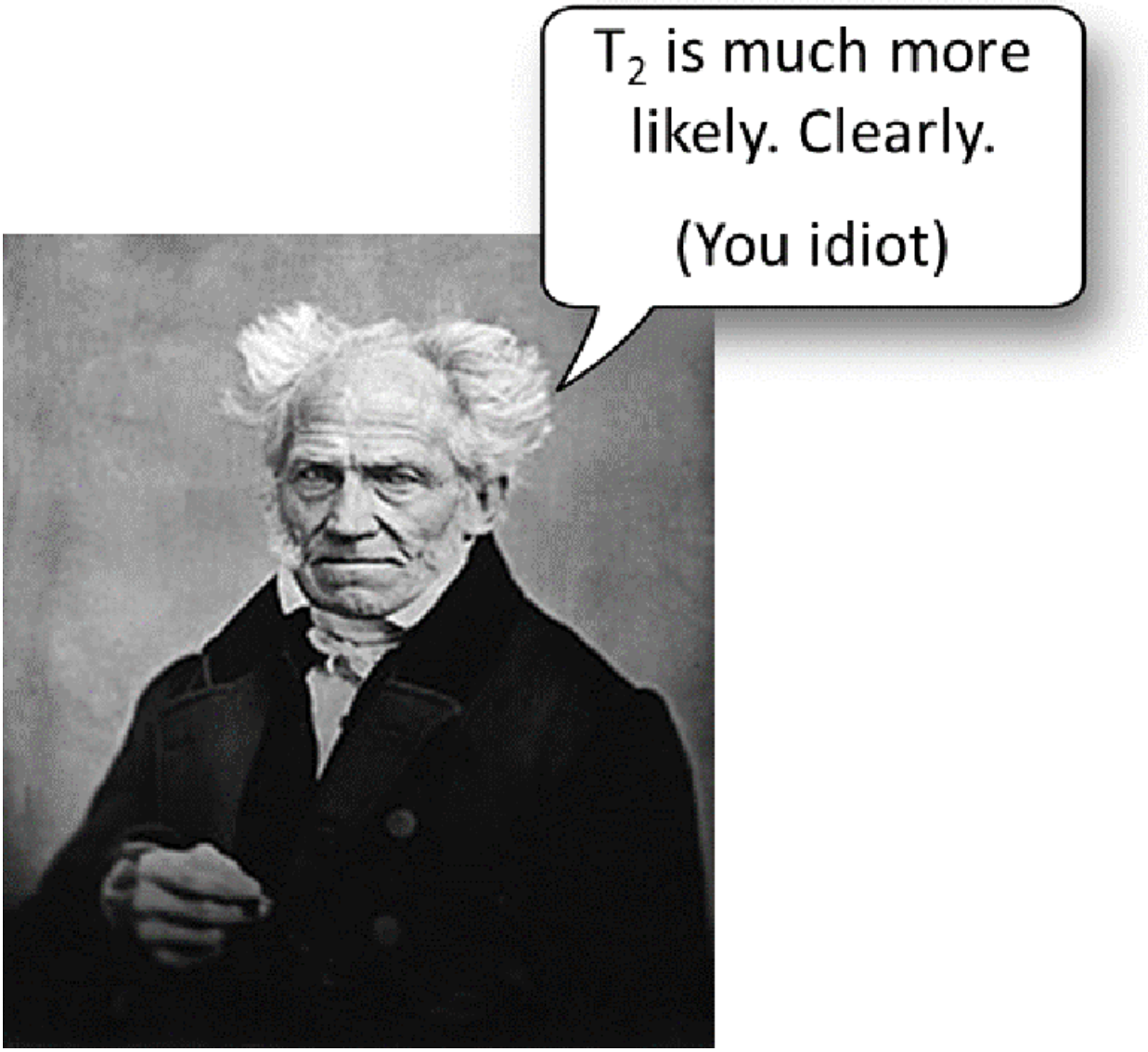
Limitations of the multiplier

Low multiplier, strong effects

It's important to note that the anthropic update can be very strong, without changing the expected population much. So a low M_{μ,σ^2} doesn't necessary mean a low impact.

