Dispelling the Anthropic Shadow

Teruji Thomas (Global Priorities Institute, University of Oxford)

Global Priorities Institute | September 2024

GPI Working Paper No . 20-2024

Dispelling the Anthropic Shadow

Teruji Thomas[*]

Abstract

There are some possible events that we could not possibly discover in our past. We could not discover an omnicidal catastrophe, an event so destructive that it permanently wiped out life on Earth. Had such a catastrophe occurred, we wouldn't be here to find out. This space of unobservable histories has been called *the anthropic shadow*. Several authors claim that the anthropic shadow leads to an 'observation selection bias', analogous to survivorship bias, when we use the historical record to estimate catastrophic risks. I argue against this claim.

1 Introduction

Estimating the probability of catastrophic events is a difficult business. We don't have much to go on when the catastrophes would be of a novel kind, arising from hypothetical social or technological developments. On the other hand, for some types of catastrophes we have a long geological record to consult, along with other forms of data and the results of scientific modelling. For example, when it comes to asteroid impacts, we can use geological dating techniques and the observed distribution of crater sizes to inform our estimates of future risks.

A curious thing about this historical data, however, is that there are some possible data points that we could not possibly observe. We could not find in the historical data an *omnicidal* catastrophe, an event so destructive that it permanently wiped out life on Earth. Had such a catastrophe happened, we wouldn't be here to find out. Ćirković,

[*] Global Priorities Institute, Faculty of Philosophy, University of Oxford. I especially thank Andreas Mogensen and Toby Ord for helpful comments.

Sandberg, and Bostrom (2010) call this space of unobservable histories *the anthropic shadow*.

Some striking claims have been made about the significance of the anthropic shadow. For example, Ćirković et al. claim that, because of the anthropic shadow, a straightforward treatment of the historical record will lead to systematically underestimating the chances of potentially omnicidal events. In the extreme case, they write,

> we should have no confidence in historically based probability estimates for events that would certainly extinguish humanity...(1497)

Tegmark and Bostrom (2005a) elaborate this line of thought in an earlier 'brief communication' published in *Nature*:

> Given that life on Earth has survived for nearly 4 billion years (4 Gyr), it might be assumed that natural catastrophic events are extremely rare. Unfortunately, this argument is flawed because it fails to take into account an observation-selection effect..., whereby observers are precluded from noting anything other than that their own species has survived up to the point when the observation is made. If it takes at least 4.6 Gyr for intelligent observers to arise, then the mere observation that Earth has survived for this duration cannot even give us grounds for rejecting with 99% confidence the hypothesis that the average cosmic neighbourhood is typically sterilized, say, every 1,000 years. The observation-selection effect guarantees that we would find ourselves in a lucky situation, no matter how frequent the sterilization events.

To avoid any straightforward appeal to the historical absence of omnicidal events, Tegmark and Bostrom develop a more complicated method based on modelling the formation times of habitable planets. More recently, Snyder-Beattie, Ord, and Bonsall (2019) estimate the lifespan of the human species and check how robust this estimate is with respect to different evolutionary hypotheses. They reason:

> [I]f human existence required a 10 million year (Myr) period of evolution free from asteroid impacts, any human observers will necessarily find in their evolutionary history

> a period of 10 Myr that is free of asteroid impacts, regardless of the true impact rate. Inferring a rate based on those 10 Myr could therefore be misleading, and methods must to be used to correct for this bias.

In this paper I argue for a deflationary position: the existence of the anthropic shadow is essentially irrelevant to estimating risks. There are several interesting points that come up along the way, but an initial reason for skepticism is easy to state. According to standard forms of Bayesianism or evidentialism, what we ought to think depends on what evidence we actually have. The fact that we could not easily have had different evidence is not important in itself. So, even if we could not easily have had evidence of past omnicidal events, our actual evidence that there were no such events should make us think that the rate of them is low.

The core of this paper, sections 2–4, analyses a stylized example close to the one in Ćirković et al. Section 5 extends my analysis applies to the models used by Tegmark and Bostrom and by Snyder-Beattie et al., while section 6 concludes. As I mentioned, Ćirković et al. focus on *potentially* omnicidal events. For present concreteness, let us say that a 'potentially omnicidal event' is one that has a 10% chance of permanently ending life on Earth. Then, the upshot of my analysis is essentially as follows.

(A) The fact that life has survived so long is evidence that the rate of potentially omnicidal events is low.

(B) Given the fact that life has survived so long, historical frequencies provide evidence for a true rate rather higher than the observed rate.

(C) These two effects cancel out, so that, overall, the historical record provides evidence for a true rate close to the observed rate.

Thus, contrary to claims about the anthropic shadow, the historical record is (in the stated sense) a reliable guide to the rate of potentially omnicidal events.

On my reading, the authors quoted above are too focused on (B). Based on (B), the suggestion is that using the historical rate as an estimate of the true rate may lead to a bad underestimate. However, I argue that (A) is true and undermines this suggestion. Effectively, focusing on (B) neglects the base-rate provided by (A). I must admit, however, that I find the exact position of Ćirković et al. somewhat difficult to decipher. So, while I will try to indicate the specific points at which I disagree with other participants in this literature, the primary aim of this paper is to lay out the true story as clearly as I can, in a way that will forestall any further confusion.

2 Supervolcano

My analysis will be based on a toy model, a small elaboration of the one considered by Ćirković et al. While this toy model is unrealistic in various ways, it should allow us to see whether the anthropic shadow has any general significance. Here is the set-up:

> **Supervolcano.** There is a sequence of Earthlings (one in each generation until extinction). Before each generation, including the first, there is chance P of a supervolcanic eruption, and, for each eruption, a chance Q that life survives, and therefore a chance $1 - Q$ that the eruption is fatal, ending life on Earth.
>
> Let Jack be the nth Earthling, for some large number n. He knows $Q = q$ but wants to estimate P. Checking the geological record, he is able to observe F, the frequency of eruptions so far. He finds $F = f$, with no fatal eruptions.
>
> Should Jack become confident that P is close to f?

