ANTHROPIC BIAS

Observation Selection Effects in Science and Philosophy

Nick Bostrom

Routledge
New York & London

Published in 2002 by
Routledge
29 West 35th Street
New York, NY 10001

Published in Great Britain by
Routledge
11 New Fetter Lane
London EC4P 4EE

Routledge is an imprint of the Taylor & Francis Group
Printed in the United States of America on acid-free paper.

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Bostrom, Nick, 1973–
Anthropic bias : observation selection effects in science and philosophy / by Nick Bostrom.
p. cm. — (Studies in philosophy)
Includes bibliographical references and index.
ISBN 0-415-93858-9
1. Methodology. 2. Anthropic principle. 3. Selectivity (Psychology) 4. Observation (Scientific method) I. Title. II. Studies in philosophy (New York, N.Y.)
BD241.B657 2002
121'.6—dc21 2001058887

This book is dedicated to my father—*tack pappa!*

Contents

Acknowledgments

This work has benefited from copious feedback generated by bits and pieces that have appeared in print earlier. Over the years, I must have corresponded with several hundreds of people about these issues. In addition, I've received comments from conference audiences, journal referees, students, and authors of replies to parts that have already been published. For all this, I am extremely grateful!

There is a website associated with the book, www.anthropic-principle.com, containing a preprint archive of relevant writings that are available online, an updated bibliography, primers on various topics, and other resources to aid scholars and interested laypersons to get up to speed with the latest research on observation selection effects.

Although I cannot name everybody who has helped me in some way with this project, there are some who must be singled out for my special thanks: Paul Bartha, Darren Bradley, John Broome, Jeremy Butterfield, Erik Carlson, Brandon Carter, Douglas Chamberlain, Robin Collins, Pierre Cruse, Wei Dai, J-P Delahaye, Jean-Michel Delhotel, Dennis Dieks, William Eckhardt, Ellery Eells, Adam Elga, Hal Finney, Paul Franceschi, Richard Gott, Mark Greenberg, Robin Hanson, Daniel Hill, Christopher Hitchcock, Richard Jeffrey, Bill Jefferys, Vassiliki Kambourelli, Loren A. King, Kevin Korb, Eugene Kusmiak, Jacques Mallah, Neil Manson, Peter Milne, Bradley Monton, Floss Morgan, Samuel Newlands, Jonathan Oliver, Ken Olum, Don N. Page, David Pearce, Elliott Sober, Richard Swinburne, Max Tegmark, Alexander Vilenkin, Saar Wilf, and Roger White. I am so grateful to all those friends, named and unnamed, without whose input this book could not have been written. (The faults that it contains, however, I was perfectly capable of producing all by myself!)

I want to especially thank John Leslie for his exceedingly helpful guidance, Colin Howson and Craig Callender for long assistance and advice, Nancy Cartwright for stepping in and removing a seemingly insurmount-

able administrative obstacle, and Milan M. Ćirković for keeping up collaboration with me on a paper whilst bombs were detonating around him in Belgrade. Finally, I want to thank Robert Nozick for encouraging rapid publication.

I gratefully acknowledge a research grant from the John Templeton Foundation that has helped fund large parts of the research. I'm thankful to *Synthese*, *Mind*, *Analysis*, and *Erkenntnis* for permitting texts to be republished.

Preface

This book explores how to reason when you suspect that your evidence is biased by observation selection effects. An explanation of what observation selection effects are has to await chapter 1. Suffice it to say here that the topic is intellectually fun, difficult, and important. We will be discussing many interesting applications; philosophical thought experiments and paradoxes aside, we will use our results to address several juicy bits of contemporary science: cosmology (how many universes are there?), evolution theory (how improbable was the evolution of intelligent life on our planet?), the problem of time's arrow (can it be given a thermodynamic explanation?), game theoretic problems with imperfect recall (how to model them?), traffic analysis (why is the "next lane" faster?) and a lot more—the sort of stuff that intellectually active people like to think about . . .

CHAPTER 1

Introduction

OBSERVATION SELECTION EFFECTS

How big is the smallest fish in the pond? You catch one hundred fishes, all of which are greater than six inches. Does this evidence support the hypothesis that no fish in the pond is much less than six inches long? Not if your net can't catch smaller fish.

Knowledge about limitations of your data collection process affects what inferences you can draw from the data. In the case of the fish-size-estimation problem, a *selection effect*—the net's sampling only the big fish—vitiates any attempt to extrapolate from the catch to the population remaining in the water. Had your net instead sampled randomly from all the fish, then finding a hundred fishes all greater than a foot would have been good evidence that few if any of the fish remaining are much smaller.

In 1936, the *Literary Digest* conducted a poll to forecast the result of the upcoming presidential election. They predicted that Alf Landon, the Republican candidate, would win by a large margin. In the actual election, the incumbent Franklin D. Roosevelt won a landslide victory. The *Literary Digest* had harvested the addresses of the people they sent the survey to mainly from telephone books and motor vehicle registries, thereby introducing an important selection effect. The poor of the depression era, a group where support for Roosevelt was especially strong, often did not have a phone or a car. A methodologically more sophisticated forecast would either have used a more representative polling group or at least fac-

[1] The *Literary Digest* suffered a major reputation loss as a result of the infamous poll and soon went out of business, being superceded by new generation of pollsters such as George Gallup, who not only got the 1936 election right but also predicted what the *Literary Digest's* prediction would be to within 1%, using a sample size just one thousandth the size of the *Digest's* but more successfully avoiding selection effects. The infamous 1936 poll has secured a place in the annals

tored in known and suspected selection effects.[1]

Or to take yet another example, suppose you're a young investor pondering whether to invest your retirement savings in bonds or equity. You are vaguely aware of some studies showing that over sufficiently lengthy periods of time, stocks have, in the past, substantially outperformed bonds (an observation which is often referred to as the "equity premium puzzle"), so you are tempted to put your money in equity. You might want to consider, though, that a selection effect might be at least partly responsible for the apparent superiority of stocks. While it is true that most of the readily available data does favor stocks, this data is mainly from the American and British stock exchanges, which both have continuous records of trading dating back over a century. But is it an accident that the best data comes from these exchanges? Both America and Britain have benefited during this period from stable political systems and steady economic growth. Other countries have not been so lucky. Wars, revolutions, and currency collapses have at times obliterated entire stock exchanges, which is precisely why continuous trading records are not available elsewhere. By looking only at the two greatest success stories, one would risk overestimating the historical performance of stocks. A careful investor, it seems, would be wise to factor in this consideration when designing her portfolio. (For one recent study that attempts to estimate this survivorship bias by excavating and patching together the fragmentary records from other exchanges, see (Jorion and Goetzmann 2000); for some theory on survivorship biases, see (Brown 1995).)

In these three examples, a selection effect is introduced by the fact that the instrument you use to collect data (a fishing net, a mail survey, preserved trading records) samples only from a proper subset of the target domain. Analogously, there are selection effects that arise not from the limitations of some measuring device but from the fact that all observations require the existence of an appropriately positioned observer. Our data is filtered not only by limitations in our instrumentation but also by the precondition that somebody be there to "have" the data yielded by the instruments (and to build the instruments in the first place). The biases that occur due to that precondition—we shall call them *observation* selection effects—are the subject matter of this book.

Anthropic reasoning, which seeks to detect, diagnose, and cure such biases, is a philosophical goldmine. Few fields are so rich in empirical implications, touch on so many important scientific questions, pose such intricate paradoxes, and contain such generous quantities of conceptual and methodological confusion that need to be sorted out. Working in this area is a lot of intellectual fun.