Consider for instance the presumptuous philosopher, slightly modified to use planetary population densities. Thus theory T_1 predicts $\rho = 1/10^{12}$ (one in a trillion) and T_2 predicts $\rho = 1$; we put initial probabilities 1/2 on both theories.

As Nick Bostrom noted, the SIA update pushes T_2 to being a trillion times more probable than T_1; a postiori, T_2 is roughly a certainty (the actual probability is $10^{12}/(10^{12} + 1)$).



However, the expected population goes from roughly 1/2 (the average of $1/10^{12}$ and 1) to roughly 1 (since *a postiori* T_2 is almost certain). This gives a M_{μ,σ^2} of roughly 2. So, despite the strong update towards T_2, the actual population update is small - and, conversely, despite the actual population update being small, we have a strong update towards T_2.

Combining multiple theories

In the previous post, note that that both T_1 and T_2 were point estimates: they posit a constant ρ. So they have a variance of zero, and hence a M_{μ,σ^2} of 1. But T_2 has a much

stronger anthropic update. Thus we can't use their M_{μ,σ^2} to compare the anthropic effects on different theories.

We also can't relate the individual Ms to that of a combined theory. As we've seen, T_1 and T_2 have Ms of 1, but the combined theory $(1/2)T_1 + (1/2)T_2$ has an M of roughly 2. But we can play around with the relative initial weight of T_1 and T_2 to get other Ms.

If we started with odds $10^{12} : 1$ on T_1 vs T_2, then this has a mean ρ of roughly 10^{-12}; the anthropic update sends it to 1 : 1 odds, with a mean of roughly 1/2. So this combined theory has an M of roughly $10^{12}/2$, half a trillion.

But, conversely, if we started with odds $1 : 10^{12}$ on T_1 vs T_2, then we have an initial mean of ρ of roughly one; its anthropic update is odds of $1 : 10^{24}$, also with a mean of roughly one. So this combined theory has an M of roughly 1.

There *is* a weak relation between M and the M_i of the various T_i. Let M_i be the multiplier of T_i has a multiplier of M_i; we can reorder the T_i so that $M_i \leq M_j$ for $i \leq j$. Let T be a combined theory that assigns probability p_i to T_i.

1. For all $\{p_i\}$, $M \geq \min_i(M_i)$.
2. For all ϵ, there exists $\{p_i\}$ with all $p_i > 0$, so that $M < \min_i(M_1) + \epsilon$.

So, the minimum value of the M_i is a lower bound on M, and we can get arbitrarily close to that bound. See the proof in this footnote[5].

---

1. As we'll see, the *population* update is small even in the presumptuous philosopher experiment itself. ↩
2. Citation partially needed: I'm ignoring Boltzmann brains and simulations and similar ideas. ↩
3. Given a fixed ρ, the probability of observing life on our own planet is exactly ρ. So Bayes's theorem implies that $f'(\rho) \propto \rho f(\rho)$. With the full normalisation, this is

$$f'(\rho) = \frac{\rho f(\rho)}{\int_0^1 \rho f(\rho) d\rho}.$$

If we want to get the mean μ' of this distribution, we further multiply by ρ and integrate:

$$\mu' = E_{f'}(\rho) = \int_0^1 \frac{\rho^2 f(\rho)}{\int_0^1 \rho f(\rho) d\rho} d\rho = \frac{\int_0^1 \rho^2 f(\rho) d\rho}{\int_0^1 \rho f(\rho) d\rho}$$

Let's multiply this by $1 = 1/1 = (\int_0^1 f(\rho)d\rho) / (\int_0^1 f(\rho)d\rho)$ and regroup the terms:

$$\mu' = \frac{\int_0^1 \rho^2 f(\rho) d\rho}{\int_0^1 f(\rho) d\rho} \cdot \frac{\int_0^1 f(\rho) d\rho}{\int_0^1 \rho f(\rho) d\rho}.$$