My answer: Yes, at least if n is sufficiently big. Indeed, Jack's situation is in all important respects the same as the situation of someone who flips a coin many times and thereby becomes confident that the chance of heads is close to the observed frequency.[1]

In contrast, Ćirković et al. seem to think that the frequency will be a significant underestimate of P. I will reconstruct their position in section 4. For now, some comments about the case $Q = 0$ may be instructive. This is the case where any eruption is fatal. So Jack, knowing

[1] More formally, the claim is that, for any $\varepsilon > 0$ and any $p < 1$, if n is large enough, Jack should end up assigning probability at least p to the hypothesis that P is within ε of f. My general convention is to use capital letters for random variables and small letters for their candidate values.

that he is the nth human, can be sure even without checking the geological record that there were no eruptions at all: $F = 0$. On my view, he should therefore be confident that P is close to 0. In apparent contrast, this case leads Ćirković et al. to the view I quoted in the introduction, that 'we should have no confidence in historically based probability estimates' for omnicidal events. Of course, 0 is the lowest possible value for P, so, even on my view, this observed frequency will probably be an underestimate. But this has nothing to do with the anthropic shadow. Consider witnessing a coin land tails 1,000 times in a row. One should then be confident that the chance of heads is close to 0, even though, for the same reason, this will probably be an underestimate. There is, I claim, nothing different going on in the $Q = 0$ version of Supervolcano.

Two Misleading Analogies

To get an initial sense of where Ćirković et al. are coming from, and why I disagree, it might be helpful to contrast Supervolcano with some superficially similar examples. Here is one:

> **Fishing.** Sarah goes fishing in a pond, using a net with six-inch holes. The fact that she only catches fish longer than six inches doesn't provide much or any evidence about the prevalence of smaller fish in the pond.[2]

There is an apparent analogy between Fishing and $Q = 0$ Supervolcano. Just as Sarah could not have found a small fish in her net, Jack could not have found a fatal eruption in Earth's past. So, one might think, just as Sarah has little or no evidence about the prevalence of small fish, Jack has little or no evidence about the prevalence of fatal eruptions. This analogy seems to be the basic reason for thinking that that anthropic shadow is important. However, this analogy may lead us astray.

Consider Fishing more carefully. An initial point is that, for the example to work, Sarah's evidence must include the fact that her net has big holes. If the net has big holes but she herself is agnostic about their size, then finding only big fish *will* give her evidence that there are no small fish. So the case does not work merely on the basis that Sarah could not have caught any small fish; it requires this fact to be part of her evidence. The only question we have to ask is what her total evidence supports.

[2] This is a version of the example that opens Bostrom (2002).

To answer this question, we note that Sarah has two key pieces of evidence: the fact that the net has big holes, and the fact that she only catches big fish. Given the first fact (and Sarah's other background evidence), the second fact provides no further evidence about the prevalence of small fish. And, importantly, we naturally imagine that the first fact itself has little or no bearing on the prevalence of small fish. Overall, then, Sarah has little or no evidence about the prevalence of small fish.

So far, so good. Note, however, that we could have imagined the details of the case in a different way. Suppose that the net was carefully designed so that it could catch all of the fish in the pond, and that Sarah knows this. Then the first fact—the fact that the net has big holes—*is* evidence that there are no small fish in the pond. On this alternative way of filling in the details, Sarah *does* end up with strong evidence that there are no small fish. Note that, in both versions of the case, Sarah's evidence could not have been different, in the sense that Sarah could not have caught any small fish because her net has big holes. So the difference in the verdicts illustrates why, as I said in the introduction, 'The fact that [Sarah] could not easily have had different evidence is not important in itself'.

Turning back to Supervolcano, Jack, like Sarah, has two key pieces of evidence: the fact that he is the nth Earthling, and the fact that $F_n = 0$, meaning that there were no fatal eruptions in the first n generations. Again, given the first fact, the second provides no further evidence about the value of P. But we still have to ask whether the first fact tells us anything about P.[3] My basic contention is that it *does* tell us something important: it is evidence that P is close to zero. In this respect, Supervolcano with $Q = 0$ is analogous to the second version of Fishing, not the first. This conclusion falls out of my analysis of Supervolcano in the next section. For now I just emphasize that the analogy with Fishing may well be misleading.

Similar points can be made with respect to

> **Survivorship Bias.** David seeks out a cohort of ten centenarians who follow a Mediterranean diet. Everyone in this cohort has lived 100 years without suffering a fatal

[3] I made a similar point in the introduction, where I said that we have to attend to (A) as well as (B). The difference is that here I am focusing on certainly, rather than potentially, omnicidal events, corresponding to $Q = 0$.

> heart-attack. But this fact provides David with little or no evidence that a Mediterranean diet is good for one's heart.[4]

Like Jack and Sarah, David has two key pieces of evidence: the fact that the people in his cohort are centenarians, and the fact that they have all gone 100 years without fatal heart-attacks. Given the first fact, the second fact provides no further evidence about the healthiness of their diet. But we must still ask whether the first fact tells us anything of use. It doesn't, the way the scenario is described: David sought out centenarians, and presumably would have found some even if the Mediterranean diet were quite unhealthy. But we need not think that Supervolcano is analogous to Survivorship Bias in this respect. We can imagine an alternative scenario, in which David chose the members of his cohort at random when they were still young, and they all happen to live to 100. Then the fact that everyone in the cohort is a centenarian *would* be good evidence about the merits of the Mediterranean diet. Neither Survivorship Bias nor this variation on it is *exactly* like Jack's situation, and until we look more closely we cannot tell which scenario provides a better guide.

3 Analysis

I am claiming that Jack's situation in Supervolcano is not in any important respect different from the situation of someone observing a sequence of coin-flips and using the observed frequency of heads to estimate the chance. One could, with a little care, give essentially the same analysis in both cases. However, there are some superficial differences between these cases that I think are worth examining closely. My strategy is to consider two intermediate cases that isolate these differences, and to argue that each difference is irrelevant.

Here is the first intermediate case:

> **The Martians.** On Mars there is a parallel sequence of generations, but no extinction risk. Jill is the nth Martian, Jack's contemporary. Jill learns all about the situation on Earth. She wants to estimate the chance P of an eruption

[4] In general, survivorship bias is the bias that arises from treating people who survived an event as if they were typical of the larger group. Ord (2020, 84), for one, says that the anthropic shadow leads to a form of survivorship bias.

