# Anthropic Reasoning in the Great Filter

by

Caitlin Grace

Submitted in partial fulfilment of the requirements for the degree of
Bachelor of Science with Honours
in the Fenner School of Environment and Society,
Australian National University
October 2010

# Candidate's Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university. To the best of the author's knowledge, it contains no material previously published or written by another person, except where due reference is made in the text.

Caitlin Grace Date:

# Acknowledgements

Many thanks to Robert Wiblin, Hugh Parsonage, Robin Hanson, Bindu Johnson, Mummy, Carl Shulman, David Dumaresq, David Chalmers, Nick Bostrom, Anders Sandberg, Anna Salamon, many blog commenters, Rob Dyball, Daniel Nolan, Stuart Armstrong, Roko Mijic, the imaginary god of truth, Michael Blume, Ramana Kumar, Toby Ord, Mike Hancock, Jason Roy, Amy Peters, Mitchell Porter, Alex Dixon, Sarah Andrews, Victoria ApSimon, Karen Warnes, and other people.

# Abstract

We apply three popular principles for reasoning about indexical information to the Great Filter model of the development of life. The aim is to discover the effect of each principle on the expected level of extinction risk humanity faces, given the Great Filter model. The principles are contentious; at most one of them is correct. We find that the principles investigated imply that human extinction is a greater risk than otherwise thought. One of the principles, the Self Sampling Assumption, implies different results with some parameters. The other two principles reliably imply that the risk of extinction has been underestimated. Other details of the results differ between principles. We explore these, discuss the implications of the findings for specific causes of extinction, and examine the plausibility of the principles. We conclude that despite continuing uncertainty in the correct method of indexical reasoning, we must increase our expectations of human extinction, as long as we are confident that one of the primary reasoning principles under contention is correct.

# Table of Contents

## List of Figures

# List of Tables



# List of Boxes



# List of Equations



# List of acronyms and abbreviations

| | |
|---|---|
| **SIA** | Self Indication Assumption |
| **SSA** | Self Sampling Assumption |
| **FNC** | Full Non-indexical Conditioning |

# Chapter 1: Introduction

## 1.1 Aims

This project will apply principles of anthropic reasoning to the Great Filter model of the development of life in the hope of learning about the level of extinction risk that humanity faces. The rest of this chapter introduces the Fermi Paradox, the Great Filter, extinction risks, and other barriers to visible space colonization, before explaining the current project in more detail.

## 1.2 The Fermi Paradox

The human race has not observed credible signs of technologically advanced extraterrestrial life (Davies 2010). Even if life started on other planets at a similar time to Earth, it could be millions of years ahead or behind us, technologically speaking. If Earth is not extremely unusual, and the technology expected by many in the coming centuries takes less than millions of years to arrive, other civilizations should have developed such technology already. If so, and if that technology allows interference with the cosmos at a large enough scale to be widely visible, as expected, we should have expected to have observed such civilizations long ago. The lack of any such observations is known as Fermi's paradox (Webb 2002).

## 1.3 The Great Filter

Any causal path between the existence of an arbitrary lifeless star, and a species near that star engaging in interstellar travel, contains a number of necessary steps. The Fermi Paradox tells us that these together are very hard to accomplish (Hanson 1998b). Robin Hanson named this set of difficult steps 'The Great Filter' (1998b). Each step 'filters out' solar systems, preventing their inhabitants proceeding further toward interstellar colonization. For instance, the development of life is plausibly a hard step, meaning the proportion of lifeless solar systems that develop life may be very small.

To be clear, it is the causal paths between dead matter and colonizing life themselves that are filtered out, not specific creatures along those paths. At any given point on that path there will be different physical structures, such as microbes or people, so filters at different points along the path will involve different events, preventing different developments. For instance a causal path may be filtered at the microbe stage by evolution not finding the next development fast enough to allow time to escape the planet before it becomes uninhabitable. In cases like this the group of creatures may be filtered without being harmed significantly at the time.

There are around $10^{22}$ stars visible to us (Craig 2003), and the Fermi Paradox demonstrates that the number of stars in our past light cone expected to reach the end of the filter's steps (visible interstellar colonization) within the time they have had is unlikely to be much greater than one. This means a given star similar to those in our past light cone has at most around a $10^{-}$

$^{22}$ chance of reaching the end of the filter during the current age of the universe, and the chance could be much, much less (Hanson 1998b). This improbability of reaching the end must be divided between the steps somehow. However the probability of passing each of these steps has been estimated, and their product is too high (Hanson 1998b). If they were correct we should expect the universe to be teeming with life. This means at least one step is harder than we realize. There may even be large steps we do not know about.

Working out the shape of the filter is important because we do not know whether the finest filtration occurs in steps humanity has already passed, or if some of it occurs in steps yet to come. If it mostly lies in steps past, mankind is extremely lucky to be here and faces a bright future. If much lies in steps yet to come, mankind will very likely be filtered out. This means if we learn that a step in our past is relatively easy, for instance through discovering primitive life on other planets, we should fear more for our own safety (Bostrom 2008; Hanson 1998b).

One might wonder if the value-laden notion of progress along a path is appropriate. Colonizing the galaxy is not everyone's idea of progress, which some suggest is part of the reason civilizations end up not doing it (Miller 2006). The valence of these concepts is not relevant to the reasoning however. 'Wanting to substantially colonize space' could be a step in itself, and if we are filtered out at that point it need not be a bad thing.

Future barriers to colonizing space need not involve danger to humans living at the time, but many plausible barriers do. Possible filter steps that appear safer seem unlikely to account for very much of the filter, as will be discussed below. For instance permanent technological limitations could be a filter step in theory, and would not lead to human extinction until Earth became uninhabitable. However, most known potential filter steps in our future are 'existential risks'.

## *1.4 Existential Risks*

### 1.4.1 What are they?

Bostrom defines an existential risk as 'one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential' (Bostrom 2002b). Many such risks are worst-case versions of threats that are most likely to be less dangerous. For instance current climate change is projected to cause damage far less than the destruction of civilization, but in the tip of the worst tail of the probability distribution, climate change is an existential risk (Matheny 2007; Weitzman 2009).

### 1.4.2 The Stakes

Existential risks put all future generations at stake on top of those killed at the time. This means they warrant far more consideration than more likely near-global catastrophes under many common ethical stances (Sandberg et al. 2008; Bostrom 2003b; Parfit 1984; Sandel 2006).

### 1.4.3 The Changing Odds

There have always been some existential risks, such as that of large asteroid impacts; however, new risks are emerging as humankind attains the technological capacity to significantly damage its habitat (Bostrom 2002b). Humankind's ecological influence is growing (Millennium Ecosystem Assessment 2005) and technologies expected in the coming century are often feared to be unusually dangerous (e.g. Joy 2000; Yudkowsky 2008; Bostrom 2002b; Sandberg & Bostrom 2008; Rees 2004). These factors suggest existential risks are increasing in number.

Academic opinion places the risk of human extinction on par with leading risks of individual death, for a person born today. John Leslie estimates a greater than 30 percent risk of human extinction within five centuries (Leslie 1998). Nick Bostrom argues for more than 25 percent risk over the twenty first century (Bostrom 2006). The median extinction risk by 2100 estimated by attendees of the Global Catastrophic Risks Conference was 19 percent (Sandberg & Bostrom 2008). Sir Martin Rees estimates 50 percent risk of civilization collapse over the twenty first century, though that does not imply extinction (Rees 2004).

### 1.4.4 Overview of the risks

Many risks could destroy humanity suddenly, but smaller threats also have the potential to destroy technological civilization, which may allow humans to be destroyed later by something currently non-threatening, for instance disease or smaller climate changes (Hanson 2008).

Non-anthropogenic extinction risks arise from planetary processes, astrophysical processes, and potentially processes in other realities (F. C. Adams 2008, p.33; Bostrom 2003a). Natural ecological change, including runaway climate change, could pose a threat in the distant future (Wills 2008, p.48; Allen & Frame 2008, p.256; Posner 2004, p.33). A disease with the right characteristics could produce a pandemic large enough to destroy humanity, helped by increasing contact between regions (Kilbourne 2008, p.287). Asteroid and comet impacts are an ongoing threat, though a relatively well monitored one (Napier 2008, p.222). Many other astrophysical processes pose some risk, for instance vacuum state decay (Coleman & De Luccia 1980). If the 'simulation argument' is true, humanity may be living in a simulation, in which case there is an extinction risk from those who simulate our world losing interest in running it (Bostrom 2003a).

Humans already have the potential to cause several extinction risks. Present technologies allow enough interference with ecological processes that humans could plausibly trigger lethally extreme runaway climate change (Allen & Frame 2008, p.265; Posner 2004, p.3). While this is very unlikely, even moderate climate change and resource depletion is expected to intensify conflicts which will amplify other risks, such as from increasingly powerful technology available to increasingly small groups (Rees 2004, p.113). Another indirect risk resulting from ecological damage is the temptation to interfere on a larger scale to fix the problems, such as

through geoengineering, which may lead to other existential risks (Cirkovic & Cathcart 2004). The nuclear winter following a war could cause catastrophic consequences, perhaps threatening humanity (Robock et al. 2007; Sandberg et al. 2008; Cirincione 2008, p.381).

Worse risks come from technologies expected to advance in the next century. Synthesizing pathogens is increasingly cheap (Williams 2006) and designing dangerous pathogens is increasingly successful (Kwik et al. 2003). Should these trends continue, it is more likely that pathogens optimized for a lethal pandemic will be released (either accidentally or intentionally). Nanotechnology is expected to pose similar risks of tiny agents capable of powerful harm becoming out of control, purposely or accidentally (Phoen & Treder 2008, p.481). Physics experiments become more advanced with time, and arguably carry tiny but nonzero risks of disastrous consequences (Dar et al. 1999). Human level artificial intelligence is argued to pose a significant risk due to its hypothesized tendency to recursively self improve, potentially becoming many orders of magnitude more intelligent than a human quite suddenly and probably having been programmed with fairly different 'values' to humans (Yudkowsky 2008).

Technological risks we foresee are mostly expected in the near future. It may be that most serious risks are in this century, or just that those are easiest for us to imagine. Unforseen risks may be a significant category.

There are many other extinction risks arguably too unlikely to be a priority, for instance various apparently safe or extremely unlikely astrophysical events, supervolcanoes, seemingly minor environmental harms, fertility loss, and voluntary human extinction (Bostrom 2002b).

### 1.4.5 Management of Specific Risks

Asteroid tracking has been an active attempt to avoid an existential risk, though smaller scale impacts presumably contribute much of motivation for this activity (Rees 2004, p.92). Avoiding a large nuclear war has received significant political attention over the years, in part due to the fear of destroying civilization and perhaps mankind (Posner 2004, p.88). Super-human level artificial intelligence, nanotechnology, and biotechnology are expected by some to be the biggest upcoming technological risks (Center for Responsible Nanotechnology 2008b; Yudkowsky 2008; Turchin 2008; Sandberg & Bostrom 2008), and several small private organizations exist to investigate them and make policy recommendations (Singularity Institute for Artificial Intelligence 2010; Center for Responsible Nanotechnology 2008a; The Foresight Institute 2010; The Lifeboat Foundation 2010). Most effort towards avoiding existential risk is a by-product of other aims, such as avoiding more likely smaller scale disaster scenarios.

## 1.4.6 Non-lethal barriers to visible expansion

There are some potential future filter steps that do not require large-scale human annihilation immediately. However, arguably none of these seem likely to contribute a lot of filter strength (Hanson 1998b). Also most of them indirectly human extinction eventually when Earth becomes uninhabitable.

### *1.4.6.1 Failure to thrive*

There are many ways humans could fail to colonize space without going extinct. We might face a non-lethal disaster and be unable to climb back to our current level of technology, perhaps due to resource depletion, or for social reasons (Bostrom 2002). Cultural precipitation of economic growth is not well understood (Cohen & Easterly 2009). A stable social equilibrium such as a powerful world government could prevent technological progress (Webb 2002, p.215). The technologies required to travel long distances in space may be much more expensive than they appear, or impossible.

### *1.4.6.2 Values*

Some suggest that most creatures may just not be interested in colonizing space when it becomes possible (Miller 2006). However expansion to use new resources has been a recurring value amongst life on Earth, presumably because those who practice it have access to vastly more resources (Hansson & Stuart 1990). A future shift in values so extreme that no capable person or group is interested in making use of resources available outside planet Earth is hard to imagine (Hanson 1998b). A lesser shift in values, where some groups would like to explore space but are prevented by a majority, could still prevent space colonization. Given how enormous the filter is, for indifference to colonization to be a large filter step, it would need to be an inevitable outcome of virtually all advanced civilizations, not just a plausible story about one. There would need to be a ubiquitous mechanism leading civilizations to change their values as well as a strong tendency for them to have effective global enforcement of this preference.

### *1.4.6.3 Invisibility*

Aliens may purposely leave our surrounds looking natural (the Zoo Hypothesis) (Ball 1973). Large scale colonization may look a lot like dead space, and colonization of our own solar system may leave it looking natural (Davies 2010). We may be misinterpreting what we see, and be quite wrong about what ‘natural’ would look like in these cases (Webb 2002; Hanson 1998b).

## 1.5 This thesis

This thesis argues that correct anthropic reasoning entails a greater risk of human extinction than usually appreciated. Chapter Two introduces anthropic reasoning. In Chapter Three we apply two popular anthropic reasoning principles and several variants or less popular principles to the Great Filter, in order to learn when along the path of development the unlocated filter strength is likely to be. We will see that each of the principles implies that more filter strength is likely in filter steps in our future than we naively estimate. Chapter Four discusses these results. In Chapter Five we will look at the plausibility of the various principles, and the potential for the best principle to be far from those we have seen. Chapter Six will elaborate the concrete implications of what was found for existential risks. Chapter Seven will summarize the discoveries of the previous four chapters.

If this thesis succeeds, we can be confident that we are underestimating the risk of human extinction, without resolving indexical reasoning. If we expect the correct indexical reasoning principle to be among the currently popular choices, we should increase our estimated risk of extinction immediately. If all of the principles we visit imply that human extinction is more likely than we thought, this is also relevant to the question of which principle is most likely to be correct, as we shall discuss later.

# Chapter 2: Anthropic Reasoning

## 2.1 Introduction

The central question of anthropic reasoning is how to treat *indexical information*, which is information about how things are relative to oneself, as opposed to how things are objectively. Your knowledge that you are alive is indexical information. 'God's coin toss' is a simple thought experiment which should illustrate the problem, taken from Olum (2000).