Thus $\mu' = E_f(\rho^2)/E_f(\rho) = (\sigma^2 + \mu^2)/\mu = \mu(1 + \sigma^2/\mu^2)$, using the fact that the variance is the expectation of ρ^2 minus the square of the expectation of ρ. ↩

4. I adapted the proof in this post.

   So, let X_i be independent random variables with means μ_i and variances σ_i^2. Let $X = \prod_i X_i$, which has mean μ and variance σ^2. Due to the independence of the X_i, the expectations of their products are the product of their expectations. Note that X_i^2 and X_j^2 are also independent if $i \neq j$. Then we have:

$$\begin{aligned}
\prod_i M_{\mu_i,\sigma_i^2} &= \prod_i \left(1 + \frac{\sigma_i^2}{\mu_i^2}\right) \\
&= \prod_i \left(\frac{\mu_i^2 + \sigma_i^2}{\mu_i^2}\right) \\
&= \prod_i \left(\frac{E(X_i^2)}{\mu_i^2}\right) \\
&= \frac{\prod_i E(X_i^2)}{\prod_i (E(X_i))^2} \\
&= \frac{E(X^2)}{E(X)^2} \\
&= \frac{\mu^2 + \sigma^2}{\mu^2} \\
&= 1 + \frac{\sigma^2}{\mu^2} = M_{\mu,\sigma^2}.
\end{aligned}$$

↩

5. Let $\{f_i\}_{1 \le i \le n}$ be probability distributions on ρ, with mean μ_i, variance σ_i^2, expectation squared $s_i = E_{f_i}(\rho^2) = \sigma_i^2 + \mu_i^2$, and $M_i = s_i/\mu_i^2$. Without loss of generality, reorder the f_i so that $M_i \le M_j$ for $i < j$.

   Let f be the probability distribution $f = p_1 f_1 + \ldots p_n f_n$, with associated multiplier M. Without loss of generality, assume $M_i \le M_j$ for $i < j$. Then we'll show that $M \ge M_1$.

   We'll first show this in the special case where $n = 2$ and $M_1 = M_2$, then generalise to the general case, as is appropriate for a generalisation. If $s_1/\mu_1^2 = M_1 = M_2 = s_2/\mu_2^2$, then, since all terms are non-negative, there exists an α such that $s_1 = \alpha^2 s_2$ while $\mu_1 = \alpha\mu_2$. Then for any given $p = p_1$, the M of f is:

$$M(p) = \frac{ps_1 + (1-p)s_2}{(p\mu_1 + (1-p)\mu_2)^2} = \frac{ps_1 + (1-p)\alpha^2 s_1}{(p\mu_1 + (1-p)\alpha\mu_1)^2} = M_1 \frac{1(p) + \alpha^2(1-p)}{(1(p) + \alpha(1-p))^2}$$

   The function $x \to x^2$ is convex, so, interpolating between the values $x = 1$ and $x = \alpha$, we know that for all $0 \le p \le 1$, the term $(1(p) + \alpha(1-p))^2$ must be lower than $1^2(p) + \alpha^2(1-p)$. Therefore $(1(p) + \alpha^2(1-p))/(1(p) + \alpha(1-p))^2$ is at most 1, and $M(p) \le M_1$. This shows the result for $n = 2$ if $M_1 = M_2$.

   Now assume that $M_2 > M_1$, so that $s_1/\mu_1^2 < s_2/\mu_2^2$. Then replace s_2 with s_2', which is lower than s_2, so that $s_1/\mu_1^2 = s_2'/\mu_2^2$. If we define $M'(p)$ as the expression for $M(p)$ with $s_2' substituting for s_2, we know that $M'(p) \le M(p)$, since $s_2' < s_2$. Then the previous result shows that $M'(p) \ge M_1$, thus $M(p) \ge M_1$ too.

   To show the result for larger n, we'll induct on n. For $n = 1$ the result is a tautology, $M_1 \le M_1$, and we've shown the result for $n = 2$. Assume the result is true for $n - 1$, and then notice that $f = p_1 f_1 + \ldots p_n f_n$ can be re-written as $f = p_1 f_1 + (1 - p_1) f'$, where $f' = (p_2' f_2 + \ldots p_n' f_n)$ for $p_i' = p_n/(1 - p_1)$. Then, by the

induction hypothesis, if M' is the M of f', then $M' \geq M_2$. Then applying the result for $n = 2$ between f_1 and f', gives $M \leq \min(M_1, M')$. However, since $M_1 \leq M_2$ and $M' \geq M_2$, we know that $\min(M_1, M') = M_1$, proving the general result.