> before each generation. She learns that eruptions have occured with frequency f, but there have been no fatal eruptions.

The notable difference between Jill's situation in The Martians and Jack's in Supervolcano is that Jill's existence, unlike Jack's, does not depend on the past absence of fatal eruptions. As a result, histories with fatal eruptions are in the anthropic shadow from Jack's point of view, but not from Jill's, so one might think that any effect of the anthropic shadow would show up in a difference between Jack's credences and Jill's.

On the other hand, an interesting feature of Supervolcano is that the total number of people who ever exist—the 'number of observers'—depends on the absence of fatal eruptions. At some points Ćirković et al. write as if this is the crucial feature of the case.[5] And it is not something that makes a difference between Jack's situation and Jill's. So I will compare The Martians with a more everyday case involving coin-flips.

> **Barking Dog.** Jacob is bored at home and decides to measure the bias of his favourite coin. That is, he wants to estimate P, the chance that it lands heads. He flips it n times and determines the frequency F of heads, finding $F = f$. Jacob's dog, Gemma, is normally calm, but gets excited by portraiture: each time she sees heads, there is a known chance $1-q$ that she barks. As it happens, Gemma does not bark at all during the sequence of flips.

Barking Dog is analogous to The Martians, with an eruption before the kth generation being analogous to the coin landing heads on the kth flip, and a fatal eruption being analogous to Gemma's barking. But (it is natural to imagine) Gemma's barking makes no difference to the number of observers.

Barking Dog is a minor variation on

> **Ordinary Coin.** Jacob is bored at home and decides to measure the bias of his favourite coin. That is, he wants to find out P, the chance that it lands heads. He flips it n times and determines the frequency F of heads, finding $F = f$. There are no other relevant considerations.

[5] See their flow-chart on p. 1501. Readers familiar with the 'doomsday argument' will also anticipate some connection here.

The difference between Ordinary Coin and Barking Dog is not philosophically interesting, but perhaps it is not instantly obvious that in *both* cases, Jacob should end up confident that P is close to f.

Ordinary Coin vs Barking Dog

In this section I explain the fairly simple mathematical point that both Ordinary Coin and Barking Dog lead to the same result: Jacob should end up confident that the chance matches the observed frequency (as usual: assuming the number of flips is sufficiently large). I begin by recalling how this works in Ordinary Coin.

Let's suppose Jacob has some prior probability measure Pr over various hypotheses. Specifically, Pr represents Jacob's credences before he observes the sequence of coin flips; it can take into account the other sorts of background evidence that one would typically have prior to witnessing a sequence of flips. The exact form of Pr will not matter, as long as it is suitably 'agnostic', not ruling out any relevant hypothesis about P or F.

As a general Bayesian norm, on gaining new evidence E, Jacob should update the probability he assigns to any other hypothesis A from $\Pr(A)$ to $\Pr(A \mid E)$. According to Bayes's Theorem,

$$\Pr(A \mid E) = \frac{\Pr(E \mid A)\Pr(A)}{\Pr(E)}.$$

The key point is that the probability he assigns to A gets multiplied by

$$\frac{\Pr(A \mid E)}{\Pr(A)} = \Pr(E \mid A) \times (\text{a factor that doesn't depend on } A).$$

In other words, learning E shifts probability towards hypotheses that make E likely, in proportion to *how* likely they make E. If $\Pr(E \mid A)$ is higher than $\Pr(E \mid B)$, we can say that E supports A rather than B. And, whatever agnostic prior Jacob starts with, if $\Pr(E \mid A)$ is sufficiently many times larger than $\Pr(E \mid B)$, Jacob should end up thinking that A is far more likely than B.[6]

Now, in our particular case, the evidence that Jacob gains is that $F = f$. This evidence supports certain hypotheses about the value of

[6] When considered as a function of A, $\mathscr{L}(A) := \Pr(E \mid A)$ is often called *the likelihood function*. I will use this terminology to some extent when we get to section 5, but, to be clear, I use the word 'likely' in the usual way as a synonym for 'probable'.

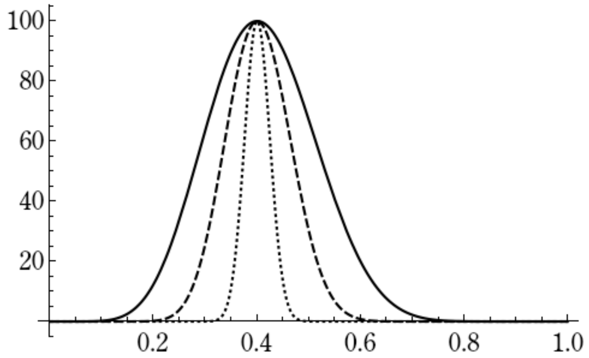

Figure 1: A coin with unknown bias P is flipped n times, and the frequency $F = 0.4$ of heads is observed. The 'likelihood' $\mathscr{L}(p) := \Pr(F = 0.4 \mid P = p)$ is given by the binomial distribution, peaking at $p = 0.4$. Plotted here is $\mathscr{L}(p)$ as a percentage of the peak value $\mathscr{L}(0.4)$, for $n = 20$ (solid line), $n = 60$ (dashed), and $n = 400$ (dotted). As n increases the peak becomes narrower, meaning that the evidence more strongly confirms that P is close to 0.4.

P. To find out which hypotheses it supports most strongly, we should consider the conditional probability $\Pr(F = f \mid P = p)$. That is, what values of P would make the frequency f particularly likely?

Since Jacob doesn't initially have any special information about the flips to begin with, the probability he assigns to each flip landing heads, given that $P = p$, should be p. It follows from some elementary combinatorics that the probability that $F = f$ is given by the *binomial distribution*:[7]

$$\Pr(F = f \mid P = p) = \frac{n!}{(fn)!(n - fn)!} p^{fn}(1 - p)^{n - fn}.$$

Holding f fixed, think of this as a function of p. It equals zero for $p = 0, 1$ (unless $f = 0, 1$) and peaks at $p = f$. The peak gets narrower as n increases. The effect, then, is that learning $F = f$ shifts probability towards $P = p$, and this effect gets stronger as n gets bigger. See Figure 1.