**God's coin toss**

> Suppose that God tosses a fair coin. If it comes up heads, he creates ten people, each in their own room. If tails, he creates one thousand people, each in their own room. The rooms are numbered 1-10 or 1-1000. The people cannot see or communicate with the other rooms. Suppose that you know all this, and you discover that you are in one of the first ten rooms. How should you reason that the coin fell?

One view is that on awakening (before learning the number of your room) you have learned nothing, since everyone would awaken in your situation under either hypothesis. Since the coin was fair, you should still be fifty percent sure that it fell heads. If that is so, upon learning you are in one of the first ten rooms you must update to be 99 percent sure the coin fell heads, since being among the first ten rooms was certain if the coin landed heads and had a one percent chance if the coin landed tails.

Another view is that on waking you should be 99 percent sure that the coin landed tails, since you had a much higher likelihood of being created if many people were created. After learning you are in one of the first ten rooms you should then be fifty percent confident of heads.

The correct answer is disputed. The two opinions above coincide with the answers given by the two main reasoning principles advanced to deal with indexical information. The first answer agrees with the Self Sampling Assumption (SSA) formulated by Nick Bostrom and used implicitly by others (Bostrom 2002a; Leslie 1992; Franceschi 2004; Leslie 1993; Lewis 2001), and the second with the Self Indication Assumption (SIA) which has been formulated, or used implicitly in some form, by several authors (Kopf et al. 1994; Dieks 1992; Olum 2002; Dieks 2007; Monton 2003; Hanson 1998a; Elga 2000). SSA and SIA will be discussed at greater length shortly.

## 2.2 Why indexical information must be treated differently

Bayes' theorem (Bayes 1958):

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \qquad (1)$$

That is, the probability of a hypothesis H being correct given evidence E is the probability of that evidence occurring if the hypothesis is true multiplied by the prior probability that the hypothesis is true divided by the unconditional probability of that evidence occurring. This is a standard way to update the probability of H when you observe evidence E. It is usually used with non-indexical information, which is information whose truth value is conditional only on which world the hypothesis describes, not your location in it. Can we use Bayes' theorem to update non-indexical hypotheses on indexical information? There are a number of possible ways to do this.

## 2.3 The naïve assumption

One suggestion is that when we obtain indexical information E, we should take p(E) to be the prior probability that E occurs to someone somewhere in the universe. For instance if I observe that I have brown eyes, the only updating I do is to exclude worlds where nobody has brown eyes. My observation does not change the relative probability that I am in a world where one person has brown eyes or everybody does.

$$P(H1|\text{I observe E}) = \frac{P(\text{anyone observes E}|H1)\,P(H1)}{P(\text{anyone observes E})} \qquad (2)$$

This is often called 'non-indexical' reasoning, because the user treats his indexical information as non-indexical. For instance 'I have brown eyes' is treated as 'someone in the world has brown eyes'. This kind of reasoning raises a big problem: it makes the scientific method nearly useless (Bostrom 2002c). To see this, imagine that you are conducting an experiment to discover whether the speed of light is $3x10^8$m/s (hypothesis 1) or $4x10^8$m/s (hypothesis 2). Your instruments record $3x10^8$m/s. Under hypothesis 1 this observation will be made at least once with near certainty. However since there is always a small chance of error, and you know such experiments are carried out many times in either world, hypothesis 2 also predicts with high likelihood that an observation of $3x10^8$m/s will be made. This is especially a problem if the universe is large, since then almost every observation can be expected to be made somewhere at some time (Knobe et al. 2006). If P(E|H) is equal to P(E|-H), you can infer nothing about hypothesis H by observing evidence E. Science becomes virtually impossible.

## 2.4 The Self Sampling Assumption

The Self Sampling Assumption (SSA) developed by Bostrom (2002a) solves the above problem (Bostrom 2002c). Intuitively, what matters in science is the assumption that whatever you observe is much more likely to be a common observation than an uncommon one. So instead of taking P(E|H) to be the probability that E is observed at least once by someone given H, when we do science we understand it as something like the expected proportion of people who observe E under hypothesis H. SSA formalizes that idea. Under SSA, P(E|H) is replaced by the proportion of your *reference class* who receive evidence E if hypothesis H is true. Your reference class is a set of people or creatures similar to you, which will be discussed further shortly. We will call the set of people you could possibly be, given your information, your *information set*. For instance in God's Coin Toss above, if your reference class were 'people in the experiment', under either hypothesis one hundred percent of them have your experience of waking up (and so are in your information set), so your posterior is the same as your prior: fifty percent to heads for a fair coin. When you learn that your number is between one and ten, under a heads hypothesis one hundred percent of your reference class shares your observations, while under tails only one percent do, so Bayes' theorem can be used to update strongly in favour of heads.

In full, SSA says that if Hj is the hypothesis that you are in World j, then $P'(H_j)$ , is the prior probability $P(H_j)$ multiplied by the number of observers World j contains in your information set $I(E_iH_j)$ divided by the number of observers it contains in your reference class, $RC(H_j)$.

$$P'(H_i) = \frac{P(H_i)\, I(H_i)}{RC(H_i)\, z} \qquad (3)$$

When you have used SSA once and wish to update on further information, the probability of a possible world is the SSA probability from the first piece of evidence multiplied by the proportion of the information set which you used for that calculation who experience the new information, normalized. This gives the same results as updating in the above SSA fashion on all of the information at once.

$$P(H_1|E_1E_2) = \frac{P(H_1|E_1)I(H_1E_1E_2)z}{I(H_1E_1)} \qquad (4)$$

Figure 1 demonstrates the use of SSA, including updating on further information.

The reference class introduced here presents problems, for both the application and the plausibility of SSA. There is no canonical procedure or grounds for selecting an appropriate reference class (Bostrom 2002a). What the reference class should refer to is an open question, so it is difficult to even estimate it. 'People' is often an intuitively appealing reference class, as are sometimes 'people who are part of the experiment' or 'people who previously did what I did' or 'conscious observers'. Unfortunately the intuitions which recommends these reference classes do not append any reasons, and mathematically they are arbitrary, except arguably the 'conscious observers' class. The reference class problem is not SSA's only problem, but we will elaborate on the others later. Bostrom defends this uncertainty as being similar to uncertainty over which prior to use, though he expects further work may narrow the range of possibilities (Bostrom 2002a, p.182).

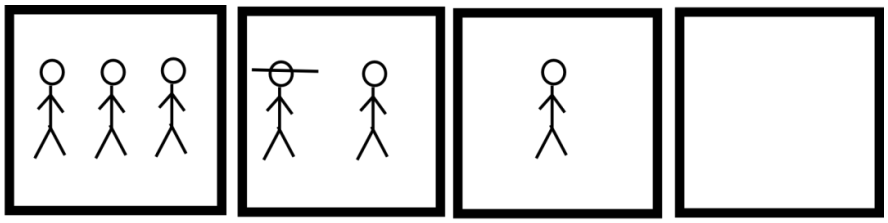
The Strong Self Sampling Assumption (SSSA) is a variant of SSA which takes time segments of observers called 'observer-moments' rather than whole temporally extended observers as its basic unit (Bostrom 2002a, p.162). By SSSA if you are a drunkard and can not remember how old you are, using the reference class of human observer-moments, you should be confident that you live to be old, because then a greater proportion of human observer-moments experience being you.

There is no requirement to keep the same reference class throughout when using SSSA. Each observer-moment is in a new indexical position, so they may have a new reference class (Bostrom 2002a, pp.168-172). If you continually use the narrowest reference class possible, always consisting of observer-moments with exactly your information, you will avoid scientific updating in a similar fashion to that of non-indexical updating discussed above. One hundred percent of your reference class always shares your observations under any hypothesis.

We will use 'SSA' to refer to both the temporally extended observer version and to SSSA, unless specified otherwise.

## 2.5 The Self Indication Assumption

The Self Indication Assumption (SIA) is the main proposed alternative reasoning principle to SSA. Variants of it have been explicitly suggested, implicitly used, or implied by the arguments of various authors e.g. (Dieks 2007; Kopf et al. 1994; Monton 2003; Hanson 1998a; Elga 2000; Olum 2002). SIA corresponds to the second set of answers to the God's Coin Toss thought experiment above.



| | | | | | |
|---|---|---|---|---|---|
| non-indexical prior | | 1/4 | 1/4 | 1/4 | 1/4 |
| SSA | per person | 1/12 1/12 1/12 | 1/8 1/8 | 1/4 | |
| | world total | 1/4 | 1/4 | 1/4 | 0 |
| | normalized | 1/3 | 1/3 | 1/3 | 0 |
| | world total on knowing you have no hat | 1/4 | 1/8 | 1/4 | 0 |
| | normalized world total on knowing you have no hat | 2/5 | 1/5 | 2/5 | 0 |
| SIA | per person | 1/4 1/4 1/4 | 1/4 1/4 | 1/4 | |
| | world total | 3/4 | 1/2 | 1/4 | 0 |
| | normalized | 1/2 | 1/3 | 1/6 | 0 |
| | world total on knowing you have no hat | 3/4 | 1/4 | 1/4 | 0 |
| | normalized world total on knowing you have no hat | 3/5 | 1/5 | 1/5 | 0 |

**Figure 1: An example of updating probabilities with SIA and SSA**

The four boxes are possible worlds, containing observers. They have equal prior chance of existing. The table underneath is divided into a procedure used under SSA and one under SIA, both assuming that you find yourself existing in one of these worlds. The top row under each principle shows the probability of being each person, if all you know is that you are a person (note, they are not normalized). The next row shows the total probability of being in each possible world. The next row shows these probabilities normalized. If you learn you are not wearing a hat, the next row shows the probabilities you should now give to the possible worlds. The (not yet normalized) probability of each possible world at this point is the sum of the probabilities of being each person who doesn't have a hat shown in the first row. The last row shows the probabilities in the second last row normalized.

SIA says that the probability of you being in possible World H is the prior probability of World H existing multiplied by the number of observers it contains in your information set, normalized. If $P(H_i)$ is the prior probability of $H_j$ occurring, $P^*(H_i)$ is the probability after using SIA, $I(H_i)$ is the cardinality of your information set under $H_j$, and z is a normalization constant,

$$P^*(H_i) = P(H_i).I(H_i).z \qquad (5)$$

For instance in God's coin toss, initially the tails world contains one hundred times as many observers in your information set as the heads world does, and both worlds are equally likely based on other information. This means SIA says the tails hypothesis is one hundred times more likely. After learning your number is between one and ten, the number of people in your situation is equal under each hypothesis, so both hypotheses are equally likely.

In general, further information can be taken into account in the same fashion as in SSA, taking P(E|H) to be the proportion of your previous information set who experiences the new information. This means that probabilities given by SIA are always equal to those given by SSA multiplied by the number of members the SSA reference class used contains, and normalized.

For a demonstration of how SIA is used, see Figure 1 above.

## 2.6 Full Non-indexical Conditioning

Full Non-indexical Conditioning (FNC) is a variation on the naïve assumption above, invented by Radford Neal (2006). It avoids the problem with science to some extent, so it is included here as arguably the most plausible incarnation of non-indexical reasoning, for comparison to the anthropic principles. The strange and concerning implications of SSA and SIA to be explored later may tempt one to abandon anthropic reasoning in favor of non-indexical reasoning, treating one's own observation as evidence only that such an observation was made somewhere. FNC is here to demonstrate that non-indexical reasoning has as many strange implications as the anthropic principles, if not more.

FNC differs from the naïve assumption in that to use it one is meant to update on every piece of information one has ever experienced, including that which seems irrelevant. Suppose under hypothesis A the world contains one person and under hypothesis B the world contains two people, and these hypotheses have no more detail. If you use the naive assumption using only your apparently relevant observation that you are a person, you cannot update. However if you update on all of your memories, almost twice the proportion of the possible two-person

worlds will have at least one person in them in your information set. This means you can update in favour of being in the two person world.

FNC allows science to work as long as the world is small enough that your memories are likely to be unique. Intuitively, updating on irrelevant information makes science more possible because it is highly unlikely that any given world contains a person with your exact memories, and more likely that those memories would be conjoined with a common experimental observation than an uncommon one. Unlike the naïve assumption, FNC gives no answer if the universe is large enough, or you are forgetful, enough for your memories to be likely to be repeated. FNC does not stop giving answers if there is a tiny chance of more than one person like you, so in practice it diverges slightly from SIA as populations increase, because SIA makes that tiny possibility of a world with two people with your experiences twice as likely as FNC does.

FNC is a weak claim. It agrees with both of the previous principles (and any plausible principle) that if there is nobody like you in a possible world, that world is not the one you are in. It agrees with SIA, and with SSA using the narrowest reference class, that if there is one person like you in a given possible world, that world has the same odds as before. If there is likely to be more than one such person, it gives no answer.

## *2.7 Applications*

SIA, FNC and SSA in many variations are the three principles we will apply to the Great Filter in Chapter Three. Before we do that, we will see how they apply to two other problems in anthropic reasoning, the Doomsday Argument and the Sleeping Beauty Problem. This will demonstrate the appeal of each of the principles, as well as provide useful background for later discussion.

### 2.7.1 The Doomsday Argument

Anthropic reasoning has made several contributions to existential risk research, e.g. (Tegmark & Bostrom 2005; Sandberg et al. 2010; Rees et al. 2008, pp.106-120). The best known and most contentious of these is the Doomsday Argument. The Doomsday Argument is an application of SSA with a broad reference class, such as ‘humans’ (Bostrom 2002a, p.89). Slightly different versions were created independently by Richard Gott and Brandon Carter, whose idea was developed by John Leslie (Gott 1993; Leslie 1990).

**The Doomsday Argument, summarized by Bostrom and Cirkovic** (2003, pp.83-84)**:**

> Your birth rank (i.e. your position in the sequence of all humans) is roughly 70 billion. That you should have such a low birth rank is less surprising, and more probable, if the total number of humans that will ever have existed is, say, 200 billion rather than, say, 200 trillion. (By 'probability' we here mean rational subjective credence.) Given these unequal conditional probabilities, one can derive from Bayes' theorem that the probability of impending doom goes up after conditionalizing on your birth rank. That is, after realizing the full evidential import of you having a relatively low birth rank, such as 70 billion, you should increase your probability estimate of hypotheses according to which there will be relatively few extra humans (such as 200 billion in total) at the expense of hypotheses according to which there will be very many more humans (such as a total of 200 trillion).

Discussion of the Doomsday Argument has made up a large part of the anthropic reasoning literature; many attempts have been made to disprove it and most of them have been rebutted e.g. (Dieks 1992; Eckhardt 1993; Leslie 1993; Eckhardt 1997; Korb & Oliver 1998; Bostrom 1999; Chambers 2001; Sowers 2002; T. Adams 2007; Sober 2003; Olum 2002).