To show M can get arbitrarily close to M_1, simply note that M is continuous in the $\{p_i\}$, define $p_1 = 1 - \epsilon$, $p_i = \epsilon/(n-1)$ for $i > 1$, and let ϵ tend to 0. ↩

Anthropics and Fermi: grabby, visible, zoo-keeping, and early aliens

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

When updating the probability of life across the universe, there are two main observations we have to build on:

- Anthropic update: we exist on Earth.
- Fermi observation: (we exist on Earth and) we don't see any aliens.

I'll analyse how these two observations affect various theories about life in the universe. In general, we'll see that the anthropic update has a pretty weak effect, while the Fermi observation has a strong effect: those theories that benefit most are those that avoid the downgrade from the Fermi, such as the Zoo hypothesis, or the "human life unusually early" hypothesis.

Grabby and visible aliens

I've argued that an anthropic update on our own existence is actually just a simple Bayesian update; here I'll explain what that means for our updates.

This paper talks about grabby aliens, who would expand across the universe, and stop humans from evolving (if they reached Earth before now). As I've argued, "we exist" and "we have not observed X" are statements that can be treated in exactly the same way. We can combine them to say "there are no visible aliens anywhere near", without distinguishing grabby aliens (who would have stopped our existence) from visible-but-not grabby aliens (who would have changed our observations).

Thus the Fermi observation is saying there are no grabby or visible aliens nearby[1]. Recall that it's so comparatively easy to cross between stars and galaxies, so advanced aliens would only fail to be grabby if they coordinated to not want to do so.

Rare Earth hypotheses

Some theories posit that life requires a collection of conditions that are very rarely found together.

But rare Earth theories don't differ much, upon updates, from "life is hard" hypotheses.

For example, suppose T_0 say that life can exist on any planet with rate ρ, while the rare Earth hypothesis, T_1, says that life can exist on Earth-like planets with rate p, while Earth-like planets themselves exist with rate r.

But neither the Fermi observation nor the anthropic update will distinguish between these theories. The Fermi observation posits that there are no visible aliens close to Earth; the anthropic updates increases the probability of life similar to us. We can see T_1 as T_0 with a different prior on $\rho = pr$ (induced from the priors on p and r), but both these updates affect ρ.

Now, T_0 and T_1 can be distinguished by observation (seeing dead planets with Earth-like features) or theory (figuring out what is needed for life). Anything that differentially change p and r. But neither the anthropic update nor the Fermi observation do this.

Independent aliens

Suppose that theory T_2 posits that there are aliens in, say, gas giants, whose existence is independent from ours. Visible gas giant alien civilizations exist at a rate ρ_g, while visible life on rocky planets exist at a rate ρ.

Then the anthropic update boosts ρ only, while the Fermi observation penalises ρ and ρ_g equally (if we make the simplifying assumption that gas giants are as common as terrestrial planets).

This gives a differential boost to ρ over ρ_g, but the effect can be mild. If we assume that there are N gas giants and N terrestrial planets in the milky way, and start with a uniform prior over both ρ and ρ_g, then after updating, we get:

$$\rho = \frac{2}{N+2}, \quad \rho_g = \frac{1}{N+1}.$$

If the rate of gas giant alien civilizations is semi-dependent on our own existence - maybe we both need it to be easy for RNA to exist - then there will be less of a difference in the update for ρ and ρ_g.

So, some differential effect due to anthropics, but not a strong one, at least for uniform priors, and not one that grows.

In the cosmic zoo

Let T_3 be a cosmic zoo hypothesis. It posits that there may be a lot of aliens, but they have agreed - or been coerced - into hiding themselves, so as not to contaminate human development (or some other reason).