That is the sense in which learning $F = f$ provides Jacob with evidence that $P \approx f$, and the sense in which the evidence gets stronger as n gets larger. Whatever agnostic prior Jacob started with, if n is large enough, Jacob should become confident that $P \approx f$.

[7] Note that fn is just the observed number of heads. The probability of any particular sequence of flips with fn heads is $p^{fn}(1 - p)^{n - fn}$, and there are $\frac{n!}{(fn)!(n-fn)!}$ such sequences.

Now let us turn to Barking Dog. In this case too, Jacob should become confident that $P \approx f$. The dog is irrelevant. Why so?

The chance P determines how likely any given number of heads is, and the number of heads (along with Q) determines how likely Gemma is to bark. But if we fix the number of heads, P plays no further role in determining how likely Gemma is to bark. More precisely, *given* that there were a total of fn heads, the probability that Gemma never barks is q^{fn}, regardless of any hypothesis about P:

$$\Pr(\textsc{no bark} \mid F = f, P = p) = q^{fn}.$$

So, once Jacob knows $F = f$, NO BARK provides no further evidence about the value of P. On learning that $F = f$ and that Gemma did not bark, Jacob's confidence that $P \approx f$ should end up exactly the same as in the previous case, Ordinary Coin.

Barking Dog vs The Martians

Now I will explain why I think Jill in The Martians, like Jacob in Barking Dog, should become confident that $P \approx f$.

As I said in section 3, the salient difference between the cases is that, in The Martians, the total number of observers is a matter of chance (it depends on whether and when there is a fatal eruption), whereas in Barking Dog we naturally imagine that the total number of observers is independent of anything that happens with the coin.

I put it this way, in terms of what we 'naturally imagine', because our analysis of Ordinary Coin and Barking Dog did not involve any explicit premise about the total number of observers, and it seems odd to claim that our ordinary chance reasoning involves a hidden premise of this type. In fact, our analysis of these cases flowed entirely from the premise that, conditional on $P = p$, Jacob should initially assign probability p to the coin's landing heads:

$$\Pr(\textsc{heads} \mid P = p) = p. \tag{*}$$

The worry must be that the analogous premise does not apply to Jill.

There is, it turns out, a reason to worry here (though not, I claim, a good one). In talking about Jacob's priors Pr, I meant *prior to observing the frequency* and (therefore) *prior to learning whether Gemma barks*. But we can imagine a more radical 'prior circumstance' in which Jacob also lacks all *self-locating* evidence: though he has all the requisite background knowledge in third-personal terms, he does not know which of the many people in the world he is; he does not know *I am Jacob*.[8]

Let $\mathrm{Pr}_{\circ}$ represent the credences of this self-ignorant version of Jacob. What probability should he assign to each coin-flip landing heads? One might think that the answer is the same as before:

$$\mathrm{Pr}_{\circ}(\textsc{heads} \mid P = p) = p. \tag{**}$$

Unfortunately, (*) and (**) are not generally compatible.

To say more, let's assume that although Jacob does not initially know *I am Jacob*, he does know *I exist* and also *I am a person* (as well as *Jacob exists* and *Jacob is a person*); beyond this, he lacks self-identifying information. Then, he should think that *I am Jacob* is relatively likely conditional on there being few people, and relatively unlikely conditional on there being many. Learning *I am Jacob* should therefore shift probability towards worlds with relatively few people.[9] And so, if the result of the coin-flip is correlated with the total number of people, learning *I am Jacob* will change the probability he assigns to heads; it will tend to favour whichever result is correlated with there being fewer people.[10] Under these circumstances, Jacob can align his credence in heads with the known chance either (**) before learning *I am Jacob* or (*) afterwards, but not both. In particular, if something like (**) is correct, then Jill

[8] Cf. Lewis (1979). There is a complication I will gloss over here. We should imagine that Jacob doesn't know other indexicals like *It is now 5 o'clock*. Because of this, the relevant thing is not ultimately the number of observers but the number of 'observer-moments' (people at times).

[9] For example, suppose Jacob's credences are initially split between two possible worlds: world 1 contains 10 people and world 2 contains 100. Assuming that he gives roughly equal probability to different hypotheses about his own identity in each world, his credence in world 1 is split roughly equally between 10 different self-locating hypotheses, and roughly equally among 100 for world 2. Then $\mathrm{Pr}_{\circ}$(I am Jacob | world 1) $\approx 1/10$ whereas $\mathrm{Pr}_{\circ}$(I am Jacob | world 2) $\approx 1/100$. So *I am Jacob* provides support for world 1 over world 2 by a factor of 10.

[10] This is a version of the 'doomsday argument' to the effect that doomsday events are more likely than one would naively think (see Thomas, 2021, for discussion). Compare also the famous case of Sleeping Beauty (Elga, 2000). There as here, purely self-locating evidence ('It's Monday') bears on third-personal hypotheses about the world ('Heads' versus 'Tails') because different third-personal hypotheses leave open different self-locating possibilities. (The 'double-halfer' view about Sleeping Beauty essentially denies that purely self-locating evidence bears on third-personal hypotheses even in this case.)

should not satisfy (*), and so the analysis of her case will be different from Jacob's.

Now, we could block this line of thought by removing the assumption that the radically ignorant version of Jacob knows *I exist* and *I am a person*. That is, we could stipulate that the Pr_0 that appears in (**) does not already incorporate these pieces of evidence. We could hold that one or the other of these pieces of evidence bears on the total number of people (and therefore on the probability of heads) in just the right way so as to reconcile (**) with (*).[11]

My own view of the matter is simpler: we should endorse (*) and not worry too much about whether (**) is true.[12] The compelling principle is that one should align one's credences with the known chances in normal situations. As to the bizarre situation of having no idea who or where or when one is in the world, let the chips fall where they may. It would be unfortunate—and from a theoretical point of view, it is completely unnecessary—for our ordinary reasoning about chances to be hostage to an analysis of the downstream effects of coin-flips on the total number of people, or hostage to speculation about the evidential weight of one's own personhood and existence.

While this is bound to be one of the points of controversy, (*) is certainly defensible both in Jacob's case and in Jill's, and from there the analysis proceeds in the same way: they should both end up confident that P is close to f.