One of the most popular counter-argument to the Doomsday Argument is that the Self Indication Assumption should be used rather than SSA (e.g. Dieks 2007; Olum 2002; Monton 2003). When SIA is applied in the Doomsday Argument scenario, the probabilities of given survival times do not change from the priors. FNC agrees with SIA here; as long as you know your birth rank the probability of there being a person with your exact memories does not vary with the total number of people.

### 2.7.2 Sleeping Beauty

The Sleeping Beauty Problem (Elga 2000) attracted a detailed discussion on anthropic reasoning, largely detached from the doomsday discussion. The two popular positions on it correspond to SIA and SSSA with a narrow reference class, though it was not identified as the same issue as the doomsday argument until later (Dieks 2007).

**The Sleeping Beauty Problem as described by Elga** (2000, p.143)**:**

> Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you back to sleep with a drug that makes you forget that waking.

> When you are first awakened, to what degree ought you believe that the outcome of the coin toss is heads?

Elga argued for what has become known as the 'thirder' position. His argument was that if you were to be told it was the first waking, you should believe there was a fifty percent chance of another awakening in your future, so P(H|first awakening) = P(T|first awakening). Also P(first awakening|T) = P(second awakening|T), by a principle of indifference. This makes all of the awakenings equally likely, implying that tails is half as likely as heads. This corresponds with the SIA position.

The 'halfer' position was forwarded by David Lewis (Lewis 2001). It says that since Sleeping Beauty should believe the coin had a fifty percent chance of heads before the experiment, and she did not learn anything new upon waking (she always knew she would wake), she should not update her credence from fifty percent. This corresponds with the SSA position if Sleeping Beauty's reference class is her wakings in the experiment. Originally when Sleeping Beauty awakes, her information set makes up one hundred percent of her reference class in either world, so under SSA she does not update her credences. If she learns that it is her first waking (and maintains the same reference class), her information set becomes only half of the reference class under tails, so heads becomes twice as likely as tails.

The FNC answer to Sleeping Beauty is very similar to the SIA answer (Neal 2006). Suppose Sleeping Beauty pays attention to all the irrelevant evidence in her surroundings, such as whether she wakes up facing left or right and the location of any insects in the room. There was a very low chance of that particular set of experiences occurring on any waking, and this chance is twice as high under the tails hypothesis because it has two waking.

We will revisit the debate over the merits of the various principles in Chapter Five.

# Chapter 3: The Filter according to the Principles

## 3.1 Introduction

This chapter will demonstrate the effects of accepting different principles on predictions about future filter steps. The principles to be applied are SIA, SSA using a variety of reference classes, and FNC. SIA and SSA are included because they are the most widely used abstract principles and correspond to most of the popular anthropic arguments. FNC is obscure and arguably implausible, but it is the strongest form of non indexical reasoning, and will demonstrate that 'non-indexical' reasoning does not avoid our conclusions here.

## 3.2 The approach taken: assumptions and explanations

There may be quite different paths from dead matter to galactic colonization, however we will only discuss the one we are on. Similarly, the steps may not always be in the same order. This does not significantly change the analysis; if there are paths other than ours with a significant chance of passing the entire filter, this does not make the filter on our own path any weaker. If that path had a weak filter, then it must be very rare, or the total filter would not be very strong, and if it is rare, that translates to a large step at getting onto it where it branches. All paths must have at least the filter strength we see.

It may seem presumptuous to think the same chances of passing particular steps prevail in different times and places. For example, you could explain the Fermi Paradox by a space phenomena we are unaware of preventing life emerging anywhere up until very recently, and our planet being among the first of many. This does not change the analysis however: if we do not know which places or times are special, such things can be treated as part of the generic difficulty of the step.

We shall from here on treat the minimum total filter strength as a single value, though in fact there is just such a sharp decline in the probability of our observations with decreasing total filter size that this simplification makes little difference.

We will use arbitrary numbers of supposedly known steps in each case. This is not because we should know how many hard steps there are: the unlocated filter strength may be in known or unknown steps. However, for the current problem it should not make much difference how we divide the path into steps, within reason. If it is suspected that two real steps are grouped together, the prior over the strength of the 'step' should reflect the possibility of it including a second step.

## 3.3 SIA

We will now look at how to apply SIA to the Great Filter, where possible worlds are combinations of strength at different filter steps. We will assume you know your own stage in

the filter. We will not account for any further indexical information, such as your species or address, as it is irrelevant. Recall that SIA weighs the odds of possible worlds in proportion to the number of members of your information set they contain.

To apply SIA we multiply the prior probability of each combination of filter strengths (a possible world) by the population at our own stage in the filter in that world, then normalize. The population at our stage is the population per star at our stage multiplied by the total number of stars, multiplied by the proportion of them that reach our stage. This means filters in our past directly decrease the total population at our stage, by decreasing the number of stars with a population at our stage. The expected population per star at our stage and the total number of stars can be cancelled out when comparing the probabilities of possible worlds with different sets of filters, as those values are the same regardless of filter strengths. So the effect of SIA is to multiply the prior probability of a combination of filters by the chance of getting to our stage with that set of filters. In the following equation, P'(x) is the probability of x after using SIA, $f_1$ is the probability of passing filter 1, k is the filter immediately before our stage, and z is a normalizing constant.

$$P'(f_1=a, f_2=b, ...f_n=y) = P(f_1=a, f_2=b, ...f_n=y).a.b.c...k.z \quad (6)$$

This is illustrated in box 1 below.

**Box 1**

Suppose you have two hypotheses which you believe to be equally likely: in hypothesis 1 (H1) all the filter steps before us led to a $10^{-7}$ decrease in the number of stars reaching our stage, in hypothesis 2 (H2) there is a decrease by $10^{-8}$.

$$P'(H1) = P(f1=10^{-7}).10^{-7}/( P(f1=10^{-7}).10^{-7} + P(f1=10^{-8}).10^{-8})$$

$$=0.5*10^{-7} / (0.5 . 10^{-7} + 0.5 . 10^{-8}))$$

$$=\sim 91\%$$

There is a strong shift toward the hypothesis containing the smaller filter. H2 has a filter ten times weaker, so it is weighted by a factor of ten. The odds move from 1:1 to 10:1. If both hypotheses have the same total filter, this means a tenfold increase in the odds of a ten times stronger late filter.

Since the SIA effect depends only on the total filter strength before our stage, all possible worlds where the steps before ours multiply to the same total unlikelihood of reaching our stage get the same weighting under SIA. See Figure 2.

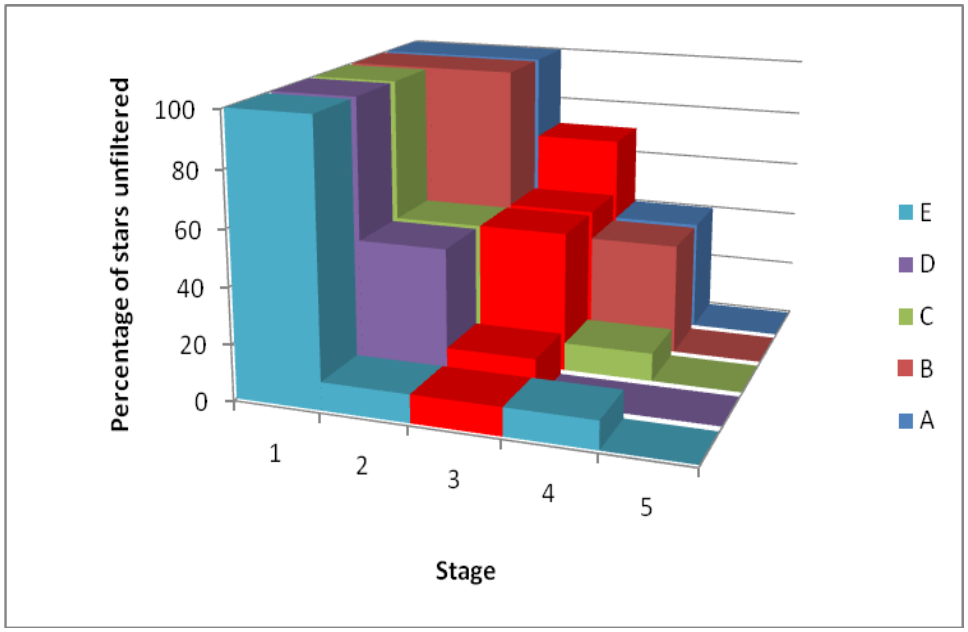



**Figure 2: Only total past filter strength matters under SIA**

1-5 are stages between filters. Five combinations of filter strengths are shown as layers A-E. The only feature of these possible worlds that affects their weighting under SIA is the height of the bright red bar in each world, which is our stage. For instance Worlds B and C will get the same weighting.

Because there is a minimum total filter strength, the effect of preferring relatively small filters in the past for the whole sequence of steps is to prefer larger filters in the future. This can be seen in Figure 2: A, B and C have larger present populations and also much stronger future filter steps than D or E.

Now we will see an example of how this effect changes the net expected strength of the filters in our future when we have a more plausible prior over a large number of filter strength combinations, rather than an arbitrary five. We use the set of filter steps, and prior over ranges of possible filter strengths shown in table 1. The results are shown in Figure 3.

| | | Range of strengths Min | Max |
|---|---|---|---|
| | Star | | |
| 1 | Life | 1 | 20 |
| 2 | Intelligence | 0 | 9 |
| 3 | Technology | 0 | 4 |
| 4 | Avoid risks | 0 | 4 |
| 5 | Visible | 0 | 4 |

**Table 1: The prior over strengths of individual steps for the example in Figure 3**

For simplicity we use a discrete prior over orders of magnitude for the filter strengths, uniform across the ranges shown. For instance our prior here over the strength of the third step, developing technology is 20 percent chance of zero orders of magnitude (virtually inevitable), 20 percent of one order of magnitude etc. Note that the prior shown here is before updating on the minimum total filter.

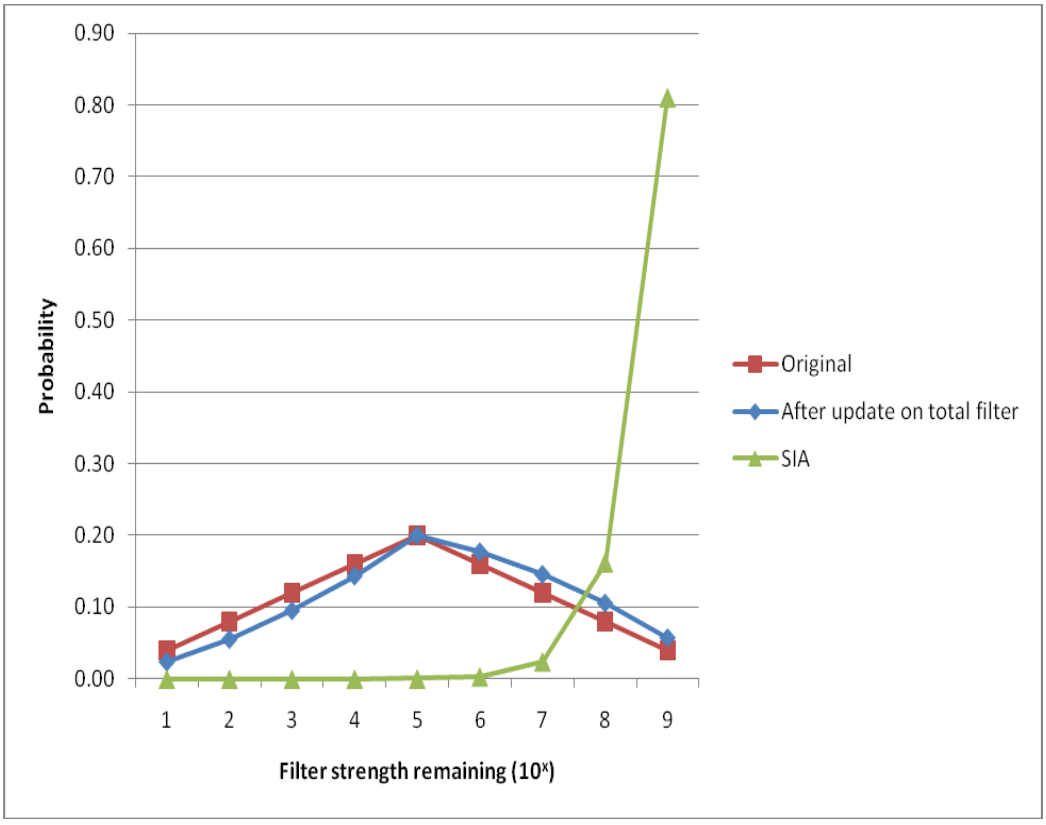



**Figure 3: Example of SIA shift in expected total future filter**

This graph shows probability distributions over the total filter strength in steps we are yet to pass, using the priors given in table 1. The red line shows distribution given by the priors alone. The blue line is the distribution after updating on the total filter being greater than $10^{-22}$. The green line is after also applying SIA, assuming we are at stage three.

## 3.4 SSA

Recall that SSA weights the odds of possible worlds by the proportion of one's reference class who share one's information set. For now we will use reference classes consisting of a set of stages in the filter, since the model of the filter contains no more detail about the inhabitants of a possible world than their filter stage. For the same reason, we will also use our own stage in the filter as the information set. We will visit other reference classes and information sets later.

To apply SSA to the Great Filter, we multiply the prior probability of each combination of filter strengths (a possible world) by the population at our own stage in that world (our information set), then divide it by the total population of all stages within our reference class combined, and normalize. The population at our own stage is the number of stars multiplied by the total strength of the filters before our stage multiplied by the average population on stars which make it to this stage. As with SIA, the number of stars can be cancelled from the equation, but with SSA the expected population per star at our stage remains relevant because its ratio to the populations at other stages plays a part. In the following equation, $\ddot{P}(x)$ is the probability of x after using SSA, '$f_x$' is the $x^{th}$ filter step, 'c' is the average population of creatures at our stage per solar system which reaches our stage, z is the normalization constant, s is the total number of steps, and N(x) is the average population of creatures at stage x per planet which reaches that stage.

$$\ddot{P}(f_1 = a, f_2 = b, \dots f_n = y) = \frac{P(f_1 = a, f_2 = b \dots f_n = y)(a.\, b.\, \dots .y)c.\, z}{\sum_{x=1}^{s} N(x) \prod_{m=1}^{x} f_m} \quad (7)$$

This is illustrated in Figure 4.

More strength in filters before our stage (but still within our reference class) decreases the total populations of both the reference class and the information set. However, such filters will always decrease the total information set population by a greater proportion, because the stages in the reference class before the extra filter are not diminished at all. See Figure 5. This means larger filters in our past always make the worlds containing them less likely under SSA, assuming our reference class stretches back far enough to contain the filter steps. Because we have a minimum total filter, this decreased probability of worlds with bigger early filters adds to the effect of increasing the probability of worlds with bigger late filters.