---

of survey research as a paradigm example of selection bias, yet just as important was a non-response bias compounding the error referred to in the text (Squire 1988).—The fishing example originates from Sir Arthur Eddington (Eddington 1939).

Let's look at an example where an observation selection effect is involved: We find that intelligent life evolved on Earth. Naively, one might think that this piece of evidence suggests that life is likely to evolve on most Earth-like planets. But that would be to overlook an observation selection effect. For no matter how small the proportion of all Earth-like planets that evolve intelligent life, we will find ourselves on a planet that did (or we will trace our origin to a planet where intelligent life evolved, in case we are born in a space colony). Our data point—that intelligent life arose on our planet—is predicted equally well by the hypothesis that intelligent life is very improbable even on Earth-like planets as by the hypothesis that intelligent life is highly probable on Earth-like planets. This datum therefore does not distinguish between the two hypotheses, provided that on both hypotheses intelligent life would have evolved somewhere. (On the other hand, if the "intelligent-life-is-improbable" hypothesis asserted that intelligent life was so improbable that is was unlikely to have evolved *anywhere* in the whole cosmos, then the evidence that intelligent life evolved on Earth *would* count against it. For this hypothesis would not have predicted our observation. In fact, it would have predicted that there would have been no observations at all.)

We don't have to travel long on the path of common sense before we enter a territory where observation selection effects give rise to difficult and controversial issues. Already in the preceding paragraph we passed over a point that is contested. We understood the explanandum, that intelligent life evolved on our planet, in a "non-rigid" sense. Some authors, however, argue that the explanandum should be: why did intelligent life evolve on *this* planet (where "this planet" is used as a rigid designator). They then argue that the hypothesis that intelligent life is quite probable on Earth-like planets would indeed give a higher probability to this fact (Hacking 1987; Dowe 1998; White 2000). But we shall see in the next chapter that that is not the right way to understand the problem.

The impermissibility of inferring from the fact that intelligent life evolved on Earth to the fact that intelligent life probably evolved on a large fraction of all Earth-like planets does not hinge on the evidence in this example consisting of only a single data point. Suppose we had telepathic abilities and could communicate directly with all other intelligent beings in the cosmos. Imagine we ask all the aliens, did intelligent life evolve on their planets too? Obviously, they would all say: Yes, it did. But equally obvious, this multitude of data would still not give us any reason to think that intelligent life develops easily. We only asked about the planets where life did in fact evolve (since those planets would be the only ones which would be "theirs" to some alien), and we get no information whatsoever by hearing the aliens confirming that life evolved on those planets (assuming we don't know the number of aliens who replied to our survey or, alternatively, that we don't know the total number of planets). An observation selection effect frustrates any attempt to extract useful information

by this procedure. Some other method would have to be used to do that. (If all the aliens also reported that theirs was some Earth-like planet, this would suggest that intelligent life is *unlikely* to evolve on planets that are *not* Earth-like; for otherwise some aliens would likely have evolved on non-Earth like planets.)

Another example of reasoning that invokes observation selection effects is the attempt to provide a possible (not necessarily the only) explanation of why the universe appears fine-tuned for intelligent life in the sense that if any of various physical constants or initial conditions had been even very slightly different from what they are, then life as we know it would not have existed. The idea behind this possible anthropic explanation is that the totality of spacetime might be very huge and may contain regions in which the values of fundamental constants and other parameters differ in many ways, perhaps according to some broad random distribution. If this is the case, then we should not be amazed to find that in our own region physical the conditions appear "fine-tuned". Owing to an obvious observation selection effect, only such fine-tuned regions are observed. Observing a fine-tuned region is precisely what we should expect if this theory is true, and so it can potentially account for available data in a neat and simple way, without having to assume that conditions *just happened* to turn out "right" through some immensely lucky—and arguably a priori extremely improbable—cosmic coincidence. (Some skeptics doubt that an explanation for the apparent fine-tuning of our universe is needed or is even meaningful. We examine the skeptical arguments in chapter 2 and consider the counterarguments offered by proponents of the anthropic explanation.)

Here are some of the topics we shall be covering: cosmic fine-tuning arguments for the existence of a multiverse or alternatively a cosmic "designer"; so-called anthropic principles (and how they fall short); how to derive observational predictions from inflation theory and other contemporary cosmological models; the Self-Sampling Assumption; observation selection effects in evolutionary biology and in the philosophy of time; the Doomsday argument, the Adam & Eve, UN^{++} and Quantum Joe paradoxes; alleged observer-relative chances; the Presumptuous Philosopher gedanken; the epistemology of indexical belief; game theoretic problems with imperfect recall; and much more.

Our primary objective is to construct a theory of observation selection effects. We shall seek to develop a methodology for how to reason when we suspect that our evidence is contaminated with anthropic biases. Our secondary objective is to apply the theory to answer some interesting scientific and philosophical questions. Actually, these two objectives are largely overlapping. Only by interpolating between theoretical desiderata and the full range of philosophical and scientific applications can we arrive at a satisfactory account of observation selection effects. At least, that is the approach taken here.

We'll use a Bayesian framework, but a reader who doesn't like formalism should not be deterred. There isn't an excessive amount of mathematics; most of what there is, is elementary arithmetic and probability theory, and the results are conveyed verbally also. The topic of observation selection effects *is* extremely difficult, yet the difficulty is not in the math, but in grasping and analyzing the underlying principles and in selecting the right models.

A BRIEF HISTORY OF ANTHROPIC REASONING

Even trivial selection effects can sometimes easily be overlooked:

> It was a good answer that was made by one who when they showed him hanging in a temple a picture of those who had paid their vows as having escaped shipwreck, and would have him say whether he did not now acknowledge the power of the gods,—'Aye,' asked he again, 'but where are they painted that were drowned after their vows?' And such is the way of all superstition, whether in astrology, dreams, omens, divine judgments, or the like; wherein men, having a delight in such vanities, mark the events where they are fulfilled, but where they fail, though this happens much oftener, neglect and pass them by (Bacon 1620)

When even a plain and simple selection effect, such as the one that Francis Bacon comments on in the quoted passage, can escape a mind that is not paying attention, it is perhaps unsurprising that *observation selection effect*s, which tend to be more abstruse, have only quite recently been given a name and become a subject of systematic study.[2]

The term "anthropic principle", which has been used to label a wide range of things only some of which bear a connection to observation selection effects, is less than three decades old. There are, however, precursors from much earlier dates. For example, in Hume's *Dialogues Concerning Natural Religion*, one can find early expressions of some ideas of anthropic selection effects. Some of the core elements of Kant's philosophy about how the world of our experience is conditioned on the forms of our sensory and intellectual faculties are not completely unrelated to modern ideas about observation selection effects as important methodological considerations in theory-evaluation, although there are also fundamental differences. In Ludwig Boltzmann's attempt to give a thermodynamic account of

[2] Why isn't the selection effect that Bacon refers to an "observational" one? After all, nobody could observe the bottom of the sea at that time.—Well, one could have observed that the sailors had gone missing. Fundamentally, the criterion we can use to determine whether something is an observation selection effect is whether a theory of observation selection effects is needed to model it. That doesn't seem necessary for the case Bacon describes.

time's arrow (Boltzmann 1897), we find for perhaps the first time a scientific argument that makes clever use of observation selection effects. We shall discuss Boltzmann's argument in one of the sections of chapter 4, and show why it fails. A more successful invocation of observation selection effects was made by R. H. Dicke (Dicke 1961), who used it to explain away some of the "large-number coincidences", rough order-of-magnitude matches between some seemingly unrelated physical constants and cosmic parameters, that had previously misled such eminent physicists as Eddington and Dirac into a futile quest for an explanation involving bold physical postulations.