The Martians vs Supervolcano

Now we get to the pair of cases that seems to speak most directly to the original worry about the anthropic shadow. The main difference between Jack and Jill is that Jill *could* have discovered a fatal eruption in Earth's past; Jack *could not* have, since a fatal eruption would have precluded his existence. Should this make a difference to their credences about P?

I claim it shouldn't. The overall point is one I already mentioned in the introduction: Jack and Jill have essentially the same evidence about

[11] In the literature, the relevant assumption about the evidential bearing of *I exist* is known as 'the self-indication assumption'; the status of *I am a person* is relevant to discussions of one's 'reference class'. See Bostrom (2002) for these ideas.

[12] I develop this view more formally in Thomas (2021), and draw out its consequences. Gallow (forthcoming) defends a similar view from a different angle.

the value of P, and the allegation that Jill but not Jack *could* have had different evidence is irrelevant. But let's dig a bit deeper.

Let me first dispel the thought that Jack can know *a priori* that there were no fatal eruptions (his very existence depends upon it!) whereas Jill has to find it out. This is not right (and even if it were right, it isn't clear why it would matter). Just like Jill, Jack's basis for thinking there were no fatal eruptions through the nth generation is his knowledge of the basic set-up and the fact that there is an nth Earthling. It's true that Jack himself is the nth Earthling, but this is not something he knows *a priori*. Indeed, *a priori*, he might have lived on Mars; he might have been Jill!

To go further, we can fill in the details of the case so that Jack and Jill end up with nearly all the same evidence. For example:

1. They both know the basic chance set-up governing eruptions on Earth, and they both know the value of Q, although neither of them knows the true value of P.

2. They both know the same things about the history of the universe up through the present time, from a third-personal perspective. They both know the frequency of eruptions, they both know that Jack and Jill exist, and they both know that Jack and Jill are the nth Earthling and the nth Martian, respectively.

3. They also have a lot of self-locating evidence in common. For example, each of them knows *I exist* and *I am Jack or Jill.*

There is one main piece of evidence that we cannot insist they both share. Jack knows the merely self-locating *I am Jack*, whereas Jill knows *I am Jill.* So, if we think there is an important difference between Jack's evidence and Jill's, it must come down to the idea that, even once we have taken into account all their considerable shared evidence, the merely self-locating *I am Jack* has some further bearing on the value of P that *I am Jill* does not.[13]

I do not see what this bearing would be. In the previous section, I pointed out that merely self-locating evidence ('I am Jacob') *can* have a

[13] Perhaps there are some other sorts of evidence that Jack and Jill cannot be stipulated to share, such as some essentially private evidence about phenomenal experiences. I will gloss over this possibility, though again, the question would be whether (given all the evidence they *do* have in common) this private evidence has some further bearing on the value of P.

bearing on the probability of third-personal hypotheses. This happens when the rival third-personal hypotheses leave open different numbers of self-locating hypotheses. For example, if Jacob knows he's a person but not which person, then learning *I am Jacob* shifts probability towards worlds with fewer people. In the present case, however, the evidence shared by Jack and Jill leaves open exactly two self-locating hypotheses, *I am Jack* and *I am Jill.* So here there is no reason to think that learning one of these hypotheses should change Jack's views about the chances.

Indeed, the claim would have to be that, given all the shared evidence listed above, *I am Jack* is more likely conditional on some values of P than it is conditional on others. But, again, conditional on each possible value of P, the shared evidence leaves open exactly two possibilities: *I am Jack* and *I am Jill.* The most natural view is that these two possibilities are equally likely. Thus, *I am Jack* has the same probability conditional on each value of P.

Now, a subjective Bayesian might think that, even if this is the most natural view, it is permissible for Jack to distribute his priors in a different way, so that learning *I am Jack* does provide some further evidence about P. In Thomas (2021) I argue against this thought: it is rationally required in cases like this to treat *I am Jack* and *I am Jill* as equally likely.[14] But even if we agree with the subjective Bayesian, we haven't identified anything about the situation that distinguishes Jack's position from Jill's. If, when all is said and done, Jill must be confident that $P \approx f$, then so must Jack.

4 Ćirković et al.

That completes my argument that Jack should become confident that the chance of an eruption is close to the observed frequency. It is just like an ordinary coin-flip case, and, in that sense, the anthropic shadow is irrelevant.

Why, then, do Ćirković et al. suggest that attending to the historical record will lead to underestimating the chances? The key example that they use to support this claim is described as follows:

> For instance, suppose $Q = 0.1$ and $P = 0.5$, correspond-

[14] See also the principle of 'indifference' defended by Elga (2004), and, for dissent, Weatherson (2005).

> ing to a fair-coin-toss chance that a [supervolcanic] event occurs once per [period], and that the probability of human survival following such an event is 0.1…[T]he actual probability [the chance?] of such an event is 5.5 times our initial estimate [the frequency?]. (1497)

As my bracketed glosses suggest, I think it is difficult to unpick what exactly is being claimed in this passage. So I will explain where the number 5.5 comes from, and what it really means, in this section. Throughout this discussion, consider Jack's position before he learns the frequency, but after he knows that there were no fatal eruptions.[15]

I think it is useful to see what is going on from two different angles. First: in this position, Jack should be inclined to think that there were *unusually few* eruptions. That is, he should be inclined to think that, if $P = p$, then the past frequency of eruptions was smaller than p. This is because the presence of few eruptions increases the chance that none of the eruptions were fatal. And the lower q is (i.e., the deadlier eruptions tend to be), the larger this discrepancy. More formally, suppose Jack were to learn that $P = p$. What probability would he then assign to there being an eruption before each generation? Not the unconditional chance of an eruption, p, but the chance *conditional on no fatal eruption*. By definition, this is the chance of a non-fatal eruption (pq) divided by the chance of no fatal eruption ($1 - p + pq$); in short,

$$p_0 := pq/(1 - p + pq).$$

The same sort of calculation we did with binomial distributions in Coin Flip leads us to this conclusion:

(a) Jack should be confident, conditional on $P = p$, that $F \approx p_0$.