More filter strength after our stage, but still between stages within our reference class, decreases the total population of observers in our reference class without decreasing the total population in our information set. See Figure 6. This makes such worlds more likely, since the information set then makes up a larger proportion of the reference class. So all things equal, larger filters in our future are more likely than we would otherwise think, assuming our reference class stretches forward far enough to contain them.

The overall effect of SSA on possible worlds with a single large filter step is shown in Figure 7.

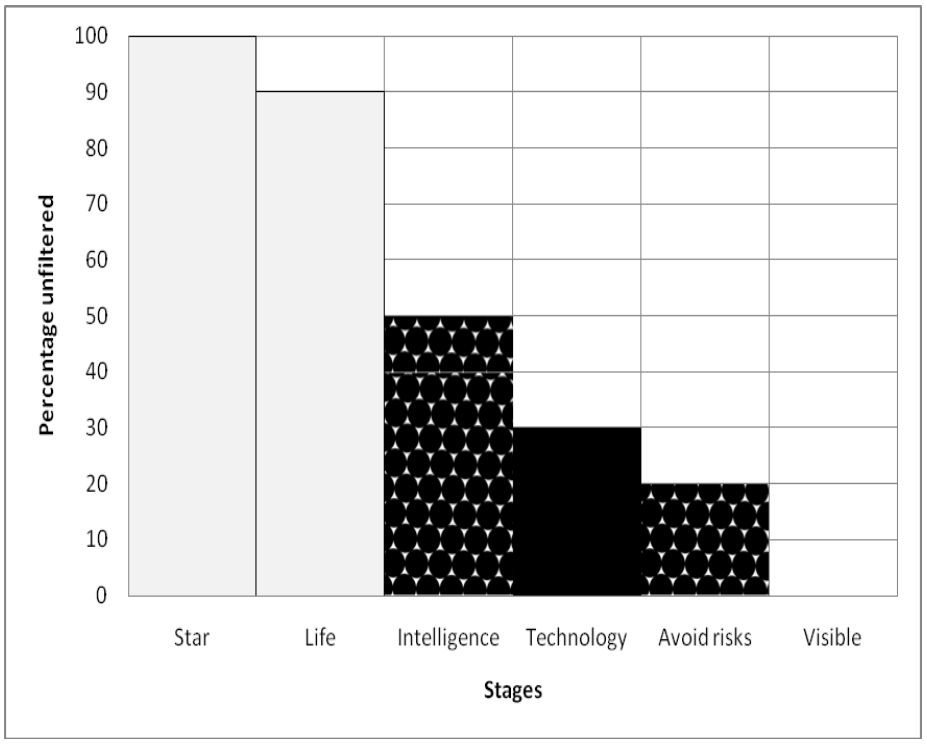



**Figure 4: How the SSA shift is determined**

This figure shows one possible combination of filter strengths, to illustrate how SSA calculates the probability of a possible world. The bars correspond to stages. The height of each bar is the proportion of solar systems which reach that stage. The width of the bar corresponds to the population at that stage per planet reaching the stage. The black bar is our information set, which together with the patterned bars comprise our reference class. The probability of this world under SSA is the black area divided by the reference class area, multiplied by the prior probability and normalized.

As Figure 7 shows, filters closer to us in the past or future have a greater effect than those in the distant past or future. A filter step just before our stage leaves as many as possible earlier stages at their full population while still diminishing our information set, while a much earlier filter step diminishes the populations of a lot of stages by the same proportion as our own. A filter step just after us reduces many more future populations than a very late step of the same size.

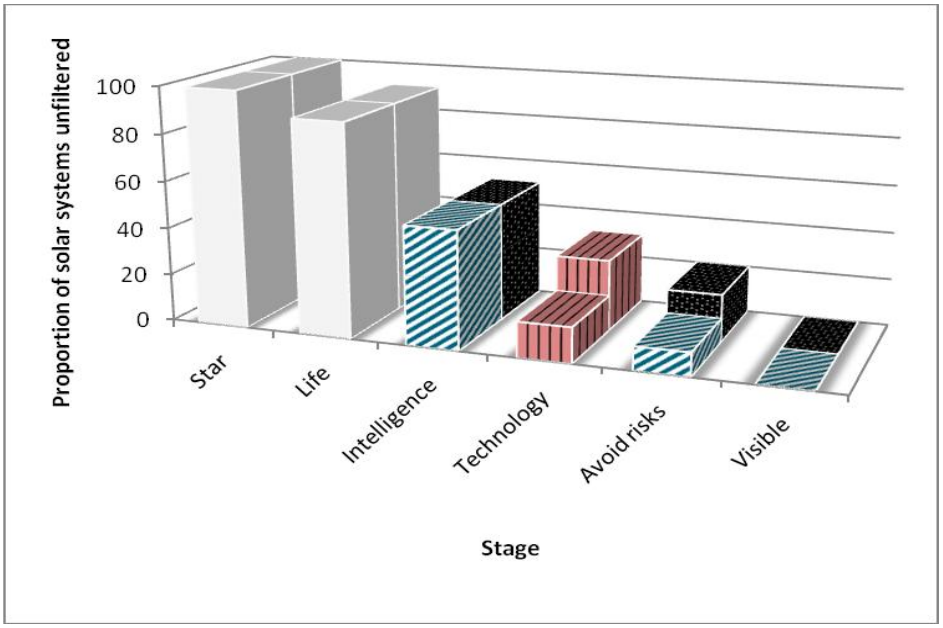



**Figure 5: Past filters within the reference class**

Here we see the combinations of filter strengths in two possible worlds. Our stage is shown in pink vertical stripes in each. Stages shown in grey are outside the reference class. The only difference between the worlds is that the green striped world has a larger filter before our stage, within the reference class. This is to illustrate that such a filter reduces the proportion of solar systems surviving at all later stages, but leaves some earlier stages at their high levels, so making the population at our stage a smaller proportion of the total population at all stages.

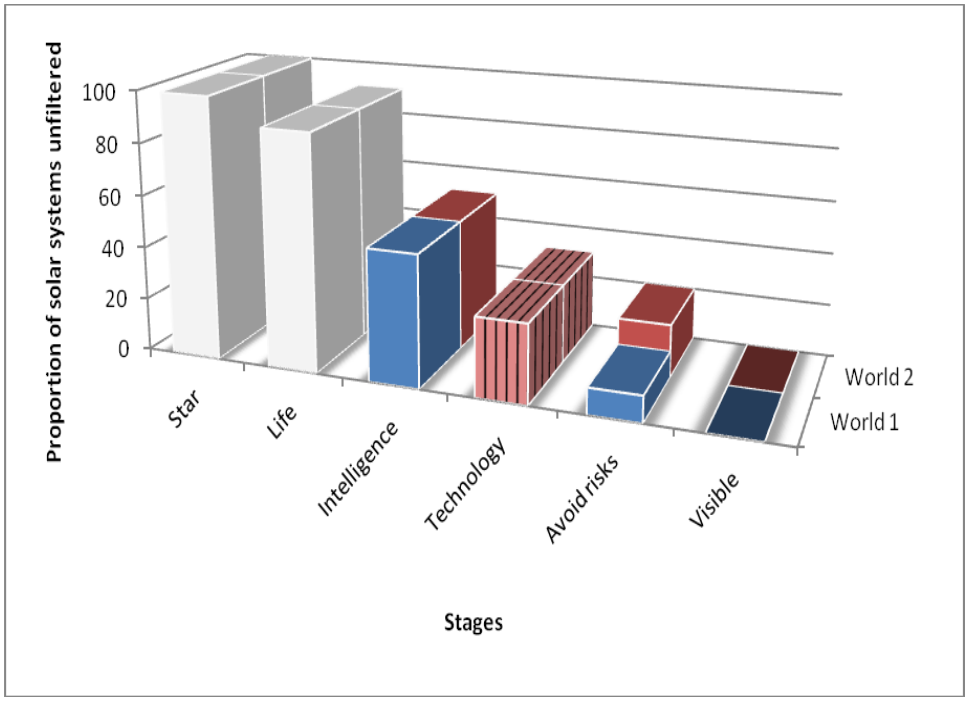



**Figure 6: Future filters within the reference class**

This is exactly the same as figure 5, except that now the world in the foreground has a larger filter than the other world in the future instead of the past, still within the reference class. This future filter only reduces the populations of other stages in the reference class, not our own. This means such a filter always makes our stage a larger proportion of the total population. Thus SSA favours worlds with such filters.

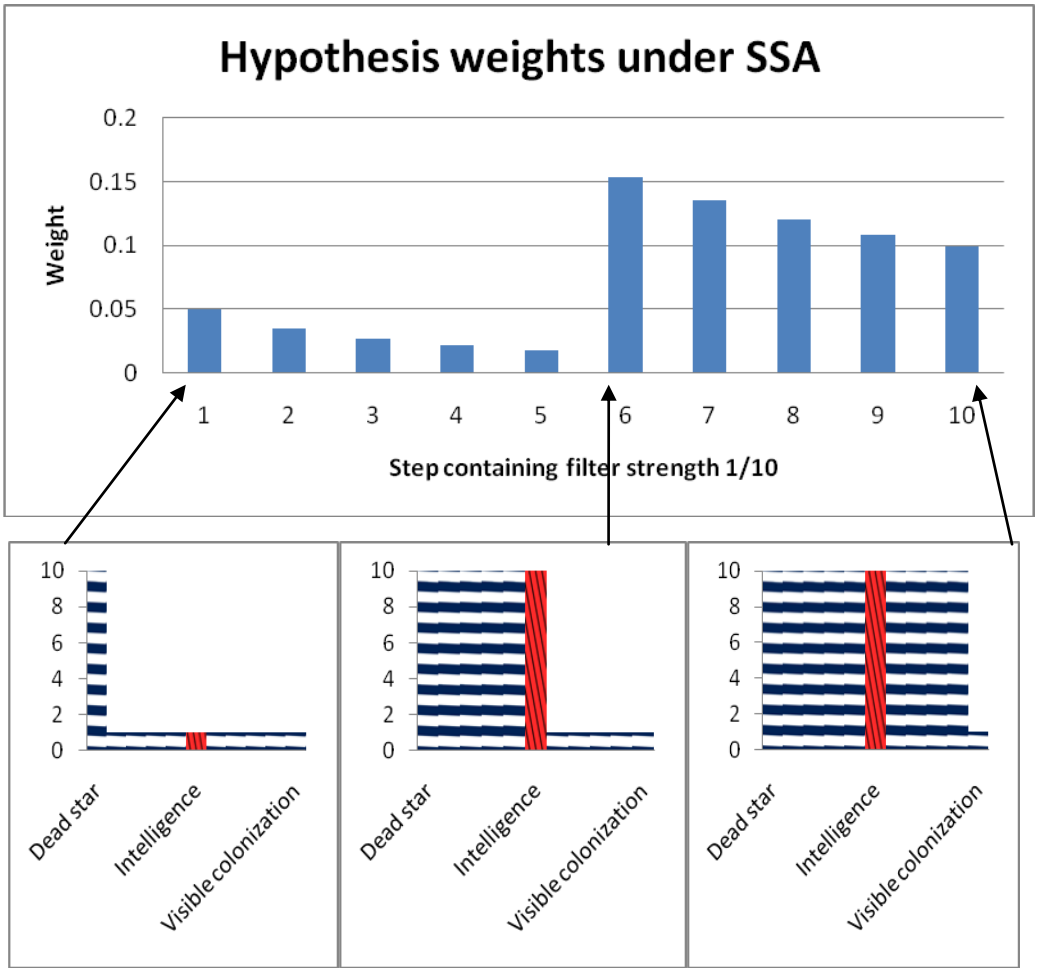



**Figure 7:How the position of a single step affects the probability of worlds under SSA**
The top graph here shows the SSA probability shift for possible worlds which contain only one filter step, but in different locations. The horizontal axis shows which step contains the filter (the other steps have zero strength). The filter allows a 1/10 chance of passing it. The vertical axis shows the factor by which the prior probability of that hypothesis is multiplied under SSA, if our stage is after filter step 5, our reference class includes all stages, and we assume for simplicity the same population per star at any stage.

The figures below illustrate the shape of the filter in three of the possible worlds featured in the top graph. They each show the proportion of unfiltered solar systems vertically, and the stages horizontally. Our stage is shown in red. The red bar as a proportion of all of the bars corresponds to the height weight given to that world in the above graph.

### 3.4.1 Population

We have ignored the complication of population size per solar system of creatures at various stages. The total population at each stage is the product of the number of stars under consideration, the proportion of those stars that reach that stage, and the population per star at that stage. We can ignore the number of stars because it is the same all stages, and cancels out. We have looked at the effect of the strengths of different filter steps, which determine the proportion of stars which reach each stage. Here we will look at the population per star that reaches a stage. This population influences the effect of filters. A larger population at a stage other than our own increases the effect of filters between that stage and our stage on the probability of the world in question. Changes in the population at a relatively sparsely populated stage will have little effect on the proportion our stage makes up of the total population. Changes of the same proportion in a large population will have a much larger effect. The effect of population is illustrated in Figure 8.

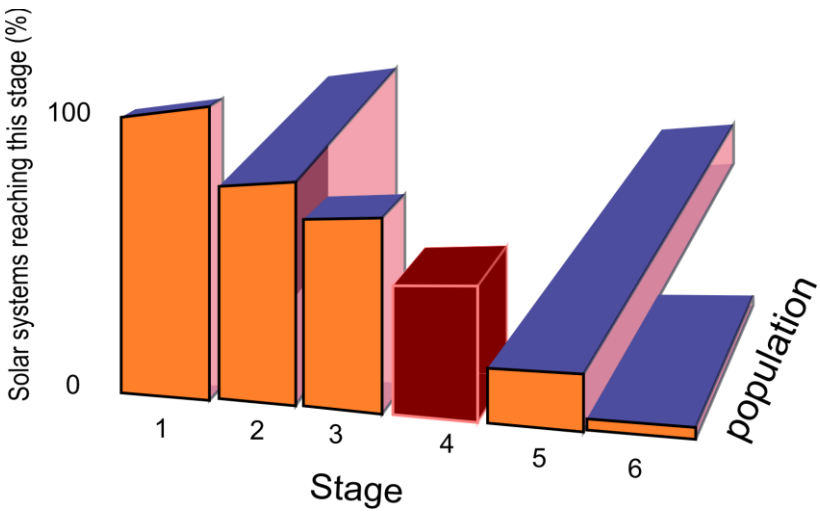



**Figure 8: Population**

The front face of this diagram is similar to that in Figure 4: it shows stages chronologically from left to right, with height signifying the proportion of stars that reach that stage. Filters are the drops in probability between the stages. The depth of each stage is the population at that stage per star, conditional on the stage being reached.