The modern era of anthropic reasoning dawned quite recently, with a series of papers by Brandon Carter, another cosmologist. Carter coined the term "anthropic principle" in 1974, clearly intending it to convey some useful guidance about how to reason under observation selection effects. We shall later look at some examples of how he applied his methodological ideas to both physics and biology. While Carter himself evidently knew how to apply his principle to get interesting results, he unfortunately did not manage to explain it well enough to enable all his followers to do the same.

The term "anthropic" is unfortunate, because reasoning about observation selection effects has nothing in particular to do with homo sapiens, but rather with observers in general. Carter regrets not having chosen a better name, which would no doubt have prevented much of the confusion that has plagued the field. When John Barrow and Frank Tipler introduced anthropic reasoning to a wider audience in 1986 with the publication of *The Anthropic Cosmological Principle*, they compounded the terminological disorder by minting several new "anthropic principles", some of which have little if any connection to observation selection effects.

A total of over thirty anthropic principles have been formulated and many of them have been defined several times over—in nonequivalent ways—by different authors, and sometimes even by the same authors on different occasions. Not surprisingly, the result has been some pretty wild confusion concerning what the whole thing is about. Some reject anthropic reasoning out of hand as representing an obsolete and irrational form of anthropocentrism. Some hold that anthropic inferences rest on elementary mistakes in probability calculus. Some maintain that at least some of the anthropic principles are tautological and therefore indisputable. Tautological principles have been dismissed by some as empty and thus of no interest or ability to do explanatory work. Others have insisted that like some results in mathematics, though analytically true, anthropic principles can nonetheless be interesting and illuminating. Others still purport to derive empirical predictions from these same principles and regard them as testable hypotheses. Obviously, we shall want to distance ourselves from most of these would-be codifications of the anthropic organon.

Some reassurance comes from the meta-level consideration that

anthropic reasoning is used and taken seriously by a range of leading physicists. One would not expect this bunch of hardheaded scientists to be just blowing so much hot air. And we shall see that once one has carefully removed extraneous principles, misconceptions, fallacies and misdescriptions, one does indeed find a precious core of methodological insights.

Brandon Carter also originated the notorious Doomsday argument, although he never published on it. First to discuss it in print was philosopher John Leslie, whose prolific writings have elucidated a wide range of other issues related to anthropic reasoning. A version of the Doomsday argument was invented independently by Richard Gott, an astrophysicist. The Doomsday argument has generated a bulky literature of its own, which sometimes suffers from being disconnected from other areas of anthropic reasoning. One lesson from this book is, I think, that different applications of anthropic reasoning provide important separate clues to what the correct theoretical account of observation selection effects must look like. Only when we put all the pieces of the puzzle together in the right way does a meaningful picture emerge.

The field of observational selection has begun to experience rapid growth in recent years. Many of the of the most important results date back only about a decade or less. Philosophers and scientists (especially cosmologists) deserve about equal parts of the credit for the ideas that have already been developed and which this book can now use as building blocks.

SYNOPSIS OF THIS BOOK

Our journey begins in chapter 2 with a study of the significance of cosmic "fine-tuning", referring to the apparent fact that if any of various physical parameters had been very slightly different then no observers would have existed in the universe. There is a sizable literature on what to make of such "coincidences". Some have argued that they provide some evidence for the existence of an ensemble of physically real universes (a "multiverse"). Others, of a more religious bent, have used arguments from fine-tuning to attempt to make a case for some version of the design hypothesis. Still others claim that comic fine-tuning can have no special significance at all. The latter view is incorrect. The finding that we live in a fine-tuned universe (if that is indeed so) would, as we shall see, provide support to explanations that essentially involve observation selection effects. Such explanations raise interesting methodological issues which we will be exploring in chapter 2. I argue that only by working out a theory of observation selection effects can we get to the bottom of the fine-tuning controversies. Using analogies, we begin to sketch out a preliminary account of how observation selection effects operate in the cosmological context, which allows us to get a clearer understanding of evidential import of fine-

tuning. Later, in chapter 11, we will return to the fine-tuning arguments and use the theory that we'll have developed in the intervening chapters to more rigorously verify the informal results of chapter 2.

Given that observation selection effects are important, we next want to know more precisely what kind of beast they are and how they affect methodology. Is it possible to sum up the essence of observation selection effects in a simple statement? A multitude of so-called "anthropic principles" attempt to do just that. Chapter 3 takes a critical look at the main contenders, and finds that they fall short. Many "anthropic principles" are simply confused. Some, especially those drawing inspiration from Brandon Carter's seminal papers, are sound, but we show that although they point in the right direction they are too weak to do any real scientific work. In particular, I argue that existing methodology does not permit any observational consequences to be derived from contemporary cosmological theories, in spite of the fact that these theories quite plainly can be and are being tested empirically by astronomers. What is needed to bridge this methodological gap is a more adequate formulation of how observation selection effects are to be taken into account. A preliminary formulation of such a principle, which we call the *Self-Sampling Assumption*, is proposed towards the end of chapter 3. The basic idea of the Self-Sampling Assumption is, very roughly put, that you should think of yourself as if you were a random observer from a suitable reference class.

Chapter 4 begins to build a "philosophical" case for our theory by conducing a series of thought experiments that show that something like the Self-Sampling Assumption describes a plausible way of reasoning about a wide range of cases.

Chapter 5 shows how the Self-Sampling Assumption enables us to link up cosmological theory with observation in a way that is both intuitively plausible and congruent with scientific practice. This chapter also applies the new methodology to illuminate problems in several areas, to wit: thermodynamics and the problem of time's arrow; evolutionary biology (especially questions related to how improbable was the evolution of intelligent life on Earth and how many "critical" steps there were in our evolutionary past); and an issue in traffic analysis. An important criterion for a theory of observation selection effects is that it should enable us to make sense of contemporary scientific reasoning and that it can do interesting work in helping to solve real empirical problems. Chapter 5 demonstrates that our theory satisfies this criterion.

The notorious Doomsday argument, which seeks to show that we have systematically underestimated the probability that humankind will go extinct relatively soon, forms the subject matter for chapter 6. We review and criticize the literature on this controversial piece of reasoning, both papers that support it and ones that claim to have refuted it. I think that the Doomsday argument is inconclusive, but the reason is complicated and must await explanation until we have developed our theory further, in

chapter 10.

The Doomsday argument deserves the attention it has attracted, however. Getting to the bottom of what is wrong or inconclusive about it can give us invaluable clues about how build a sound methodology of observation selection effects. It is therefore paramount that the Doomsday argument not be dismissed for the wrong reasons. Lots of people think that they have refuted the Doomsday argument, but not all these objections can be right—many of the "refutations" are inconsistent with one another, and many presuppose ideas that can be shown unacceptable when tried against other criteria that a theory of anthropic reasoning must satisfy. Chapter 7 examines several recent criticisms of the Doomsday argument and shows that they all fail.

In chapter 8, we refute an argument purporting to show that anthropic reasoning gives rise to paradoxical observer-relative chances. We then give an independent argument showing that there are cases where anthropic reasoning does generate probabilities that are "observer-relative" in an interesting but non-paradoxical sense.