Now, p_0 is always smaller than p, and the ratio between them increases as q gets smaller (further, one might say, into the anthropic penumbra). In our particular case, if $q = 1/10$ and $p = 1/2$, then $p_0 = 1/11$. The discrepancy betwen p and p_0 is the factor of 5.5 cited by Ćirković et al.

However, it would be wrong to conclude from this calculation that Jack should treat the observed frequency as a 5.5-fold underestimate of the chance. The question about how Jack should treat the observed frequency is the question of what estimate Jack should adopt for P if he learns $F = f$. The answer I have given is

[15] By the way, precisely analogous points apply to Jill in The Martians and to Jacob in Barking Dog.

(b) Jack should be confident, conditional on $F = f$, that $P \approx f$.

It may be surprising that (a) and (b) are compatible. Here's the point. According to (a), if Jack learns $P = p$, the fact that there were no fatal eruptions provides evidence that there were unusually few eruptions (p_0 rather than p per period). This is because there are more likely to be no fatal eruptions if the total number of eruptions is smaller. However, if Jack learns $F = f$, the fact that there were no fatal eruptions provides no additional information about P, and does nothing to suggest that P is 'unusually large' (larger than f). This is because, once we have fixed the total number of eruptions, the underlying chance P has no bearing on the number of *fatal* eruptions.

Let us look at this same issue from a different angle, which will bring out the connection with claims (A)–(C) in the introduction, and the issue of base-rate neglect.[16] Again, consider Jack's position when he knows there were no fatal eruptions, but does not yet know the value of F. At that point, I claim, learning $F = f$ provides strong evidence that $P > f$. More precisely, define

$$f_0 := \frac{f}{q + (1-q)f}.$$

(This is the value of P that makes it most likely that $F = f$; that is, if $p = f_0$, then $p_0 = f$.) One can check that f_0 is always greater than f, and, I claim,

(c) Learning $F = f$ would provide Jack with strong evidence that $P \approx f_0$.

For example (really the same example as before), if $q = 1/10$, then $F = 1/11$ is strong evidence that $P \approx 1/2$, 5.5 times bigger than the observed frequency! And this evidence gets stronger as n gets larger.

[16] Here is a classic example of base-rate neglect. Let D be the hypothesis that Jones has a certain disease, and Neg the hypothesis that Jones tests negative. If $\Pr(\text{Neg} \mid D) < 0.01$, then a negative test provides strong evidence that Jones does not have the disease. In fact, it 'rules out the hypothesis that Jones has the disease, at 99% confidence'. However, this does not mean we should become anything like 99% confident that Jones is disease-free. It does mean that we should strongly update in that direction. But if we started off highly confident that Jones has the disease (a high base-rate), we might still end up quite confident that she has it.

It may be surprising that (c) is compatible with (b). Indeed, one might naively conclude from (c) that, if n is sufficiently large, Jack should become confident that $P \approx f_0$, rather than that $P \approx f$. However, this conclusion would again be too hasty. For there are two key pieces of evidence:

(i) There were no fatal eruptions.

(ii) $F = f$.

The point is that (i) is evidence that $P \approx 0$. The bigger n is, the stronger the evidence. Since we are considering Jack's position when he knows (i) and is yet to find out (ii), he should be confident that P is close to 0, which is less than f. Claim (c) tells us that conditionalizing on (ii) will shift probability towards $P \approx f_0$, which is greater than f. Claim (b) tells us that the *overall* effect is to make Jack confident that $P \approx f$. It is true that, by considering larger and larger values of n, we can make the shift towards $P \approx f_0$ more dramatic. However, increasing n also makes Jack antecedently more confident that $P \approx 0$. So, overall, increasing n just makes Jack more confident that $P \approx f$, not that $P \approx f_0$. We can see this overall effect very clearly if we imagine what would happen if someone learnt (i) and (ii) in the opposite order. Learning (ii) first would provide powerful evidence that $P \approx f$. And once one knows (ii), (i) has no further bearing on the value of P. I made exactly this point about Jacob's situation in Barking Dog: once he knows $F = f$, NO BARK provides no further evidence about the value of P.

5 Other Literature

So far I have focused on the model used by Ćirković et al. Here I sketch how my analysis extends to the models used by Tegmark and Bostrom and by Snyder-Beattie et al.

Snyder-Beattie et al.

Snyder-Beattie et al. try to estimate the annual extinction risk for the species *Homo sapiens*. Setting anthropogenic risks aside, and focusing on natural background risks, their model assumes an unknown, constant rate μ of extinctions per year.[17] Let T be the lifespan of our species. Then, given any candidate value μ_0 of μ, the prior distribution of T is $\Pr(T \geq t \mid \mu = \mu_0) = e^{-\mu_0 t}$. Henceforth let $t_{\text{now}} = 200\,\text{kyr}$ be an estimate for the current age of humanity. Then the 'likelihood function'

$$\mathscr{L}(\mu_0) := \Pr(T \geq t_{\text{now}} \mid \mu = \mu_0) = e^{-\mu_0 t_{\text{now}}}$$

gives the relative strength with which the evidence $T \geq t_{\text{now}}$ supports different hypotheses about μ. The authors find that $T \geq t_{\text{now}}$ supports a low value of $\mu \approx 10^{-8}$ a million times more strongly than it supports a high value of $\mu \approx 7 \times 10^{-5}$. In that sense, 7×10^{-5} provides an upper estimate for μ.[18]

So far, so good. But Snyder-Beattie et al. worry that such estimates are subject to an 'observation selection bias' related to the anthropic shadow. This is the passage I quoted in the introduction. Translated into the present context, the worry is that a species like *Homo sapiens* might require a long gestation time (as it were) before its members count as 'observers' in the sense that they are able to inspect the historical record.[19] Any human observer would necessarily find such a long period in the human past, and the authors worry that it would be improper or misleading to use the existence of this time-period as evidence about μ.

Let me reiterate the basic reason this worry is misguided, before turning to the formal treatment of observation selection biases. Even if human observers will 'necessarily' find a long gestation time in humanity's past, this isn't a matter of *epistemic* necessity. The need for a gestation time is something we might discover through such empirical sciences as geology, evolutionary biology, history, and so on. Together with the fact that *we* are human observers, it implies that there was a long gestation time in our past; and this empirical conclusion is perfectly good evidence that the rate of extinction is low. Again, the fact that (as a matter of nomological necessity) we could not easily have had

[17] To ease comparisons with the literature, I use 'μ' and later 'τ' to denote random variables, in exception to my general convention of using capital letters.