This figure illustrates that filters which change the relative size of larger populations have more of an effect on probability than filters which change the relative size of smaller populations. All stages are in the reference class, and stage four is our own stage. The probability of this world under SSA is proportional to the side area (population per planet*number of planets) of block 4 divided by the side areas of all blocks. A larger proportion of solar systems reaching a stage other than ours adds to the denominator in proportion to the size of the population at that stage. For instance if stage 5 were a little higher, it would increase the size of the denominator much more than if stage 3 were a little higher. This would decrease the probability of the world in question accordingly. For similar reasons, an increase in filter size will have more of an effect on stages with less total filter before them already.

## 3.4.2 Reference classes and information sets

In the above we have assumed a broad reference class. If we narrow the reference class to a set of filter stages which is still broader than our own, we see a similar pattern to that in the widest reference class case shown in Figure 7, but only across stages within the chosen reference class. This is illustrated in Figure 9. Filters that aren't between stages in that reference class have no effect on the weighting of a possible world. Figure 10 demonstrates this.

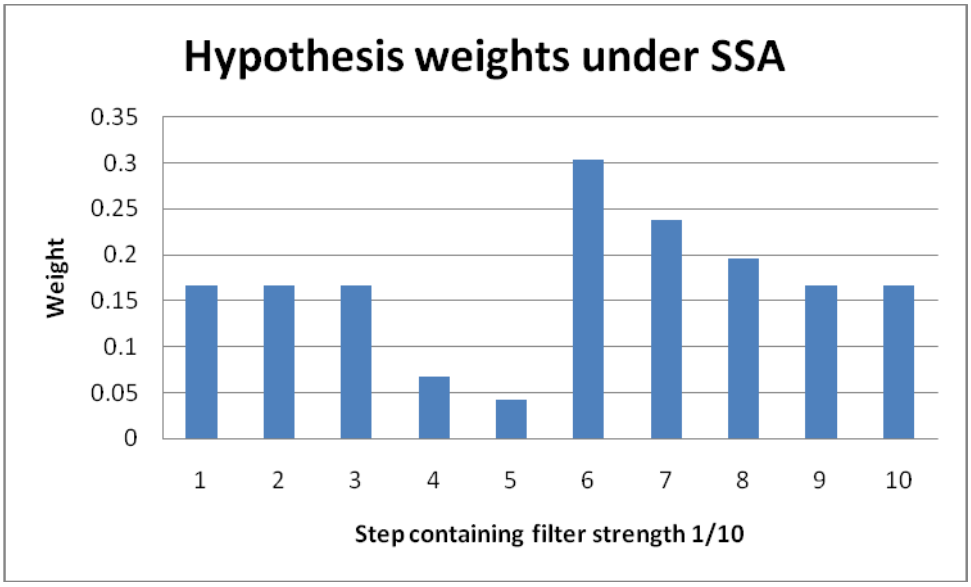



**Figure 9: How reference class choice influences the effect of filter steps on SSA probability weighting**

This is the same as the top graph in figure 7 above, but with a reference class encompassing only stages 4-8. It shows the weightings SSA gives to a set of worlds which are identical except for having a single filter; the horizontal axis shows the stage before the filter step. The population per star at each stage is assumed to be the same.

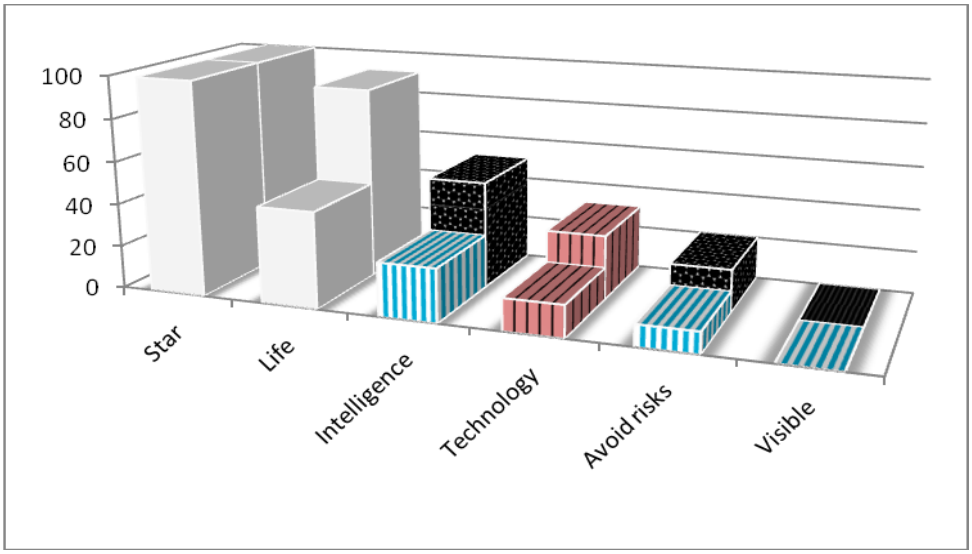



**Figure 10: Filters outside of the reference class make no difference**

This is similar to Figures 5 and 6, but compares a world with a larger filter before any stages that include reference class members (foreground) to one without (background). Our information set (pink stripes) comprises the same proportion of the reference class (patterned) in both worlds. Both worlds have the same probability because all bars are smaller by the same proportion in the foreground world.

This means for instance that possible worlds may have very large filters in our past, as long as they are before any stages in our reference class, and those worlds will not be diminished in probability under SSA, unlike those with such filters within our reference class.

You may use a more specific reference class or information set than a set of stages you are part of. For instance your reference class could make up some proportion of our stage, or of several stages. This will generally give similar results to inclusion of the whole stage in the reference class. We will not consider further possible reference classes here.

#### *3.4.2.1 Narrowest reference class*

If you use a reference class which only includes your information set, the information set makes up 100 percent of the reference class under every hypothesis, so SSA has no effect except to exclude worlds which contain nobody in your reference class. If the information you are taking into account about yourself is only your stage in the filter, as long as there is any chance of a solar system reaching our stage there is no effect. However if your reference class takes into account more detailed, apparently irrelevant information, such as 'I am from a species with advanced technology, and I have a pet Iguana', we see an effect similar to that of SIA. This is because of the finely detailed possible worlds with many creatures at our stage rather than few, many more of those with a large number of observers at our stage will have at least one creature at our stage with some arbitrary unlikely characteristic, such as a pet Iguana.

### 3.4.3 Overall result

We will now show the effect of SSA on the same simple prior as we used to exemplify the effect of SIA above (shown again in table 2). The results are shown in Figure 11. SSA pushes the distribution far toward the larger future filters, but not as far as SIA.

| | | | Total strength | |
|---|---|---|---|---|
| | | Population | Min | Max |
| | Star | 0 | | |
| 1 | Life | 100 | 1 | 20 |
| 2 | Intelligence | 1 | 0 | 9 |
| 3 | Technology | 10 | 0 | 4 |
| 4 | Avoid risks | 100 | 0 | 4 |
| 5 | Visible | 1000000 | 0 | 4 |

**Table 2: Priors and populations for the example in Figure 11**

These are the prior probabilities and populations used in the example shown in Figure 11. They are reproduced from Table 1 above, except we new include population per star at a given stage, as that is required to use SSA. For simplicity we use a discrete prior over orders of magnitude for the filter strengths, uniform across the ranges shown. For instance our prior here over the strength of the third step, developing technology is 1/5 to 0 orders of magnitude (virtually inevitable), 1/5 to 1 order of magnitude etc. Note that the prior shown is before updating on the minimum total filter.

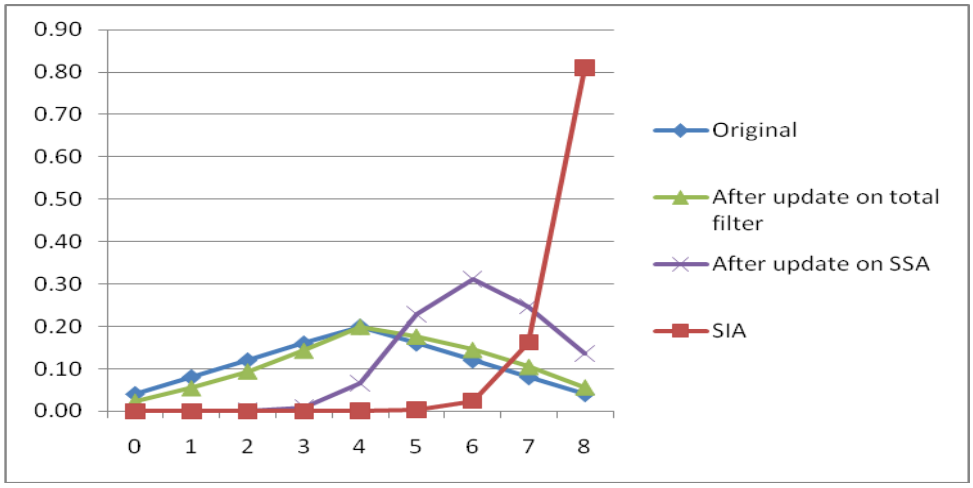



**Figure 11: Example of SIA shift in expected total future filter**

This is figure adds the SSA distribution to figure 3. Each line is a probability distribution over the total filter strength in steps we are yet to pass, using the priors over the strength of each step given in table 2. The blue line shows the distribution given by the priors in Table 2 alone. The green line is the distribution after updating on the total filter being stronger than $10^{-22}$. The purple and red lines are the distributions after also applying SSA and SIA respectively, if we are at stage three, and our reference class includes creatures at stages 2-5, with the populations at each stage shown in table 2.

## 3.5 FNC

When we apply FNC to the Great Filter, the result is very similar to that of SIA, given some assumptions. The universe must be small enough that any particular experience is unlikely to occur more than once, and you must have a large amount of apparently irrelevant indexical information with which to define a highly specific information set. The reasoning here is identical to that elaborated above for SSA with a highly specific information set and the narrowest possible reference class. That is, FNC does nothing more than exclude worlds where nobody has your experiences, but larger numbers of people at our stage make a single occurrence of your highly specific experiences more likely. When the likelihood of your experiences is very small, the increase in probability from an increase in population is almost proportional to the increase in population, so the result is very similar to SIA. FNC calls for as narrow an information set as possible, which brings the effect as close as possible to that of SIA. Radford Neal used FNC in this fashion to conclude that the human race is less likely to colonize space than we would otherwise think (Neal 2006).

## 3.6 Summary

This chapter has demonstrated that SIA always implies filters in our future are likely to be stronger than we otherwise estimate, as long as we have any uncertainty about their strength. FNC has very similar results, which we discussed but did not see in detail. SSA often agrees with SIA that future filters will be stronger than we have thought, though it has more complicated results due to the freedom in reference class choice and influence of population per solar system. Population per solar system isn't relevant to the SIA result because it can be cancelled out between possible worlds. We saw different aspects of SSA's effect individually. The details of the probability shifts induced by the principles were different. The next chapter will discuss the similarities mentioned here, examine the differences, and further explain the relationships between the effects we saw above.

# Chapter 4: Discussion

## 4.1 Future filters are more likely

In the last chapter we saw that SIA, FNC and SSA using many reference classes agree that we should expect larger filter steps in our future than we imagined. Since these principles agree, we can be confident in this result even though the principles are contentious, as long as we are confident that one of them is correct.

### 4.1.1 The implications for human survival

Since not all possible future filter steps endanger the human species, it is possible to increase our credence in larger later filters without anticipating human extinction. However, the extent to which the shift leads to particular later filters being more likely depends on your prior probability distribution over their strengths. If you have any uncertainty about the strength of the many potential filter steps that are future extinction risks, any of the principles we discussed will favour possible worlds where those filters are stronger, making those risks more likely. So we can extend the above findings to conclude that almost any future filter step involving human extinction is more likely than we think.

## 4.2 The SIA result

According to SIA, our future is far more likely to contain large filters than we naively think (or hope). This is because the Fermi Paradox implies that small filters in our past require large filters in our future, and under SIA smaller filters in our past get a boost in probability. This is because smaller filters in our past mean more solar systems reach our stage, so there are more observers at our stage, making us more likely to exist. Since we do exist, SIA favors the hypothesis which predicted that with higher credence.

The shift under SIA can be very large. Even if we thought the probability of half of the filter being in our future was one in a billion, and all the rest of our distribution was on future filters with a total of one order of magnitude or less strength, after using SIA we would be fairly confident in the large future filter.

The SIA result is very similar to the results of FNC and SSA using a highly specific information set and the narrowest reference class, as we discussed earlier, so those principles not be discussed individually here. Further details of the SIA result will be discussed below, in comparison to the SSA result.

## 4.3 The SSA result

SSA has more varied conclusions than SIA on the Great Filter due to its dependence on the extra variables of reference class and population per star at every stage. SSA agrees with SIA that future filters within our reference class are more likely than we would otherwise think.

Unlike SIA, SSA favors future filters even when there is no minimum total filter. SSA also makes smaller filters before us more likely, so long as they are within our reference class. This also makes future filters in general more likely, along with past filters before our reference class. It is possible to choose reference classes that avoid any shift toward expecting later filters, but reference classes broad enough to include members outside our stage generally do have such an effect.

## 4.4 The Doomsday Argument rises again

The SSA effect for broad enough reference classes is partially equivalent to the standard doomsday argument. Our current stage in the filter is analogous to our birth rank, so SSA reduces the likelihood of future stages and future generations alike.

The filter scenario differs from the standard Doomsday Argument in an important respect, however. In the Doomsday Argument we know our birth rank, so there is one person in our current situation. The filter leaves us quite unsure of the number of observers at our stage – in fact this is a key question we are trying to answer. This uncertainty about the population of our information set creates two differences.

First, it means the doomsday effect can extend backwards, making past populations small relative to ours. SSA favours hypotheses where observers at past stages are rare relative to observers at our stage. This argues for weaker filters between reference class members at past stages and our stage, since strong filters would leave our stage less populous relative to theirs.

Second, uncertainty about the population of our stage means that the Doomsday Argument for future people works on an added dimension. Assume that you know when you live, but nothing else relevant. SSA favors worlds where the ratio of future people to current people is small. SSA is indifferent about whether the ratio is relatively small because the world in question contains more people now or fewer people in the future. The Doomsday Argument used a scenario where the current population is fixed across worlds, so the entire doomsday effect came from favoring smaller numbers of future people. In the filter scenario both current and future populations are variable. This means worlds with relatively large current populations are also favored. This is what makes the difference between SIA being able to undermine the Doomsday Argument perfectly in the original case, and making even stronger predictions of doom in the filter case.