Paradoxes lie in ambush in chapter 9. We explore the thought experiments *Adam & Eve*, *UN*$^{++}$, and *Quantum Joe*. These reveal some counterintuitive aspects of the most straightforward version of the Self-Sampling Assumption.

Is there a way out? At the end of chapter 9 we find ourselves in an apparent dilemma. On the one hand, something like the Self-Sampling Assumption seems philosophically justified and scientifically indispensable on the grounds explained in chapters 4 and 5. On the other hand, we seem then to be driven towards a counterintuitive (albeit coherent) position vis-à-vis the gedanken experiments of chapter 9. What to do?

Chapter 10 goes back and reexamines the reasoning that led to the formulation of the original version of the Self-Sampling Assumption. But now we have the benefit of lessons gleaned from the preceding chapters. We understand better the various constraints that our theory needs to satisfy. And we have a feel for what is the source of the problems. Combining these clues, we propose a solution that enables us to escape the paradoxes while still catering to legitimate methodological needs. The first step of the solution is to strengthen the Self-Sampling Assumption so that it applies to "observer-moments" rather than just observers. This increases our analytical firepower. A second step is to relativize the reference class. The result is a general framework for modeling anthropic reasoning, which is given a formal expression in an equation that specifies how to take into account evidence that has an indexical component or that has been subjected to an observation selection effect.

In chapter 11, we illustrate how this theory of observation selection effects works by applying it to a wide range of philosophical and scientific problems. We show how it confirms (and makes more precise) the preliminary conclusions that were arrived at by less rigorous analogy-based

arguments in earlier chapters. Chapter 11 also provides an analysis of the Sleeping Beauty problem (and a fortiori its closely related game-theoretic analogues, the Absent-Minded Driver problem and the Absent-Minded Passenger problem). It is argued that the solution is more complex than previously recognized and that this makes it possible to reconcile the two opposing views that dominate the literature. We close with a discussion of the element of subjectivity that may reside in the choice of a prior credence function for indexical propositions. We compare it with the more widely admitted aspect of subjectivity infesting the non-indexical component of one's credence function, and we suggest that the issue throws light on how to rank various applications of anthropic reasoning according to how scientifically rigorous they are. At the very end, there are some pointers to avenues for further research.

CHAPTER 2

Fine-Tuning Arguments in Cosmology

One aspect of anthropic reasoning that has attracted plenty of attention, from both philosophers and physicists, is its use in cosmology to explain the apparent fine-tuning of our universe. "Fine-tuning" refers to the supposed fact that there is a set of cosmological parameters or fundamental physical constants that are such that had they been very slightly different, the universe would have been void of intelligent life. For example, in the classical big bang model, the early expansion speed seems fine-tuned. Had it been very slightly greater, the universe would have expanded too rapidly and no galaxies would have formed; there would only have been a very low density hydrogen gas getting more and more dispersed as time went by. In such a universe, presumably, life could not evolve. Had the early expansion speed been very slightly less, then the universe would have recollapsed within a fraction of a second, and again there would have been no life. Our universe, having just the right conditions for life, appears to be balancing on a knife's edge (Leslie 1989). A number of other parameters seem fine-tuned in the same sense—e.g. the ratio of the electron mass to the proton mass, the magnitudes of force strengths, the smoothness of the early universe, the neutron-proton mass difference, perhaps even the metric signature of spacetime (Tegmark 1997).

Some philosophers and physicists take fine-tuning to be an explanandum that cries out for an explanans. Two possible explanations are usually envisioned: the design hypothesis and the ensemble hypothesis. Although these explanations are compatible, they tend to be viewed as competing: if we knew that one of them were correct, there would be less reason to accept the other.

The design hypothesis states that our universe is the result of purposeful design. The "agent" doing the designing need not be a theistic God, although of course that is one archetypal version of the design hypothesis. Other universe-designers have been considered in this context. For exam-

ple, John Leslie (Leslie 1972; Leslie 1979; Leslie 1989) discusses the case for a neoplatonist "causally efficacious ethical principle", which he thinks might have been responsible for creating the world and giving physical constants and cosmological parameters the numerical values they have. Derek Parfit (Parfit 1998) considers various "universe selection principles", which, although they are very different from what people have traditionally thought of as "God" or a "Designer," can nevertheless suitably be grouped under the heading of design hypotheses for present purposes. We can take "purposeful designer" in a very broad sense to refer to any being, principle or mechanism external to our universe responsible for selecting its properties, or responsible for making it in some sense probable that our universe should be fine-tuned for intelligent life. Needless to say, it is possible to doubt the meaningfulness of many of these design hypotheses. Even if one admits that a given design hypothesis represents a coherent possibility, one may still think that it should be assigned an extremely low degree of credence. For people who are already convinced that there is a God, however, the design hypothesis is likely to appear as an attractive explanation of why our universe is fine-tuned. And if one is not already convinced about the existence of a Designer, but thinks that it is a coherent possibility, one may be tempted to regard fine-tuning as reason for increasing one's credence in that hypothesis. One prominent champion of the fine-tuning argument for God's existence is Richard Swinburne (Swinburne 1991). Several other theologians and philosophers also support this position (see e.g. (Polkinghorne 1986; Craig 1988; Manson 1989; Craig 1997)).

The main rival explanation of fine-tuning is the ensemble hypothesis, which states that the universe we observe is only a small part of the totality of physical existence. This totality itself need not be fine-tuned. If it is sufficiently big and variegated so that it was likely to contain as a proper part the sort of fine-tuned universe we observe, then an observation selection effect can be invoked to explain why we see a fine-tuned universe. The usual form of the ensemble hypothesis is that our universe is but one in a vast ensemble of actually existing universes, the totality of which we can call "the multiverse". What counts as a universe in such a multiverse is a somewhat vague matter, but "a large, causally fairly disconnected spacetime region" is precise enough for our aims. If the world consists of a sufficiently huge number of such universes, and the values of physical constants vary among these universes according to some suitably broad probability distribution, then it may well be the case that it was quite probable that a fine-tuned universe like ours would come into existence. The actual existence of such a multiverse—an ensemble of "possible universes" would not do—provides the basis on which the observation selection effect operates. The argument then goes like this: Even though the vast majority of the universes are not suitable for intelligent life, it is no wonder that we should observe one of the exceptional universes which are

fine-tuned; for the other universes contain no observers and hence are not observed. To observers in such a multiverse, the world will look as if it were fine-tuned. But that is because they see only a small and unrepresentative part of the whole. Observers may marvel at the fact that the universe they find themselves in is so exquisitely balanced, but once they see the bigger picture they can realize that there is really nothing to be astonished by. On the ensemble theory, there *had* to be such a universe (or at least, it was not so improbable that there would be), and since the other universes have no observers in them, a fine-tuned universe is precisely what the observers should expect to observe given the existence of the ensemble. The multiverse itself need not be fine-tuned. It can be robust in the sense that a small change in its basic parameters would not alter the fact that it contains regions where intelligent life exists.

In contrast to some versions of the design hypothesis, the meaningfulness of the ensemble hypothesis is not much in question. Only those subscribing to a very strict verificationist theory of meaning would deny that it is possible that the world might contain a large set of causally fairly disconnected spacetime regions with varying physical parameters. And even the most hardcore verificationist would be willing to consider at least those ensemble theories according to which other universes are in principle physically accessible from our own universe. (Such ensemble theories have been proposed, although they represent only a special case of the general idea.) But there are other philosophical perplexities that arise in this context. One can wonder, for example, in what sense the suggested anthropic explanation of fine-tuning (it is "anthropic" because it involves the idea of an observation selection effect) is really explanatory and how it would relate to a more directly causal account of how our universe came to be. Another important issue is whether fine-tuning provides some evidence for a multiverse. The first question that we shall consider, however, is whether fine-tuning stands in any need of explanation at all.