Then T_3 gets a boost from the anthropic update, and no penalty from the Fermi observation. Since most theories get a big downgrade from the Fermi observation, this can raise its probability quite a lot relative to other theories.

A few caveats, however:

1. In the zoo hypothesis, aliens are hiding themselves from us. This is close to a "Descartes's demon" hypothesis, in that powerful entities are acting to feed us erroneous observations. Pure Descartes's demon hypotheses are not differentially boosted by anything, since they explain nothing (once you've posited a demon, you also have to explain why we see what we think we see). The zoo hypothesis is not quite as bad - "keep everything hidden" is more likely than other ways aliens could be messing with our observations. Still, it should be a low prior.
2. Though the Fermi observation doesn't downgrade the zoo hypothesis directly, the more carefully we observe the universe, the more unlikely it becomes, since the aliens would have to work harder to conceal any evidence.
3. Conversely, the more visible we become, the less likely the zoo hypothesis becomes, because we have to explain why the zookeepers haven't intervened to keep us concealed (if we suppose that these aliens are powerful enough to intercept light and other signals between the stars, then we're very close to the Descartes demon territory). Once we successfully launched replicating AIs to the stars, then we'd be pretty sure the zoo hypothesis was wrong.

Time enough for aliens

So far, we've neglected time in the equation, talking about a rate ρ that was per planet, but not stretched over time. But consider theory T_4: advanced life starts appearing around 13.77 billion years after the Big Bang, but not before.

This theory might be unlikely, but it gets a mild boost from anthropics (since it's compatible with our existence) and avoids the downgrade from the Fermi observation (since it says there are no visible aliens - **yet**).

Since that downgrade has been quite powerful for most theories, T_4 can get boosted relative to them - and the more dead planets we observe or infer, the stronger the relative boost is.

Now, T_4 may seem unlikely, since the Earth is a late planet among the Earth-like planets: "Thus, the average earth in the Universe is 1.8 ± 0.9 billion years older than our Earth". But there are some theories that make more plausible T_4, such as some versions of panspermia. Specifically, if we imagine that life had to go through several stages, on several planets - maybe RNA/DNA was the result of billions of years of evolution on a planet much older than the Earth, and was then spread here, where it allowed another stage of evolution.

Conversely, theories T_5 that posit that advanced life started much earlier than the present day, pay a much higher price via the Fermi observation.

---

1. The grabby alien paper uses "loud" to designate aliens that "expand fast, last long, and make visible changes to their volumes". Visible aliens are more general; in particular, they need not expand (though this may make them less visible). ↩

Practical anthropics summary

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is a simple summary of how to "do anthropics":

- If there are no issues of exact copies, or advanced decision theory, and the questions you're asking aren't weird, then use SIA. And by "use SIA", I mean "ignore the definition of SIA, and just do a conventional Bayesian update on your own existence". Infinite universes won't be a problem (any more than they are with conventional probabilities). And this might not increase your expected population by much.

First of all, there was the realisation that different theories of anthropic probability correspond to correct answers to different questions - questions that are equivalent in non-anthropic situations.

We can also directly answer "what actions should we do?", without talking about probability. This anthropic decision theory gave behaviours that seem to correspond to SIA (for total utilitarianism) or SSA (for average utilitarianism).

My personal judgement, however, is that the SIA-questions are more natural than the SSA-questions (ratios of totals rather than average of ratios), including the decision theory situation (total utilitarianism rather than average utilitarianism). Thus, in typical situations, using SIA is generally the way to go.

And if we ignore exact duplicates, Boltzmann brains, and simulation arguments, SIA is simply standard Bayesian updating on our existence. Anthropic probabilities can be computed exactly the same way as non-anthropic probabilities can.

And there are fewer problems than you might suspect. This doesn't lead to problems with infinite universes - at least, no more than standard probability theories do. And anthropic updates tend to increase the probability of larger populations in the universe, but that effect can be surprisingly small - 7 to 32 given the data we have.

Finally, note that anthropic effects are generally much weaker than Fermi observation effects. The fact that we don't see life, on so many planets, tells us a lot more than the fact we see life on this one.