As always the SIA conclusion is equal to the probability distribution found by SSA weighted by the total number of people in the SSA reference class. In the Doomsday Argument, where the current population is constant, the SSA probability for a world is 1/(reference class population), so SIA brings the probability of every world back to its original prior. In an opposite case where the future population is held constant and the current population varies across possible worlds, those worlds with a greater number of current people also have both a greater proportion of current people and more people in total, so the SIA shift will be toward the

same worlds as the SSA shift, strengthening rather than countering it. The Great Filter case is a combination of these situations: the past, current, and future populations are all unknown. This is why SIA cannot remove the doomsday effect in the Great Filter case, and can instead predict future doom more strongly.

Whether the SIA conclusion predicts a future filter more weakly or strongly than the SSA conclusion depends on the populations at different stage. SIA is indifferent to populations per star, while it is possible to make the SSA effect arbitrarily strong by increasing the population of beings at a later stage within your reference class relative to the population at your stage.

## *4.5 Differences in the details*

Beyond similar implications of impending human extinction, the effects of the principles differ both in structure and in the detail of their conclusions. Here we shall visit these differences.

### 4.5.1 Timing

Between locations in steps we have passed SSA prefers earlier filters to later ones, and the same for possible locations in our future. This is visible in Figure 7. This is unsurprising; the same pattern is seen for the future in the standard Doomsday Argument, and a later filter in the past means a greater number of larger populations at the intervening steps. On the other hand SIA is indifferent to the timing of filter steps beyond their location in our past or future, as shown in Figure 2.

### 4.5.2 Size of error

Under SIA, the difference in final probability between a world with a filter in our past and a world with the same size filter in our future is equal to the strength of the filter. According to SSA, the difference is between zero (if the early filter were before any stages containing reference class members and the late one after any) and less than the size of the filter (if the filters were just before and after our stage). This means that SIA will always claim that future filters are more likely by a larger factor than SSA. Figure 11 exemplifies this; while SSA shifts the curve substantially toward larger filters in our future, SIA shifts it so far that 80 percent of the probability is on the maximum future filter strength possible in the model.

### 4.5.3 Certainty

SSA's specific conclusions can vary wildly with the choice of reference class, and even be escaped completely. In contrast, SIA depends on no unknown variables, so given any prior over the strengths of filter steps it can give a precise update. This update will reliably include larger future filters if there is any uncertainty about their strength in the prior.

### 4.5.4 Reference class dependent possibilities

In the above we have usually used reference classes and information sets comprising of a set of stages. Reference classes defined by more apparently extraneous information, such as our number of limbs, can be accounted for by reducing the expected population per planet reaching a stage. We saw in section 3.4.2.1 that reference classes narrower than a single stage approach SIA in their results. If the reference class contains only some creatures at a given level, possible worlds where those creatures tend to change into or be replaced by other creatures outside the reference class will be equivalent to those where they are destroyed.

### 4.5.5 The importance of the past

If we take into account the time at which we live, SIA implies directly only that there are many people at our level at this time. However without any reason to think our time is special prior to knowing we exist in it, most possible worlds where there are many people at our stage now also have many people at our stage in the past. Thus we can infer if there are many people at our stage now, there were probably many people at our stage in the past. Fermi's paradox tells us that they tend to meet a filter, and from that we may infer that we will probably meet a filter too. This information does not increase our risk of demise from causes which could not be a filter any more than seeing more people enter a forest than leave should increase your expectation of the sun exploding while you are in the forest, or of you having a heart attack there. If we discover evidence to suggest that we are in a relevantly different situation to most of the stars in our past light cone then, the filter may be much reduced under SIA. It is of course very unlikely that we will discover this. The SSA effect on the other hand does not rely on inferring your prospects from past experiences of others, as evidenced by the Doomsday Argument. It decreases the proportion of future people by any means possible. If we learned we were in a different situation to the rest of the universe somehow, SSA would continue to predict our doom, given suitable reference class choices.

### 4.5.6 Population interactions

We have treated the populations of creatures per solar system at each stage as independent of the chances of passing the steps before and after that stage, but in fact the strength of filters is likely to influence the population. This is only relevant to SSA. For instance, if a step is easy to pass, the presence of creatures that pass it may cause the population of those at previous stages to be reduced, by replacing them. This need not be a large effect, for instance the population of animals on earth has not obviously decreased since humans have taken power. This means SSA may sometimes equally prefer a world where a filter is successfully passed and one where it is not, but the previous population tends to continue for longer instead.

### 4.5.7 Filters and other catastrophes

SSA increases the likelihood of all hypotheses where future members of our reference class are destroyed, not only those where we are destroyed by a filter. Non filter catastrophes include those that destroy a large portion of the universe at once or are specific to a small area or our own situation for instance.

## *4.6 Conclusion*

Despite similarities in the key finding that extinction is more likely than we would otherwise think, it is important to know which anthropic principle is correct. The principles predict disasters at different times, and an imminent disaster demands more concern than a disaster at some time in the next few centuries. The size and reliability of the effect is important for reassessing how much of a priority dealing with existential risks should be. Knowing whether we should be addressing all disasters, or only those which may be filter steps, makes a substantial difference. The concrete implications of these details will be discussed in Chapter Six, after we address the plausibility of the anthropic principles the arguments rely on.

# Chapter 5: The Anthropic Principles

We have seen that SIA, FNC, and generally SSA, forecast a higher probability of doom for the human race than we would otherwise expect, but that their predictions differ significantly in the details. These facts raise the question of which, if any, of these principles we should accept. This is the question of this chapter.

## 5.1 Why these principles?

So far we have been discussing SIA and SSA as the main alternative anthropic principles. In this chapter we shall ignore FNC, since it has few advocates and was included mainly for comparison.

A *centered world* consists of a possible world and a 'center', the part of that world where the observer in question is located. An anthropic principle needs only to be a function which takes a prior over possible worlds, and derives a prior over centered worlds. There are infinitely many such functions, so why do we focus on these two?

A big reason indexical reasoning is difficult is that there is a one-to-many relationship between possible worlds and the centered possible worlds corresponding to them. This means that either the probability of being person X is proportional to the probability of that person existing, or the probability of being in World Y is proportional to the probability of World Y existing, or neither of these things is true. Both of these seem intuitively correct, but they can't be true at once, as is illustrated in figure 12. If the probability of being person X is proportional to them existing, SIA is true. If the probability of being in World Y is proportional to World Y existing, SSA is true. This is one reason SIA and SSA are prominent principles.

SIA and SSA also embody the two obvious answers to each of these questions: 'Is it the number or the proportion of people like you in a world that makes you more likely to be in that world?', 'Does your existence alone, before learning any details of what you are, give you evidence?', 'Was it necessary that you would exist if anyone did?'.

## 5.2 The case for confusion

On the other hand, both principles have some quite counterintuitive consequences. We will overview some known concerns with SSA and SIA in turn before considering some new arguments.

### 5.2.1 Problems with SSA

Olum lists many unintuitive aspects of SSA (2000). He notes the asymmetry between people implied by SSA; if you consider yourself a random member of each possible world, then some other people must not exist at all in the smaller worlds. Relatedly, there is an incongruity in thinking you are as likely to be in a world of any population size, except for zero. SSA

depends on the content of regions which are causally disconnected from the event in question. Lastly, Olum points out that SSA's dependence on apparently unrelated observers means that SSA will often give different answers if an event occurs once and if it is repeated.

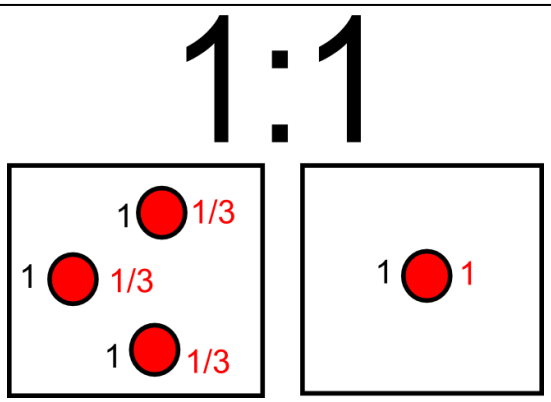



**Figure 12: SIA and SSA**

The two boxes represent possible worlds. They contain different numbers of observers, and so correspond to different numbers of centered worlds. Both worlds have the same prior probability of existing. The smaller numbers beside the centers represent the probability of being that center under two methods of assigning probabilities. The red numbers show the probability of being in each center if the probability of being in a world is equal to probability of that world existing (and if the total probability of a world is evenly divided between centers in that world). The black numbers show the probability of each centered world if the probability of being that person is equal to the probability of that person existing. Evidently, both of these cannot be true at once. The red figures correspond to SSA and the black figures to SIA.

SSA allows for backward causation and incredible predictive powers. For instance, if upon hearing that gamma rays were likely to hit Earth and kill many people (but not all), a strong central government could protect the world by committing to embarking on aggressive population growth policies if and only if the rays hit. Even if the rays' path was already determined, everyone could be confident that there would be more people in their reference class but outside their information set if the rays hit, decreasing the likelihood of that hypothesis under SSA. This is Nick Bostrom's UN++ thought experiment (2001). There are a variety of similar thought experiments (Bostrom 2002a, pp.141-158; Bostrom 2001). These unintuitive outcomes all rely on certain reference classes, for instance in the above a broad enough reference class to include humans born in future.

### 5.2.1.1 Reference class problems

Bostrom suggests avoiding the above problems by using a reference class which does not cause them (2002a, p.171). This wide choice of reference classes may solve any specific unintuitive consequences, but it leads to strange theoretical consequences.

The reference class is defined by a set of information about yourself, which you will not update on later. If you use information A to define your reference class, then update on information B, you will tend to get a different result than if you use information B as your reference class and update on information A. For instance if I am a human who plays the drums, and I want to guess how many drummers there are in the world, if I start with the reference class 'humans' then update on being a drummer, I come to think there are a lot of drummers. If I start with the reference class 'drummers' then update on being a human, I hardly move at all from my initial expectation. This is despite the information being theoretically equivalent – there are not different 'kinds' of information for these purposes. Bostrom suggests disallowing arbitrary differences from defining a reference class (2002a, p.181), but any distinction between arbitrary and non arbitrary information appears unjustified.

The choice of reference class is potentially so open that a huge range of probability distributions can be chosen. There is no requirement to maintain the same reference class throughout the calculations when using SSSA, since at different moments you are at different indexical positions (e.g. Bostrom 2007). This gives the SSSA user the freedom to apparently ignore Bayesian updating in many situations, as we shall investigate later. The reference class concept treats information differently according to how well you feel like it could have applied to you, though there is no theoretical difference between the senses in which you could have been me and you could have been a road. Treating them differently means that structurally identical situations viewed from the same perspective lend themselves to different reference class choices, and so different probability distributions, though the true outcome will be the same and equivalent information is known.

### 5.2.1.2 Continuity with uncontroversial situations

Another unintuitive feature of SSA is that it agrees with SIA and common sense in saying that if the possible worlds under consideration are all in existence somewhere, one is more likely to be in the more populous one if all else is equal, but comes to an entirely different conclusion if one learns that only one of these worlds exists (Finkelstein 2008; Bostrom 2002a, pp.194-198). For instance suppose a group of one hundred people were divided into a group of ten and a group of ninety, assigned the letters A and B, but left ignorant about whether they were in the large or the small group. You are told you are in group B. SSA says this means group B is likely to be the large one. Then everyone in the other group is killed. SSA still says group B was likely to be the larger one, if you use a reference class of people in the experiment. However suppose instead foetuses were divided into the groups before they were born, then one

group randomly chosen for abortions before anyone was conscious. When you are born, and old enough to think about the question, SSA says you should think yourself equally likely to be in either world.

This difference means that if the many worlds interpretation of quantum mechanics is true, SSA agrees with SIA on questions concerning events where worlds would have split, rather than probability due to ignorance (Olum 2002).

Existing is treated differently to other evidence by SSA. If some of a group of people will receive a piece of evidence in one possible world and not in another, not receiving that evidence is incontrovertibly reason to update in favour of the world where nobody received it. One kind of evidence a person could receive is instantaneous death. It would seem not coming into being originally is a quite similar piece of evidence. However, it follows if one treats this in the same as any other evidence that one should generally think oneself more likely to be in a larger world if one finds oneself existing, and so accept SIA, or a principle like it (Armstrong 2009). SSA implies that somewhere in the continuum between being killed and not being born there is a point where the probability calculation comes out entirely differently. For instance, if you were born a minute before a disaster destroyed one of two cities, killing all inhabitants, you should assume you live in what was the larger one, given an appropriate reference class. If you were born a minute after, it is equally likely that yours was the smaller one.

### 5.2.2 Problems with SIA

SIA can also seemingly permit backward causation and incredible predictions (Shulman 2010). To do this one must commit to creating identical replicas of one's mind having exactly the present experience, if you discover in future that the desired event has occurred. That way a greater population has your current experience in the possible world you prefer, so you can be more confident that you are in that world (this is close to the method suggested by Elga to defeat Dr Evil (2004)). Notice that conditional on your successfully carrying out this plan, rather than being a deluded mind emulation, the probability of the event is as before, however you cease to be able to tell whether you are a deluded mind emulation or not.

SIA may also seem to depend on the nature of causally disconnected observers, similar to SSA. This is because it increases the chance of a universe with a large population of observers causally disconnected from you. However SIA only does this when you may belong to either group of observers, so while one of them is causally disconnected from you in reality, they all have a chance of containing you. If you know a given observer *in the hypothesis* was not you, even if they were only an inch away from you, they would be irrelevant under SIA.

The main counterintuitive results of SIA are the Presumptuous Philosopher (Bostrom 2002a, p.124), and variations on it:

**The Presumptuous Philosopher**

> It is the year 2100 and physicists have narrowed down the search for a theory of everything to only two remaining plausible candidate theories, $T_1$ and $T_2$ (using considerations from super-duper symmetry). According to $T_1$ the world is very, very big but finite, and there are a total of a trillion, trillion observers in the cosmos. According to $T_2$, 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 to you that $T_2$ is about a trillion times more likely to be true than $T_1$ (whereupon the philosopher […] appeals to SIA)!"

More counterintuitive versions of this exist. Olum points to a hypothetical crank who claims every planet exists on a huge number of 'other planes' (2002). Under SIA no matter how unlikely this hypothesis begins, it can posit a number of planes high enough that it should become more likely than the alternatives after applying SIA, unless ones prior for large numbers of planes shrinks fast enough.