DOES FINE-TUNING NEED EXPLAINING?

First a few words about the supposition that our universe is in fact fine-tuned. This is an empirical assumption that is not trivial. It is certainly true that our current best physical theories, in particular the Grand Unified Theory of the strong, weak, and electromagnetic forces and the big bang theory in cosmolog,y have a number (twenty or so) of free parameters. There is quite strong reason to think at least some of these parameters are fine-tuned—the universe would have been inhospitable to life if their values had been slightly different.[1] While it is true that our knowledge of

[1] A good overview of the case for fine-tuning can be found in chapter 2 of (Leslie 1989). For a recent discussion of some complications, see (Aguirre 2001).

"exotic" life forms possible under different physical laws than the ones that hold in the actual world is very limited (Feinberg and Shapiro 1980; Smith 1985; Wilson 1991), it does seem quite reasonable to believe, for instance, that life would not have evolved if the universe had contained only a highly diluted hydrogen gas or if it had recollapsed before the temperature anywhere had dropped below 10,000 degrees (referring to the seeming fine-tuning in the early expansion speed) (Hawking 1974; Leslie 1985). What little direct evidence we have supports this suggestion. Life does not seem to evolve easily even in a universe like our own, which presumably has rather favorable conditions—complex chemistry, relatively stable environments, large entropy gradients etc. (Simpson 1964; Papagiannis 1978; Hart 1982; Carter 1983; Mayr 1985; Raup 1985; Hanson 1998). There are as yet no signs that life has evolved in the observable universe anywhere outside our own planet (Tipler 1982; Brin 1983).

One should not jump from this to the conclusion that our universe is fine-tuned. For it is possible that some future physical theory will be developed that uses fewer free parameters or uses only parameters on which life does not sensitively depends. Even if we *knew* that our universe were not fine-tuned, the issue of what fine-tuning would have implied could still be philosophically interesting. But in fact, the case for fine-tuning is quite strong. Given what we know, it is reasonable to doubt that there is a plausible physical theory on which our universe is not fine-tuned. Inflation theory, which was originally motivated largely by a desire to avoid the fine-tuning regarding the flatness and smoothness of the universe required by the ordinary big bang theory, seems to require some fine-tuning of its own to get the inflation potential right. More recent inflation theories may overcome this problem, at least partly; but they do so by introducing a multiverse and an observation selection effect—in other words by making exactly the kind of move that this chapter will scrutinize. The present best candidate for a single-universe theory that could reduce the number of free parameters may be superstring theories (e.g. (Kane 2000), but they too seem to require at least some fine-tuning (because there are many possible compactification schemes and vacuum states). The theories that currently seem most likely to be able to do away with fine-tuned free parameters all imply the existence of a multiverse. On these theories, *our* universe might still be fine-tuned, although the multiverse as a whole might not be, or might be fine-tuned only to a less degree.

However, since the empirical case for fine-tuning is separate from the philosophical problem of how to react if our universe really is fine-tuned, we can set these scruples to one side. Let's assume the most favorable case for fine-tuning enthusiasts: that the physics of our universe has several independent free parameters which are fine-tuned to an extremely high degree. If that is so, is it something that cries out for explanation or should we be happy to accept it as one of those brute facts that just happen to obtain?

I suggest that there are two parts to the answer to this question, one of which is fairly unproblematic. This easier part of the answer is as follows: In general, simplicity is one desideratum on plausible scientific theories. Other things equal, we prefer theories which make a small number of simple assumptions to ones that involve a large number of ad hoc stipulations. This methodological principle is used successfully in all of science and it has in particular a strong track record in cosmology. For example, think of the replacement of the complicated Ptolomaic theory of planetary motion by the far simpler Copernican heliocentric theory. (Some people might regard Einstein's relativity theory as more complicated than Newton's theory of gravitation, although "more difficult" seems a more accurate description in this case than "more complicated". But note that the *ceteris paribus* includes the presupposition that the two theories predict known data equally well, so this would not be a counterexample. Newton's theory does not fit the evidence.) Thus, one should admit that there is something intellectually dissatisfying about a cosmological theory which tells us that the universe contains a large number of fine-tuned constants. Such a theory might be true, but we should not be keen to believe that until we have convinced ourselves that there is no simpler theory that can account for the data we have. So if the universe looks fine-tuned, this can be an indication that we should look harder to see if we cannot find a theory which reduces the number of independent assumptions needed. This is one reason for why a universe that looks fine-tuned (whether or not it actually *is* fine-tuned) is crying out for explanation.

We should note two things about this easy part of the answer. First, there might not be an explanation even if the universe is "crying out" for one in this sense. There is no guarantee that there is a simpler theory using fewer free parameters which can account for the data. At most, there is a prima facie case for looking for one, and for preferring the simpler theory if one can be found.

Second, the connection to fine-tuning is merely incidental. In this part of the answer, it is not fine-tuning *per se*, only fine-tuning *to the extent that it is coupled to having a wide range of free parameters*, that is instigating the hunt for a better explanation. Fine-tuning is neither necessary nor sufficient for the hunting horns to sound in this instance. It is not sufficient, because in order for a theory to be fine-tuned for intelligent life, it needs to have but a single free parameter. If a theory has a single physical constant on which the existence of intelligent life very sensitively depends, then the theory is fine-tuned. Yet a theory with only one free parameter could be eminently simple. If a universe cries out for explanation even though such a theory accounts for all available evidence, it must be on some other basis than that of a general preference for simpler theories. Also, fine-tuning is not necessary for there to be a cry for explanation. One

can imagine a cosmological theory that contains a large number of free parameters but is not fine-tuned because life does not sensitively depend on the values assigned to these parameters.

The easy part of the answer is therefore: Yes, fine-tuning cries out for explanation to the extent to which it is correlated with an excess of free parameters and a resultant lack of simplicity.[2] This part of the answer has been overlooked in discussions of fine-tuning, yet it is important to separate out this aspect in order to rightly grasp the more problematic part to which we shall now turn. The problematic part is to address the question of whether fine-tuning *especially* cries out for explanation, beyond the general desideratum of avoiding unnecessary complications and ad hoc assumptions. In other words, is *the fact that the universe would have been lifeless* if the values of fundamental constants had been very slightly different (assuming this is a fact) relevant in assessing whether an explanation is called for of why the constants have the values they have? And does it give support to the multiverse hypothesis? Or, alternatively, to the design hypothesis? The rest of this chapter will focus on these questions (though the design hypothesis will be discussed only as it touches on the other two questions).

Let's begin by critically examining some answers given in the literature.

No "Inverse Gambler's Fallacy"

Can an anthropic argument based on an observation selection effect together with the assumption that an ensemble of universes exists explain the apparent fine-tuning of our universe? Ian Hacking has argued that this depends on the nature of the ensemble. If the ensemble consists of all possible big-bang universes (a position he ascribes to Brandon Carter) then, says Hacking, the anthropic explanation works:

> Why do we exist? Because we are a possible universe [sic], and all possible ones exist. Why are we in an orderly universe? Because the only universes that we could observe are orderly ones that support our form of life . . . nothing is left to chance. Everything in this reasoning is deductive. (Hacking 1987), p. 337

---

[2] The simplicity principle I'm using here is not that every phenomenon must have an explanation (which would be version of the principle of sufficient reason, which I do not accept). Rather, what I mean is that we have an a priori epistemic bias in favor of hypotheses which are compatible with us living in a relatively simple world. Therefore, if our best account so far of some phenomenon involves very non-simple hypotheses (such as that a highly remarkable coincidence happened just by chance), then we may have prima facie reason for thinking that there is some better (simpler) explanation of the phenomenon that we haven't yet thought of. In that sense, the phenomenon is crying out for an explanation. Of course, there might not be a (simple) explanation. But we shouldn't be willing to believe in the complicated account until we have convinced ourselves that no simple explanation would work.