[18] Of course, this isn't the full Bayesian analysis; if one's prior over values of μ strongly favoured high values, then one might still end up confident in values higher than 7×10^{-5}.

[19] They write: 'If desired, we could more crisply define this observerhood property as the ability for a species to collect reliable data on its own track record of survival (e.g. via fossil dating) and analyse it.'

different evidence does not change the weight of the evidence we do have.

In thinking this through, it may help to return again to the perspective of Jill the Martian. Consulting the historical record on Earth, she observes that *Homo sapiens* has been around for 200 kyr and that within that time it has reached observerhood. Given a constant-rate model of extinction, she estimates μ in exactly the way I described above, using the likelihood function $\mathscr{L}$. But, as I've argued, what goes for Martians goes for Earthlings too. That's because merely self-locating facts like 'I'm Jack the human' as opposed to 'I'm Jill the Martian' provide no evidence about μ, given the rich background information that Jack and Jill hold in common.

If these general comments are correct, they raise a puzzle: what is going wrong with the more formal analysis by Snyder-Beattie et al. that purports to show observation selection bias in action? Let me sketch that analysis along with my diagnosis.

The authors extend their model to include a random variable S giving the time from the origin of humanity to the point at which humans become observers (we can take $S = \infty$ if humans never do). They define $F_S(t) = \Pr(t \geq S \mid T \geq t, \mu = \mu_0)$ to be the probability that humans become observers before time t, given that they survive until at least time t; for the types of risks under consideration, we can assume that $F_S(t)$ does not depend on μ_0. The authors then reach modified estimates for μ using the likelihood function

$$\mathscr{L}^*(\mu_0) := \Pr(T \geq t_{\text{now}} \mid T \geq S, \mu = \mu_0).$$

That is, they treat as background evidence the fact that humans at some point become observers ($T \geq S$) and ask what, against that background, the additional evidence $T \geq t_{\text{now}}$ tells us about μ. While they consider a variety of distributions for S, we can get a sense of their results by considering their 'Model 4', according to which observerhood occurs at a fixed time t_{obs}. Thus

$$\mathscr{L}^*(\mu_0) = \Pr(T \geq t_{\text{now}} \mid T \geq t_{\text{obs}}, \mu = \mu_0) = e^{-\mu_0(t_{\text{now}} - t_{\text{obs}})}.$$

Let's consider the extreme case in which we only recently passed the threshold t_{obs} for observerhood, so $t_{\text{obs}} = t_{\text{now}} = 200$ kyr. Then $\mathscr{L}^*(\mu_0) = 1$ independent of μ_0. This conclusion is supposed to represent mathematically the thought that if observerhood requires at least 200 kyr of gestation, then observing such a past is uninformative about the rate of extinction risk.

There are two problems with this analysis, as far as I can see. First, $\mathscr{L}^*$ tells us to what extent the observation $T \geq t_{\text{now}}$ supports various values of μ, given background evidence $T \geq S$. However, $T \geq S$ is itself significant evidence, and what we want to know is what *all* this evidence tells us about μ. Second, we in fact know something stronger than $T \geq t_{\text{now}}$ and $T \geq S$; we know $T \geq t_{\text{now}} \geq S$.[20]

For these two reasons, the likelihood function that accounts for all our evidence in the model is not $\mathscr{L}^*$ but

$$\mathscr{L}^{**}(\mu_0) := \Pr(T \geq t_{\text{now}} \geq S \mid \mu = \mu_0).$$

Using the definition of conditional probability, we can factor this as

$$\begin{aligned}\mathscr{L}^{**}(\mu_0) &= \Pr(t_{\text{now}} \geq S \mid T \geq t_{\text{now}}, \mu = \mu_0) \times \Pr(T \geq t_{\text{now}} \mid \mu = \mu_0) \\ &= F_S(t_{\text{now}}) \times \mathscr{L}(\mu_0).\end{aligned}$$

Recall that $F_S(t_{\text{now}})$ does not depend on μ_0, so $\mathscr{L}^{**}$ is just $\mathscr{L}$ times a constant. Therefore the estimates for μ that we get using $\mathscr{L}^{**}$ will be exactly the same as the ones we got using $\mathscr{L}$, before we tried to take observation selection effects into account. For example, it is still the case that our evidence supports a low value $\mu \approx 10^{-8}$ a million times more strongly than it supports a high value of $\mu \approx 7 \times 10^{-5}$. There are no observation selection effects, as far as this model goes.

Tegmark and Bostrom

The story about Tegmark and Bostrom (2005a) is similar in its essentials, and I will treat it more briefly, with less conceptual discussion.[21] These authors want to estimate the rate μ of events like vacuum decay, measured in events per gigayear, so that the typical timescale $\tau := 1/\mu$ has units of gigayears. Such an event would annihilate any planets nearby and prevent any new ones from forming. The authors attempt to estimate μ (or equivalently τ) in a way that is immune to the effects of the anthropic shadow.

[20] For simplicity, I'm assuming that we can rely on 200 kyr as a precise estimate of humanity's current age.

[21] I refer in this section to the extended version of their paper, Tegmark and Bostrom (2005b).

Let T_v be the time of the first vacuum decay after the big bang. Vacuum decay is assumed to happen at a constant rate $1/\tau$, so that $\Pr(T_\text{v} \geq t \mid \tau = \tau_0) = e^{-t/\tau_0}$. Now, naively, we could take $t_\text{now} := 13.7\,\text{Gyr}$ as an estimate of Earth's current age, and generate estimates for τ using the likelihood function

$$\mathscr{L}(\tau_0) := \Pr(T_\text{v} \geq t_\text{now} \mid \tau = \tau_0) = e^{-t_\text{now}/\tau_0}.$$

However, in an effort to circumvent the anthropic shadow, the authors complicate things by introducing two further events: a 'formation' event, occuring at time T_f, which results in the formation of Earth *if* there has not yet been vacuum decay ($T_\text{f} \geq T_\text{v}$); and an 'observerhood' event, occuring at time $T_\text{f} + T_\text{o}$, which results in the evolution of observers on Earth, *if* there has still not yet been vacuum decay ($T_\text{v} \geq T_\text{f} + T_\text{o}$). The authors use a particular model of planetary formation to give a prior probability distribution for T_f; the distribution for T_o turns out not to matter, assuming that T_v, T_f, and T_o are independent.