If we accept SIA we must also accept an infinite universe if it is possible that we are in one. If one originally has any finite credence in the universe being infinitely large and populous, SIA will weight this theory infinitely relative to any finite theories. This resembles the Presumptuous Philosopher, but is applicable in humanity's current circumstances.

#### *5.2.2.1 Infinity*

This brings us to one last problem with using SIA. If SIA is correct, then we should be almost sure that we are in an infinite universe, populated by infinitely many observers having similar experiences to us, assuming before using SIA we have some non-infinitesimal prior on such an infinite universe. If we are certain we are in this infinite universe, it's not clear whether SIA has any further conclusions to make (Chalmers 2010). For instance in Sleeping Beauty, there are not more waking under tails than under heads – there are infinitely many waking under either, with the same prior history, across the universe. This is a problem for any probabilistic reasoning principle; in SSA it is not possible to determine the proportion of observers who have a characteristic, and FNC makes no predictions at all in even a large world, let alone an infinite one. Even principles that are uncontroversial have problems here. For instance it is uncontroversial that if your class is divided into one person and one hundred people, you are more likely to be amongst the hundred. In an infinite universe, even this is hard to justify. This problem is just more pressing for SIA because SIA implies that we are almost certainly in an infinite universe.

## 5.3 What can be relied upon?

The above issues are all counterintuitive; however, none conclusively demonstrates either principle to be wrong. A recurring problem in the anthropic reasoning literature has been that it's not clear which principles or intuitions can be relied upon. As Lewis points out (Lewis 2001), Elga's original Sleeping Beauty argument relies on the Principal Principle, which is known to have limits, whereas Lewis' response relies on the promising but unproven principle that only new relevant evidence should change credences. Which of these principles should be dismissed here isn't obvious without evidence from something more reliable (which Lewis attempts to provide but is rebutted by Bostrom (2007)). As a result of this uncertainty, many arguments seem to do little more than beg the question, relying on the same weak intuitions that recommend the principle they purport to prove, and are happily rejected by the supporter of the opposing argument. For instance White shows the thirder position is wrong, assuming that if you reduce the chance of Sleeping Beauty awakening on each 'waking', her actually waking becomes more evidence that tails came up, which relies on something like the naïve assumption, which is already disregarded by thirders (White 2006). Jenkins believes Elga (2000) is doing something similar (2005). There are many arguments from analogy to situations where the answer is clearer, but those who disagree with the reasoning used simply dispute the analogy. For instance Bovens points out the analogy between Sleeping Beauty and the thought experiment of Judy Benjamin (Bovens 2010), which suggests the thirder answer to Sleeping Beauty, but the analogy only holds if the required symmetry exists between parts of worlds and different worlds, which SSA claims there is not. Finkelstein shows that the thirder position must be correct, assuming your credences aren't affected by the existence of uninvolved people (2008). However, SSA proponents accept these people do make a difference if they are within your reference class (Bostrom 2002a, pp.196-198). Franceschi (2004) makes another argument like this, and Eckhardt (1997) accuses Leslie (1997)of it.

Two possible ways to move forward in this situation are to find cases where these principles contravene truths we are more certain of, or to find cases where unintuitive results held against one principle would also condemn the other, and so the intuition can't count against either principle, relative to the other. The self indication conclusion outlined in Chapter Three potentially provides an example of the latter.

## 5.4 Is self indication in the filter more plausible than doomsday?

The conclusion of SIA on the filter is similar to the Doomsday argument. They both infer from a tiny amount of evidence, of a kind experienced by everyone in the relevant situation, that we have greatly underestimated the risk of human extinction. The Doomsday Argument has been seen as evidence against the reasoning that supports it, SSA, and in favor of the reasoning

that undermines it, SIA (Olum 2002; Dieks 2007; Dieks 1992). If SIA creates a similar enough conclusion, it is no better than SSA in this regard. Also the Doomsday Argument is more likely to be correct in that case, which has further implications for managing existential risks.

Whether the SIA conclusion is similar enough to the doomsday argument to have these results depends on what features of Doomsday make it unintuitive. We shall not attempt to judge the relevant intuitions further here.

While this argument may be evidence against SIA, it would leave SIA and SSA on equal footing on such issues. Now we will look at another case where an argument against one principle can equally ensnare the other.

## 5.5 The Unpresumptuous Philosopher is as extreme as her colleague

Consider a variation on the Sleeping Beauty scenario where there are one million waking on tails and one on heads. The probability you initially put on heads is determined by the reasoning principle you use, but the probability shift when you learn that this is the first awakening is the same either way. You will have to shift your odds by a million to one toward heads. Bostrom points out that either before or after this shift, you will have to be extremely certain one way or the other, and that such extreme certainty in either position seems intuitively unjustified (Bostrom 2007). However, he points out that the only alternative to this certainty is to keep credences near fifty percent both before and after receiving the evidence, apparently giving up Bayesian conditionalization. The latter is what Bostrom proposes, as a 'hybrid model' of Sleeping Beauty (2007), though he argues that this does not violate Bayesian conditionalization. We will look more at this claim in the next subsection.

The thirder in 'Extreme Sleeping Beauty' is analogous to the Presumptuous Philosopher. Both are considering two possible worlds containing people they could be, one of them far more populated than the other. Both believe themselves to be very likely in the populated world. The Unpresumptuous Philosopher then shall be analogous to the halfer. When the Unpresumptuous Philosopher learns there are a trillion times as many observers in T2 she remains cautiously unmoved. However, when the physicists later tell her exactly where in the cosmos our planet is under both theories, the Unpresumptuous Philosopher becomes virtually certain that the sparsely populated T1 is correct while the Presumptuous Philosopher hops back on the fence. In the infinite case, the Unpresumptuous Philosopher believes with probability one that we are in a finite world if she knows her location is within any finite region. For instance if she knows the age of the universe she is *certain* that it will not continue for infinitely long. Whether these SSA-based views are as unintuitive as the SIA-based ones depends on what feature of them is unintuitive. If it is the extreme certainty in the face of very limited evidence, they match. So it seems the Presumptuous Philosopher thought experiment is no reason to prefer SSA with

consistent updating over SIA. However, there is still an alternative form of SSA that may be preferred: the Perpetually Fence-Sitting Philosopher who follows Bostrom's preferred solution to Extreme Sleeping Beauty: the hybrid model.

### 5.5.1 Both philosophers are more conservative than their alternative

The hybrid model uses SSSA with the narrowest possible reference class both before and after receiving evidence (different reference classes in each case), which means one hundred percent of the reference class shares the same experience in both cases, so Sleeping Beauty stays with the fifty percent chance given by the coin (Bostrom 2007). Bostrom argues that this does not contravene Bayesian conditionalization because Sleeping Beauty is in different indexical positions when she has the conflicting beliefs, so her observer-moments need not agree.

However, these arguments prove too much without further constraints. Everyone can count themselves in a different indexical situation at every moment, and Bostrom does not say what distinguishes cases where this inconsistent updating is warranted.

If observer-moments reason inconsistently like this, it does not merely mean they come to different conclusions about their own locations, but that they believe their past self was wrong. This is despite their past self using the same reasoning principle, and neither being in a privileged position. For instance, suppose before Sleeping Beauty knows what day it is she assigns 50 percent probability to heads having landed. Suppose she then learns that it is Monday, and still believes she has a 50 percent chance of heads. She also knows that her past self's beliefs imply that conditional on her past self being followed by a Sleeping Beauty observer-moment who knew that it was Monday, there was a 2/3 chance of heads having come up. She also knows that while her index has changed, she is in the same objective world as her past self. Yet she must knowingly disagree with her past self about how likely their world is to be one where heads landed, even while they both know every bit of information the other used, and agree on the reasoning principle.

If the case for Bayesian conditioning without updating on evidence is unpersuasive, we have a choice of disregarding Bayesian conditioning, or disregarding the Presumptuous Philosopher thought experiment as an argument against principles that come to such conclusions. We have greater reason to trust Bayesian conditioning in ordinary situations than to trust our direct intuitions about whether the Presumptuous Philosopher or his unpresumptuous friend are taking positions too extreme. Consequently the Presumptuous Philosopher thought experiment is not strong evidence against anything. Even if it was, we saw that it can equally count against SSA. Even principles we haven't considered would not avoid

analogous results, since they will need to make the same extreme update regardless of the starting prior.

## 5.6 Summary

We have seen that SIA and SSA stand out as the main contending anthropic principles. Both have some implausible results and worthy arguments against them, but none so strong as to conclusively settle debate. One way to assess the strength of arguments in battles of weak intuitions is to see whether the same argument can be made against all sides equally. If so, that argument can be dismissed as a means to adjudicate between them. The SIA result on the Great Filter may share enough features with the Doomsday Argument to put SIA and SSA on equal footing in this way, on that issue. This depends on which intuitions it is that recoil at the Doomsday Argument. We saw as a side note that a similar move can be made with the Presumptuous Philosopher argument against SIA. SSA, and in fact any principle which appears to heed Bayesian conditionalization in the normal way, comes to very similar conclusions. Again the similarity depends on exactly why the Presumptuous Philosopher seemed absurd to begin with. Arguably the main argument for SIA is that it avoids the Doomsday Argument, and the main argument against it is the Presumptuous Philosopher thought experiment. If the arguments of this chapter hold, SIA has less to recommend or condemn it than thought.

# Chapter 6: Implications

Here we apply the abstract discoveries from the past three chapters to the existential risks and other potential filters we met in the introduction, and consider policy implications. We have seen that SSA and SIA tend to predict later filters, but they also influence the type of filters we should fear. We will discuss the implications of SIA and SSA, though the SIA outcome is the most important given the findings of Chapter Five. In cases where both principles agree, there is more reason to heed their predictions.

Of the possible risks and obstacles to human expansion, some are more likely to be filter steps than others, and some are more likely under one principle or another. Here we will look at what is required to be a filter step, and some of the details of the anthropic predictions visited in Chapter Five, and how they influence the types of disasters and obstacles we should expect.

## *6.1 Filters vs. non-filter barriers*

### 6.1.1 Big disasters aren't filters

Some barriers would be expected to occur at a continual low frequency, or regularly to creatures at a certain stage of development. Others would be correlated so that creatures at all stars meet doom at the same time. This for instance includes vacuum decay and our being in a simulation which is shut down. These cannot be filter steps, since our information about the filter is from seeing that other stars have so far failed to pass it in our vicinity. Under SSA, possible worlds where all future observers are destroyed can become much more likely, given a broad reference class and assuming we have some information such as our time, to differentiate ourselves from the future people who may be destroyed. This would be true with or without a known filter. So if SSA is correct we should worry about all these kinds of disasters more than we are, while if SIA is correct we should worry only about those which may be filters.

### 6.1.2 Aliens might filter us, but then the filter is found

Destruction by extraterrestrials is an existential risk; however, it is conditional on capable aliens existing, and if capable aliens exist we are done: there is already a strong enough filter between colonizing space and being visible to explain the Fermi Paradox, so little need to expect further filters.

## *6.2 Destruction and change*

Another disaster that may seem to fall into the above category of 'too large to be a filter' is artificial intelligence explosion; however, one large enough to destroy life across large distances would not count as a filter for a different reason. If we give rise to AI that colonizes space, our star has not been filtered. Generally, the destruction of humanity is not a filter if we are replaced by something else with spacefaring potential. This means we should also not worry a lot about

uploaded minds causing a dangerous intelligence explosion or triggering any other extremely destructive transition.

Under SSA our reference class may be narrower than the stage we are at, and there are such cases where a change in the type of creature living will cause the population to pass outside the reference class. SSA weights such scenarios in the same way as if the creatures were destroyed at that point. A future where the universe is filled with robots is given similar weight by SSA with a reference class that does not include robots, as a future where the universe is dead.

## 6.3 Convergence

The future of the human race is much more detailed to us than the future of an arbitrary planet. What seems like a potential barrier to our own development can only be a filter step if it is a problem most civilizations would face. It is very likely that most civilizations meet resource shortages and pollute their environment, so if these things cause problems powerful enough to destroy us they are very feasible filter steps. Continual conflict and increasingly powerful technology are basic enough that threats could converge; however, if we perceive a great threat from a particular geopolitical tension, that cannot be the filter we await unless it can be seen as a manifestation of a general trend that should affect other extraterrestrial civilizations. Most of the above mentioned disasters are general enough that we should expect many civilizations to face the same risks. A change in values to the extent that nobody with the power wants to colonize space is easier to see as an outcome on one planet than a necessary outcome across beings of all kinds. Long term failure to thrive and destruction of humanity by smaller problems after an initial non-lethal disaster both require a combination of factors: the original disaster and the later inability to recover. For either of these to be a strong filter both factors would need to be very likely across civilizations.

### 6.3.1 Priorities

As we saw in Chapter Four, SSA can predict destruction of humanity via catastrophes that are not filters as well as those that are. If the future is otherwise expected to be long and full of members of our reference class, SSA can become highly certain that we are doomed. SIA does not increase the probability of non-filter existential risks. This should make averting filters less of a priority if SSA is correct than if SIA is correct. Some effort should be redirected to other existential risks, and there is less to gain from averting filters if you are confident humanity will not survive long anyway.

## 6.4 Timing

A big difference between the predictions of SIA and SSA is that SSA weights sooner disasters more than later ones, while SIA is indifferent about time. We should keep in mind that

both make all future disasters more likely to some degree. We may expect the greatest risks from nanotechnology and biotechnology to be near the start of their proliferation, and for them to become better controlled over time. Unforeseen risks are more likely farther in the future. Not being widely visible after colonizing other stars is clearly a very late filter.

Humans will likely gain more understanding of AI and physics experiments with time which should make them safer insofar as their risks are mostly from accidents. However, a later AI explosion is more likely to be fast, which makes it more dangerous (Sandberg & Shulman 2010).

Natural events such as non-anthropogenic climate change and astrophysical events are similarly likely to be early as late over long periods. The long term trends in likelihood of potentially doom-inducing warfare and ecological destruction are unclear. So SSA says all other things equal we should be most concerned about new technologies and risky science, less concerned about natural events, unsure about ecological destruction and war, and not expect invisibility to be a large part of barrier. In practice even under SIA it generally makes sense to focus more on early risks for other reasons. We are less likely to meet late risks, especially if early ones aren't managed, and there is more time to deal with them.

## 6.5 Population interactions

As discussed in Chapter Five, if failing to pass a filter leads to the population at the last stage being much larger than it otherwise would, failing to pass that filter can potentially become less favoured over passing it under SSA. This is not an issue with most of the impediments considered above. Failing to avoid an existential risk does not allow the population at the previous stage to continue. Failure to thrive allows the population to continue, but probably at a much lower level than if it were technologically advanced. However a change in values to prefer not colonizing space wouldn't obviously affect the population at all as long as the civilization in question was filtered before they actually colonized space. This means unless changing values takes a population out of the reference class being used, SSA should generally not prefer people changing their values to not, assuming they do not pass the next step either way.