Hacking contrasts this with a seemingly analogous explanation that seeks to explain fine-tuning by supposing that a Wheeler-type multiverse exists. In the Wheeler cosmology, there is a never-ending sequence of universes each of which begins with a big bang and ends with a big crunch which bounces back in a new big bang, and so forth. The values of physical constants are reset in a random fashion in each bounce, so that we have a vast ensemble of universes with varying properties. The purported anthropic explanation of fine-tuning based on such a Wheeler ensemble notes that, given that the ensemble is large enough, it could be expected to contain at least one fine-tuned universe like ours. An observation selection effect can be invoked to explain why we observe a fine-tuned universe rather than one of the non-tuned ones. On the face of it, this line of reasoning looks very similar to the anthropic reasoning based on the Carter multiverse, which Hacking endorses. But according to Hacking, there is a crucial difference. He thinks that the version using the Wheeler multiverse commits a terrible mistake, which he dubs the "Inverse Gambler's Fallacy". This is the fallacy of a dim-witted gambler who thinks that the apparently improbable outcome he currently observes is made more probable if there have been many trials preceding the present one.

> [A gambler] enters the room as a roll is about to be made. The kibitzer asks, 'Is this the first role of the dice, do you think, or have we made many a one earlier tonight? . . . slyly, he says 'Can I wait until I see how this roll comes out, before I lay my bet with you on the number of past plays made tonight?' The kibitzer . . . agrees. The roll is a double six. The gambler foolishly says, 'Ha, that makes a difference—I think there have been quite a few rolls.' (Hacking 1987), p. 333

The gambler in this example is clearly in error. But so is Hacking in thinking that the situation is analogous to the one regarding fine-tuning. As pointed out by three authors (Leslie 1988; McGrath 1988; Whitaker 1988) independently replying to Hacking's paper, there is no observation selection effect in his example—an essential ingredient in the purported anthropic explanation of fine-tuning.

One way of introducing an observation selection effect in Hacking's example is by supposing that the gambler has to wait outside the room until a double six is rolled. Knowing that this is the setup, the gambler does obtain some reason upon entering the room and seeing the double six for thinking that there probably have been quite a few rolls already. This is a closer analogy to the fine-tuning case. The gambler can only observe certain outcomes—we can think of these as the "fine-tuned" ones—and upon observing a fine-tuned outcome he obtains reason to think that there have been several trials. Observing a double six would then be surprising on the

hypothesis that there were only one roll, but it would be expected on the hypothesis that there were very many. Moreover, a kind of *explanation* of why the gambler is seeing a double six is provided by pointing out that there were many rolls and the gambler would be let in to observe the outcome only upon rolling a double six.

When we make the kibitzer example more similar to the fine-tuning situation, we thus find that it supports, rather than refutes, the analogous reasoning based on the Wheeler cosmology.

What makes Hacking's position especially peculiar is that he thinks that the anthropic reasoning works with a Carter multiverse but not with a Wheeler multiverse. Many think the anthropic reasoning works in both cases, some think it doesn't work in either case, but Hacking is probably alone in thinking it works in one but not the other. The only pertinent difference between the two cases seems to be that in the Carter case one *deduces* the existence of a universe like ours whereas in the Wheeler case one infers it probabilistically. The Wheeler case can be made to approximate the Carter case by having the probability that a universe like ours should be generated in some cycle be close to 1 (which is, incidentally, the case in the Wheeler scenario if there are infinitely many cycles and there is a fixed finite probability in each cycle of a universe like ours resulting). It is hard to see the appeal of a doctrine that drives a methodological wedge between the two cases by insisting that the anthropic explanation works perfectly in one and fails completely in the other.

ROGER WHITE AND PHIL DOWE'S ANALYSIS

Recently, a more challenging attack on the anthropic explanation of fine-tuning has been made by Roger White (White 2000) and Phil Dowe (Dowe 1998). They eschew Hacking's doctrine that there is an essential difference between the Wheeler and the Carter multiverses as regards the prospects for a corresponding anthropic fine-tuning explanation. But they take up another idea of Hacking's, namely that what goes wrong in the Inverse Gambler's Fallacy is that the gambler fails to take into account the most specific version of the explanandum that he knows when making his inference to the best explanation. If all the gambler had known were that *a* double six had been rolled, then it need not have been a fallacy to infer that there probably were quite a few rolls, since that would have made it more probable that there would be at least one double six. But the gambler knows that *this* roll, the latest one, was a double six; and that gives him no reason to believe there were many rolls, since the probability that that specific roll would be a double six is one in thirty-six independently of how many times the dice have been rolled before. So Hacking argues that when seeking an explanation, we must use the most specific rendition of the explanandum is in our knowledge:

> If F is known, and E is the best explanation of F, then we are supposed to infer E. However, we cannot give this rule carte blanche. If F is known, then FvG is known, but E* might be the best explanation of FvG, and yet knowledge of F gives not the slightest reason to believe E*. (John, an excellent swimmer, drowns in Lake Ontario. Therefore he drowns in either Lake Ontario or the Gulf of Mexico. At the time of his death, a hurricane is ravaging the Gulf. So the best explanation of why he drowned is that he was overtaken by a hurricane, which is absurd.) We must insist that F, the fact to be explained, is the most specific version of what is known and not a disjunctive consequence of what is known. (Hacking 1987), p. 335

Applying this to fine-tuning, Hacking, White, and Dowe charge that the purported anthropic explanation of fine-tuning fails to explain the most specific version of what is known. We know not only that *some* universe is fine-tuned; we know that *this* universe is fine-tuned. Now, if our explanandum is, why is *this* universe fine-tuned? (where "this universe" is understood rigidly) then it would seem that postulating many universes cannot move us any closer to explaining that; nor would it make the explanandum more probable. For how could the existence of many other universes make it more likely that this universe be fine-tuned?

At this stage it is useful to introduce some abbreviations. In order to focus on the point that White and Dowe are making, we can make some simplifying assumptions.[3] Let us suppose that there are n possible configurations of a big bang universe $\{T_1, T_2, \ldots, T_n\}$ and that they are equally "probable", $P(T_i) = 1/n$. We assume that T_1 is the only configuration that permits life to evolve. Let x be a variable that ranges over the set of actual universes. We assume that each universe instantiates a unique T_i, so that $\forall x \exists! i(T_i x)$. Let be the number of actually existing universes, and let "α" rigidly denote our universe. We define

$E := T_1\alpha$	("α is life-permitting.")
$E' := \exists x (T_1 x)$	("Some universe is life-permitting.")
$M := m>>0$	("There are many universes."—the multiverse hypothesis)

White claims that, while there being many universes increases the probability that there is a life-permitting universe, $(P(E' \mid M) > P(E' \mid \neg M))$, it is not the case that there being many universes increases the probability that our universe is life-permitting. That is, $P(E \mid M) = P(E \mid \neg M) = 1/n$. The argument White gives for this is that

[3] I will adopt White's formalism to facilitate comparison. The simplifying assumptions are also made by White, on whose analysis we focus since it is more detailed than Dowe's.