Henceforth take $t_\text{f} = 9.1\,\text{Gyr}$ as an estimate of Earth's actual formation time. To get an estimate for τ, Tegmark and Bostrom consider the likelihood function

$$\mathscr{L}^*(\tau_0) := \Pr(T_\text{f} \geq t_\text{f} \mid T_\text{v} \geq T_\text{f} + T_\text{o}, \tau = \tau_0).$$

$\mathscr{L}^*(\tau_0)$ is the chance that Earth would form at least as late as it actually did, given that observers managed to evolve. If τ is low, then this chance is low (more likely, Earth evolved early, allowing more time for evolution to do its work). Thus, the fact that Earth formed so late is (in this context) evidence that τ is high. Tegmark and Bostrom find the smallest τ_0 such that $\mathscr{L}^*(\tau_0) \leq 0.05$. This provides a lower estimate for τ, namely, $\tau \geq 2.5\,\text{Gyr}$.[22]

Although the method here is not quite parallel to the one used by Snyder-Beattie et al., it raises essentially the same two issues.

First, the above analysis effectively treats the evolution of observers ($T_\text{v} \geq T_\text{f} + T_\text{o}$) as background evidence, and against this background asks what the late formation of the Earth tells us about the value of τ. But we should also consider what the background evidence tells us. In fact,

[22] As they put it in standard frequentist terminology, 'we can rule out the hypothesis that $\tau < 2.5\,\text{Gyr}$ at 95% confidence' (p. 3). They give estimates at other levels of confidence as well, and some other considerations about the robustness of their model.

second, we should ask what is supported by *the entirety* of our evidence about the parameters of the model. We have ballpark figures for both T_f and T_o, and, more importantly, we know that Earth is $t_\mathrm{f} = 13.7\,\mathrm{Gyr}$ old, so $T_\mathrm{v} \geq 13.7\,\mathrm{Gyr} \geq T_\mathrm{f} + T_\mathrm{o}$. Thus the likelihood function that captures all our evidence in this model is

$$\mathscr{L}^{**}(\tau_0) = \Pr(T_\mathrm{v} \geq 13.7\,\mathrm{Gyr}, T_\mathrm{f} \approx 9.1\,\mathrm{Gyr}, T_\mathrm{o} \approx 4.5\,\mathrm{Gyr} \mid \tau = \tau_0).$$

On the standing assumption that T_f and T_o are independent of T_v, we find that $\mathscr{L}^{**}$ is the same as $\mathscr{L}$ up to a constant scale-factor. We could, then, just as well have used $\mathscr{L}$ to generate estimates for τ. For example, the condition $\mathscr{L}(\tau_0) \leq 0.05$ gives a lower estimate of $\tau \geq 4.5\,\mathrm{Gyr}$.

6 Conclusion

My analysis of the anthropic shadow highlights three main points.

First (as discussed throughout, but especially in section 3), the basic worry about the anthropic shadow is misguided. Jack and Jill differ with respect to whether their existence depends on the absence of a fatal supervolcano eruption, but their evidence does not differ in any way relevant to the rate of eruptions.

Second (as discussed mainly in section 3), a full analysis of the situation depends on theoretical choices about the circumstances under which one should align one's credences with the known chances. My own view is that Jacob and Jill are both in the right circumstances to do so. Even if the reader is not convinced of this view, the general point is interesting because these theoretical choices are not explicitly discussed in the small literature on the anthropic shadow, even as they are contentious within the wider literature on 'anthropics'.

Third, analyses that do claim to find an effect of the anthropic shadow are misleading because of a form of base-rate neglect. For example, Ćirković et al. do not take into account the fact that a long time without omnicidal events is good evidence that the rate of potentially omnicidal events is low. And Snyder-Beattie et al. do not take into account the fact that a long gestation period for humanity is good evidence that the extinction rate for humanity is low. Of course, the cogency of this critique depends on the first point made above. Jill the Martian can obviously appeal to these long 'safe' periods as evidence about extinction rates—and what goes for Jill should go for us as well.

Thus is the anthropic shadow finally dispelled.

References

Bostrom, N. (2003). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*, Routledge.

Ćirković, Milan M., Anders Sandberg, and Nick Bostrom (2010), 'Anthropic shadow: Observation selection effects and human extinction risks', *Risk Analysis* 30(10):1495–1506. doi:10.1111/j.1539-6924.2010.01460.x

Elga, Adam (2000). 'Self-locating belief and the Sleeping Beauty problem', *Analysis* 60(2):143–47.

Elga, Adam (2004). 'Defeating Dr. Evil with self-locating belief', *Philosophy and Phenomenological Research* 69(2):383–396.

Gallow, J. Dmitri (forthcoming). 'Two-dimensional *de se* chance deference', *Australasian Journal of Philosophy*.

Lewis, David (1979). 'Attitudes *de dicto* and *de se*', *Philosophical Review* 88(4):513–543.

Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury.

Snyder-Beattie, Andrew E., Toby Ord, and Michael B. Bonsall (2019). 'An upper bound for the background rate of human extinction', *Scientific Reports* 9, 11054. doi:10.1038/s41598-019-47540-7

Tegmark, Max and Nick Bostrom (2005a), 'Is a doomsday catastrophe likely?', *Nature* 438: 754 (with supplementary materials). doi:10.1038/438754a

Tegmark, Max and Nick Bostrom (2005b), 'How unlikely is a doomsday catastrophe?', https://arxiv.org/abs/astro-ph/0512204.

Thomas, Teruji (2021). 'Doomsday and objective chance', GPI Working Paper No. 8-2021. https://globalprioritiesinstitute.org/doomsday-and-objective-chance-teruji-thomas/

Weatherson, Brian (2005). 'Should we respond to Evil with Indifference?', *Philosophy and Phenomenological Research* 70(3):613–635.