## 6.6 Summary

Most of the potential filters we may meet involve the end of the human race. Some ways the human race could end are not filters. Depending on the reference class, SSA gives as much weight to the human race ending by these other means, whereas SIA only gives weight to disasters that are filters. This means vacuum decay, simulations being shut down, artificial intelligence explosions, and Earth-specific problems are only given more weighting than you would otherwise expect under SSA, with various reference classes. This makes averting filters more important if you believe SIA. SSA also gives more weight than SIA to disasters in the

nearer future, such as new technologies being misused, relative to impediments in the far future, such as changing our values or being invisible. The effect of SSA on likely filters is mediated by the populations at the different levels, and how being filtered affects the earlier population. This means SSA may not put extra weight on people changing their values being a strong filter. To SIA these population considerations are irrelevant, so under SIA a change in values is treated like any other filter.

# Chapter 7: Conclusion

SIA and FNC imply that Great Filter steps in our future are likely to be larger than we naively believe. SSA has the same implications with a variety of reference classes, though not all of them. If we increase our expectation of stronger filters at future steps, we must also increase our expectation of human extinction from a variety of causes. We must do this because human extinction accounts for many possible filter steps, the strength of which we are unsure. These things together mean that standard estimates of human extinction risks are systematically underestimated regardless of how the debate over indexical updating is resolved.

SSA's mechanism of increasing the probability of future filters is similar to the Doomsday Argument. SIA does not cancel the doomsday effect in the Great Filter case because possible worlds contain different numbers of people at our own stage, unlike possible worlds in the Doomsday Argument.

The details of SIA's and SSA's predictions differ significantly. SIA forecasts a greater probability of any potential filter step with uncertainty surrounding its strength. SSA generally predicts a greater probability of filter steps in the nearer future as well as disasters that aren't filter steps. The SIA probability shift for a given possible world is larger than the SSA shift. SIA gives a precise prediction for a given set of priors, while SSA can be used with a wide range of reference classes, for different results. The SSA result also depends on populations at various stages, while the SIA result does not. SIA relies more strongly on the fate of humanity being similar to that of any other civilizations in a similar situation.

We saw that SSA and SIA stand out as plausible and popular principles. If the Doomsday Argument reduced the plausibility of SSA, SIA's conclusion from minimal evidence that we are highly likely to be destroyed may be similar evidence against SIA, or in favour of the Doomsday Argument's legitimacy. By related reasoning, the Presumptuous Philosopher thought experiment is not as strong evidence against SIA as it seems. The best principle remains unclear. This leaves uncertainty over whether to expect the detailed predictions of SSA or of SIA. However as long as those principles are the most likely contenders, we can be confident that we have previously underestimated the risk of human extinction.

# References

Adams, F.C., 2008. Long-term astrophysical processes. In *Global Catastrophic Risks*. United States: Oxford University Press.

Adams, T., 2007. Sorting out the Anti-Doomsday Arguments: A Reply to Sowers. *Mind*, 116(462), 269-273.

Allen, M. & Frame, D., 2008. Climate change and global risk. In *Global Catastrophic Risks*. United States: Oxford University Press.

Armstrong, S., 2009. Avoiding doomsday: a "proof" of the self-indication assumption. *Less Wrong*. Available at: http://lesswrong.com/lw/18r/avoiding_doomsday_a_proof_of_the_selfindication / [Accessed October 26, 2010].

Ball, J., 1973. The Zoo Hypothesis. *Icarus*, 19(3), 347-349.

Bayes, 1958. An essay towards solving a Problem in the Doctrine of Chances. *Biometrika*, 45(3-4), 296-315.

Bostrom, N., 1999. The Doomsday Argument Is Alive and Kicking. *Mind*, 108(431), 539-553.

Bostrom, N., 2001. The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe. *Synthese*, 127(3), 359-387.

Bostrom, N., 2002a. *Anthropic bias*, New York: Routledge.

Bostrom, N., 2002b. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, 9. Available at: http://www.nickbostrom.com/existential/risks.pdf.

Bostrom, N., 2002c. Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation. *The Journal of Philosophy*, 99(12), 607-623.

Bostrom, N., 2003a. Are We Living in a Computer Simulation? *The Philosophical Quarterly*, 53(211), 243-255.

Bostrom, N., 2003b. Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*, 15(03), 308-314.

Bostrom, N., 2006. Dinosaurs, Dodos, Humans? *Global Agenda*, 230-231.

Bostrom, N., 2007. Sleeping Beauty and Self-Location: A Hybrid Model. *Synthese*, 157(1), 59-78.

Bostrom, N., 2008. Where Are They? Why I hope the search for extraterrestrial life finds nothing. *MIT Technology Review*. Available at: http://www.technologyreview.com/infotech/20569/ [Accessed July 21, 2010].

Bostrom, N. & Ćirković, M.M., 2003. The Doomsday Argument and the Self-Indication

Assumption: Reply to Olum. *The Philosophical Quarterly*, 53(210), 83-91.

Bovens, L., 2010. Judy Benjamin is a Sleeping Beauty. *Analysis*, 70(1), 23-26.

Center for Responsible Nanotechnology, 2008a. Nanotechnology Research. *Center for Responsible Nanotechnology*. Available at: http://www.crnano.org/ [Accessed July 22, 2010].

Center for Responsible Nanotechnology, 2008b. Nanotechnology: Dangers of Molecular Manufacturing. Available at: http://www.crnano.org/dangers.htm [Accessed July 20, 2010].

Chalmers, D., 2010. Personal Communication.

Chambers, T., 2001. Do Doomsday's Proponents Think We Were Born Yesterday? *Philosophy*, 76(297), 443-450.

Cirincione, J., 2008. The continuing threat of nuclear war. In *Global Catastrophic Risks*. United States: Oxford University Press.

Cirkovic, M.M. & Cathcart, R., 2004. Geo-engineering Gone Awry: A New Partial Solution of Fermi's Paradox. *Journal of the British Interplanetary Society*, 57, 209-215. Available at: http://arxiv.org/abs/physics/0308058 [Accessed July 25, 2010].

Cohen, J. & Easterly, W., 2009. *What works in development: thinking big and thinking small*, Brookings Institution Press.

Coleman, S. & De Luccia, F., 1980. Gravitational effects on and of vacuum decay. *Physical Review D*, 21(12), 3305-3315.

Craig, A., 2003. Astronomers count the stars. *BBC News*. Available at: [Accessed April 8, 2010].

Dar, A., De Rujula, A. & Heinz, U., 1999. Will relativistic heavy-ion colliders destroy our planet? *Physics Letters B*, 470(1-4), 142-148.

Davies, P., 2010. *The Eerie Silence: Are We Alone In The Universe?*, Allan Lane.

Dieks, D., 1992. Doomsday--Or: The Dangers of Statistics. *The Philosophical Quarterly*, 42(166), 78-84.

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

Eckhardt, W., 1997. A Shooting-Room View of Doomsday. *The Journal of Philosophy*, 94(5), 244-259.

Eckhardt, W., 1993. Probability Theory and the Doomsday Argument. *Mind*, 102(407), 483-488.

Elga, A., 2000. Self-locating belief and the Sleeping Beauty problem. *Analysis*, 60(266),

143-147.

Elga, A., 2004. Defeating Dr. Evil with Self-Locating Belief. *Philosophy and Phenomenological Research*, 69(2), 383-396.

Finkelstein, J., 2008. Sleeping Beauty: theme and variations. Available at: http://philsci-archive.pitt.edu/archive/00004318/ [Accessed July 20, 2010].

Franceschi, P., 2004. Sleeping Beauty in Flatland. Available at: http://philsci-archive.pitt.edu/archive/00001580/ [Accessed July 20, 2010].

Gott, J.R., 1993. Implications of the Copernican principle for our future prospects. *Nature*, 363(6427), 315-319.

Hanson, R., 2008. Catastrophe, Social Collapse, and Human Extinction. In *Global Catastrophic Risks*. United States: Oxford University Press, pp. 363-376. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.63.1401 [Accessed July 21, 2010].

Hanson, R., 1998a. Critiquing the Doomsday Argument. Available at: http://hanson.gmu.edu/nodoom.html [Accessed October 24, 2010].

Hanson, R., 1998b. The Great Filter - Are We Almost Past It? Available at: http://hanson.gmu.edu/greatfilter.html [Accessed May 5, 2010].

Hansson, I. & Stuart, C., 1990. Malthusian Selection of Preferences. *American Economic Review*, 80(3), 529-544.

Jenkins, C.S., 2005. Sleeping Beauty: A Wake-Up Call. *Philosophia Mathematica*, 13(2), 194-201.

Joy, B., 2000. Why the future doesn't need us. *Wired Magazine*. Available at: http://www.wired.com/wired/archive/8.04/joy_pr.html [Accessed July 21, 2010].

Kilbourne, E., 2008. Plagues and pandemics: past, present, and future. In *Global Catastrophic Risks*. United States: Oxford University Press.

Knobe, J., Olum, K.D. & Vilenkin, A., 2006. Philosophical Implications of Inflationary Cosmology. *The British Journal for the Philosophy of Science*, 57(1), 47-67.

Kopf, T., Krtous, P. & Page, D.N., 1994. Too Soon for Doom Gloom? *Arxiv*. Available at: http://arxiv.org/abs/gr-qc/9407002 [Accessed July 25, 2010].

Korb, K.B. & Oliver, J.J., 1998. A Refutation of the Doomsday Argument. *Mind*, 107(426), 403-410.

Kwik, G. et al., 2003. Biosecurity: Responsible Stewardship of Bioscience in an Age of Catastrophic Terrorism. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 1(1), 27-35.

Leslie, J., 1993. Doom and Probabilities. *Mind*, 102(407), 489-491.

Leslie, J., 1997. Observer-relative chances and the doomsday argument. *Inquiry: An Interdisciplinary Journal of Philosophy*, 40(4), 427.

Leslie, J., 1990. Risking the world's end. *Interchange*, 21(1), 49-58.

Leslie, J., 1992. The doomsday argument. *The Mathematical Intelligencer*, 14(2), 48-51.

Leslie, J., 1998. *The end of the world: the science and ethics of human extinction*, Routledge.

Lewis, D., 2001. Sleeping Beauty: reply to Elga. *Analysis*, 61(271), 171-76.

Matheny, J.G., 2007. Reducing the Risk of Human Extinction. *Risk Analysis*, 27(5), 1335-1344.

Millennium Ecosystem Assessment, W.V., 2005. Ecosystems and Human Wellbeing. In *Millennium Ecosystem Assessment*. Washington, DC: Island Press.

Miller, G., 2006. Why we haven't met any aliens. *Seed Magazine*.

Monton, B., 2003. The Doomsday Argument without Knowledge of Birth Rank. *The Philosophical Quarterly*, 53(210), 79-82.

Napier, W., 2008. Hazards from comets and asteroids. In *Global Catastrophic Risks*. United States: Oxford University Press.

Neal, R.M., 2006. Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning. *Arxiv*. Available at: http://arxiv.org/abs/math/0608592 [Accessed July 23, 2010].

Olum, K.D., 2002. The Doomsday Argument and the Number of Possible Observers. *The Philosophical Quarterly*, 52(207), 164-184.

Parfit, D., 1984. *Reasons and persons*, Oxford, UK: Clarendon Press.

Phoenix, C. & Treder, M., 2008. Nanotechnology as global catastrophic risk. In *Global Catastrophic Risks*. United States: Oxford University Press.

Posner, R.A., 2004. *Catastrophe: Risk and Response* First Edition and First Printing., Oxford University Press, USA.

Rees, M., 2004. *Our Final Hour*, Basic Books.

Rees, M.J., Bostrom, N. & Cirkovic, M.M., 2008. *Global Catastrophic Risks* First Edition., OUP Oxford.

Robock, A., Oman, L. & Stenchikov, G.L., 2007. Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences. *Journal of Geophysical Research*, 112(D13107), 14.

Sandberg, A. & Bostrom, N., 2008. Global Catastrophic Risks Survey.

Sandberg, A., Bostrom, N. & Cirkovic, M., 2010. Anthropic Shadow: Observation Selection Effects and Human Extinction Risks. *Risk Analysis*.

Sandberg, A., Matheny, J.G. & Cirkovic, M.M., 2008. How can we reduce the risk of human extinction? Available at: http://www.thebulletin.org/web-edition/features/how-can-we-reduce-the-risk-of-human-extinction [Accessed July 21, 2010].

Sandberg, A. & Shulman, C., 2010. Implications of a Software-limited Singularity. In European Conference on Computing and Philosophy 2010.

Sandel, M.J., 2006. *Public philosophy: essays on morality and politics*, Harvard University Press.

Shulman, C., 2010. Personal Communication.

Singularity Institute for Artificial Intelligence, 2010. Singularity Institute for Artificial Intelligence | AI. Available at: http://singinst.org/ [Accessed July 21, 2010].

Sober, E., 2003. An Empirical Critique of Two Versions of the Doomsday Argument: Gott's Line and Leslie's Wedge. *Synthese*, 135(3), 415-430.

Sowers, G.F., 2002. The Demise of the Doomsday Argument. *Mind*, 111(441), 37-46.

Tegmark, M. & Bostrom, N., 2005. Astrophysics: Is a doomsday catastrophe likely? *Nature*, 438(7069), 754.

The Foresight Institute, 2010. The Foresight Institute. Available at: http://www.foresight.org/ [Accessed July 22, 2010].

The Lifeboat Foundation, 2010. Lifeboat Foundation BioShield. Available at: http://lifeboat.com/ex/bio.shield?background=white [Accessed July 22, 2010].

Turchin, A., 2008. *Structure of the Global Catastrophe: Risks of human extinction in the XXI century*, Moscow. Available at: http://www.scribd.com/doc/6250354/STRUCTURE-OF-THE-GLOBAL-CATASTROPHE-Risks-of-human-extinction-in-the-XXI-century- [Accessed July 20, 2010].

Webb, S., 2002. *If the universe is teeming with aliens - where is everybody?*, United States: Springer.

Weitzman, M., 2009. On modeling and interpreting the economics of catastrophic climate change. *Review of Economics and Statistics*, 91(1), 1-19.

White, R., 2006. The generalized Sleeping Beauty problem: a challenge for thirders. *Analysis*, 66(290), 114-119.

Williams, M., 2006. The Knowledge. *Technology Review (MIT)*, (March/April).

Wills, C., 2008. Evolution theory and the future of humanity. In *Global Catastrophic Risks*. United States: Oxford University Press.

Yudkowsky, E., 2008. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risks*. United States: Oxford University Press.