> the probability of [E, i.e. the claim that α instantiates T1] is just 1/n, regardless of how many other universes there are, since α's initial conditions and constants are selected randomly from a set of n equally probable alternatives, a selection which is independent of the existence of other universes. The events which give rise to universes are not causally related in such a way that the outcome of one renders the outcome of another more or less probable. They are like independent rolls of a die. (White 2000), pp. 262–3

Since we should conditionalize on the most specific information we have when evaluating the support for the multiverse hypothesis, and since E is more specific than E', White concludes that our knowledge that our universe is life-permitting gives us no reason to think there are many universes.

This argument has some initial plausibility. Nonetheless, I think it is fallacious. We get a strong hint that something has gone wrong if we pay attention to a certain symmetry. Let α, β_1, . . . , β_{m-1} be the actually existing universes, and for $i = \alpha$, β_1. . . , β_{m-1}, let E_i be the proposition that if some universe is life-permitting then i is life-permitting. Thus, E is equivalent to the conjunction of E' and E_α. According to White, if all we knew was E' then that would count as evidence for M; but if we know the more specific E then that is not evidence for M. So he is committed to the following ((White 2000), p. 264):

$P(M \mid E') > P(M)$, and

$P(M \mid E) = P(M)$

Since by definition $P(M \mid E'E_\alpha) = P(M \mid E)$, this implies:

$$P(M \mid E'E_\alpha) < P(M \mid E') \qquad (*)$$

Because of the symmetry of the β_j:s, $P(M \mid E'E_{\beta j}) = c$, for every β_j, for no ground has been given for why *some* of the universes β_j would have given more reason, had it been the fine-tuned, for believing M, than would any other β_j similarly fine-tuned. Since E' implies the disjunction $E'E_\alpha \vee E'E_{\beta 1} \vee E'E_{\beta 2} \vee \ldots \vee E'E_{m-1}$, this together with (*) implies:

$$P(M \mid E'E_{\beta j}) > P(M \mid E') \text{ for every } \beta_j \qquad (**)$$

In other words, White is committed to the view that, given that some uni-

verse is life-permitting, then: conditionalizing on α being life-permitting *decreases* the probability of M, while conditionalizing on any of $\beta_1 \ldots, \beta_{m-1}$, *increases* the probability of M.

But that seems wrong. Given that some universe is life-permitting, why should the fact it is *this* universe that is life-permitting, rather than any of the others, lower the probability that there are many universes? If it had been some other universe instead of this one that had been life-permitting, why should that have made the multiverse hypothesis any more likely? Clearly, such discrimination could be justified only if there were something special that we knew about *this* universe that would make the fact that it is this universe rather than some other that is life-permitting significant. I can't see what sort of knowledge that would be. It is true that *we* are in this universe and not in any of the others—but that fact *presupposes* that this universe is life-permitting. It is not as if there is a remarkable coincidence between our universe being life-permitting and us being in it. So it's hard to see how the fact that we are in this universe could justify treating its being life-permitting as giving a lower probability to the multiverse hypothesis than any other universe's being life-permitting would have.

So what, precisely, is wrong in White's argument? His basic intuition for why P(M | E) = P(M) seems to be that "The events which give rise to universes are not causally related in such a way that the outcome of one renders the outcome of another more or less probable.". Yet a little reflection reveals that this assertion is highly problematic for several reasons.

First, there's no empirical warrant for it. Very little is yet known about the events which give rise to universes. There are models on which the outcomes of some such events *do* causally influence the outcome of others. To illustrate, in Lee Smolin's (admittedly highly speculative) evolutionary cosmological model (Smolin 1997), universes create "baby-universes" whenever a black hole is formed, and these baby-universes inherit, in a somewhat stochastic manner, some of the properties of their parent. The outcomes of chance events in one such conception can thus influence the outcomes of chance events in the births of other universes. Variations of the Wheeler oscillating universe model have also been suggested where some properties are inherited from one cycle to the next. And there are live speculations that it might be possible for advanced civilizations to spawn new universes and transfer some information into them by determining the values of some of their constants (as suggested by Andrei Linde, of inflation theory fame), by tunneling into them through a wormhole (Morris, Thorne et al. 1988), or otherwise (Ćirković and Bostrom 2000; Garriga, Mukhanov et al. 2000).

Even if the events which give rise to universes are not causally related in the sense that the outcome of one event causally influences the outcome of another (as in the examples just mentioned), that does not mean that one universe cannot carry information about another. For instance, two universes can have a partial cause in common. This is the case in the mul-

tiverse models associated with inflation theory (arguably the best current candidates for a multiverse cosmology). In a nutshell, the idea is that universes arise from inflating fluctuations in some background space. The existence of this background space and the parameters of the chance mechanism that lead to the creation of inflating bubbles are at least partial causes of the universes that are produced. The properties of the produced universes could thus carry information about this background space and the mechanism of bubble creation, and hence indirectly also about other universes that have been produced by the same mechanism. The majority of multiverse models that have actually been proposed, including arguably the most plausible one, directly negate White's claim.

Second, even if we consider the hypothetical case of a multiverse model where the universes bear no causal relations to one another, it is *still* not generally the case that $P(M|E) = P(M)$. This holds even setting aside any issues related to anthropic reasoning. We need to make a distinction between objective chance and epistemic probability. If there is no causal connection (whether direct or indirect via a common cause) between the universes, then there is no correlation in the physical chances of the outcomes of the events in which these universes are created. It does not follow that the outcomes of those events are uncorrelated in one's rational epistemic probability assignment. Consider this toy example:

> Suppose you have some background knowledge K and that your prior subjective probability function P, conditionalized on K, assigns non-negligible probability to only three possible worlds and assigns an equal probability to these: $P(w_1|K) = P(w_2|K) = P(w_3|K)$. In w_1 there is one big universe, *a*, and one small universe, *d*; in w_2 there is one big, *b*, and one small, *e*; and in w_3 there is one big, *c*, and one small, *e*. Now suppose you learn that you are in universe *e*. This rules out w_1. It thus gives you information about the big universe—it is now more likely to be either *b* or *c* than it was before you learnt that the little universe is *e*. That is, P("The big universe is *b* or *c*"|K&"The little universe is *e*") > P("The big universe is *b* or *c*"|K).

No assumption whatever is made here about the universes being causally related. White presupposes that any such subjective probability function P must be irrational or unreasonable (independently of the exact nature of the various possible worlds under consideration). Yet that seems implausible. Certainly, White provides no argument for it.

Third, White's view that $P(M|E') > P(M)$ seems to commit him to denying just this assumption. For how could E' (which says that some universe is life-permitting) be probabilistically relevant to M unless the outcome of one universe-creating event *x* (namely that event, or one of those events, that created the life-permitting universe(s)) can be probabilistically relevant

to the outcome of another *y* (namely one of those events that created the universes other than *x*)? If *x* gives absolutely no information about *y*, then it is hard to see how knowledge that there is some life-permitting universe, the one created by *x*, could given us grounds for thinking that there are many other universes, such as the one created by *y*. So on this reasoning, it seems we would have P(M | E') = P(M), pace White.

This last point connects back to our initial observation regarding the symmetry and the implausibility of thinking that because it is *our* universe that is life-permitting there is less support for the multiverse hypothesis than if it had been some other universe instead that were life-permitting. All these problems are avoided if we acknowledge that that not only P(M | E') > P(M) but also P(M | E) > P(M).

I conclude that White's argument against the view that fine-tuning lends some support to the multiverse hypothesis fails. And so do consequently Phil Dowe's and Ian Hacking's arguments, the latter failing on other accounts as well, as we have seen.

SURPRISING VS. UNSURPRISING IMPROBABLE EVENTS

If, then, the fact that our universe is life-permitting *does* give support to the multiverse hypothesis, i.e. P(M | E) > P(M), it follows from Bayes' theorem that P(E | M) > P(E). How can the existence of a multiverse make it more probable that *this* universe should be life-permitting? One may be tempted to say: By making it more likely that this universe should exist. The problem with this reply is that it would seem to equally validate the inference to many universes from any sort of universe whatever. For instance, let E* be the proposition that α is a universe that contains nothing but chaotic light rays. It seems wrong to think that P(M | E*) > P(M). Yet, if the only reason that P(E | M) > P(E) is that α is more likely to exist if M is true,

[4] Some authors who are skeptical about the claim that fine-tuning is evidence for a multiverse still see a potential role of an anthropic explanation using the multiverse hypothesis as a way of reducing the surprisingness or amazingness of the observed fine-tuning. A good example of this tack is John Earman's paper on the anthropic principle (Earman 1987), in which he criticizes a number of illegitimate claims made on behalf of the anthropic principle by various authors (especially concerning those misnamed "anthropic principles" that don't involve any observation selection effects and hence bear little or no relation to Brandon Carter's original ideas on the topic (Carter 1974; Carter 1983; Carter 1989; Carter 1990). But in the conclusion he writes: "There remains a potentially legitimate use of anthropic reasoning to alleviate the state of puzzlement into which some people have managed to work themselves over various features of the observable portion of our universe. . . . But to be legitimate, the anthropic reasoning must be backed by substantive reasons for believing in the required [multiverse] structure." (p. 316). Similar views are espoused by Ernan McMullin (McMullin 1993), Bernulf Karnitscheider (Kanitscheider 1993), and (less explicitly) by George Gale (Gale 1996). I agree that anthropic reasoning reduces puzzlement only given the existence of a suitable multiverse, but I disagree with the claim that the potential reduction of puzzlement is no ground whatever for thinking that the multiverse hypothesis is true. My reasons for this will become clear as we proceed.

then an exactly analogous reason would support $P(E^*|M) > P(E^*)$, and hence $P(M|E^*) > P(M)$. This presents the anthropic theorizer with a puzzle. Somehow, the "life-containingness" of α must be given a role to play in the anthropic account. But how can that be done?

Several prominent supporters of the anthropic argument for the multiverse hypothesis have sought to base their case on a distinction between events (or facts) that are surprising and ones that are improbable but not surprising (see e.g. John Leslie (Leslie 1989) and Peter van Inwagen (van Inwagen 1993).[4]

Suppose you toss a coin one hundred times and write down the results. Any particular sequence *s* is highly improbable ($P(s) = 2^{-100}$), yet most sequences are not surprising. If *s* contains roughly equally many heads and tails in no clear pattern then *s* is improbable and unsurprising. By contrast, if *s* consists of 100 heads, or of alternating heads and tails, or some other highly patterned outcome, then *s* is surprising. Or to take another example, if *x* wins a lottery with one billion tickets, this is said to be unsurprising ("someone had to win . . . it could just as well be *x* as anybody else . . . shrug."); whereas if there are three lotteries with a thousand tickets each, and *x* wins all three of them, this is surprising. We evidently have some intuitive concept of what it is for an outcome to be surprising in cases like these.

The idea, then, is that a fine-tuned universe is surprising in a sense in which a particular universe filled with only chaotic electromagnetic radiation would not have been. And that's why we need to look for an explanation of fine-tuning but would not have had any reason to suppose there were an explanation for a light-filled universe. The two potential explanations for fine-tuning that typically are considered are the design hypothesis and the multiple universe hypothesis. An inference is then made that at least one of these hypotheses is quite likely true in light of available data, or at least more likely true than would have been the case if this universe had been a "boring" one containing only chaotic light. This is similar to the 100 coin flips example. An unsurprising outcome does not lead us to search for an explanation, while a run of 100 heads does cry out for explanation and gives at least some support to potential explanations such as the hypothesis that the coin flipping process was biased. Likewise in the lottery example. The same person winning all three lotteries could make us suspect that the lottery had been rigged in the winner's favor.

A key assumption in this argument is that fine-tuning is indeed surprising. Is it? Some dismiss the possibility out of hand. For example, Stephen Jay Gould writes:

> Any complex historical outcome—intelligent life on earth, for example—represents a summation of improbabilities and becomes therefore absurdly unlikely. But something has to happen, even if any particular "something" must stun us by its improbability. We could look at any outcome

> and say, "Ain't it amazing. If the laws of nature had been set up a tad differently, we wouldn't have this kind of universe at all." (Gould 1990), p. 183

From the other side, Peter van Inwagen mocks that way of thinking:

> Some philosophers have argued that there is nothing in the fact that the universe is fine-tuned that should be the occasion for any surprise. After all (the objection runs), if a machine has dials, the dials have to be set some way, and any particular setting is as unlikely as any other. Since any setting of the dial is as unlikely as any other, there can be nothing more surprising about the actual setting of the dials, whatever it may be, than there would be about any possible setting of the dials if that possible setting were the actual setting. . . . This reasoning is sometimes combined with the point that if "our" numbers hadn't been set into the cosmic dials, the equally improbable setting that did occur would have differed from the actual setting mainly in that there would have been no one there to wonder at its improbability. (van Inwagen 1993), pp. 134–5

Opining that this "must be one of the most annoyingly obtuse arguments in the history of philosophy", van Inwagen asks us to consider the following analogy. Suppose you have to draw a straw from a bundle of 1,048,576 straws of different lengths. It has been decreed that unless you draw the shortest straw you will be instantly killed so that you don't have time to realize that you didn't draw the shortest straw. "Reluctantly—but you have no alternative—you draw a straw and are astonished to find yourself alive and holding the shortest straw. What should you conclude?" According to van Inwagen, only one conclusion is reasonable: that you did not draw the straw at random but that instead the situation was somehow rigged by an unknown benefactor to ensure that you got the shortest straw. The following argument to the contrary is dismissed as "silly":

> Look, you had to draw some straw or other. Drawing the shortest was no more unlikely than drawing the 256,057th-shortest: the probability in either case was .000000954. But your drawing the 256,057th-shortest straw isn't an outcome that would suggest a 'set-up' or would suggest the need for any sort of explanation, and, therefore, drawing the shortest shouldn't suggest the need for an explanation either. The only real difference between the two cases is that you wouldn't have been around to remark on the unlikelihood of drawing the 256,057th-shortest straw. (van Inwagen 1993), p. 135

Given that the rigging hypothesis did not have too low a prior probability and given that there was only one straw lottery, it is hard to deny that this argument would indeed be silly. What we need to ponder though,

is whether the example is analogous to our epistemic situation regarding fine-tuning.

Erik Carlson and Erik Olsson (Carlson and Olsson 1998), criticizing van Inwagen's argument, argue that there are three points of disanalogy between van Inwagen's straw lottery and fine-tuning.

First, they note that whether we would be willing to accept the "unknown benefactor" explanation after drawing the shortest straw depends on our prior probability of there being an unknown benefactor with the means to rig the lottery. If the prior probability is sufficiently tiny—given certain background beliefs it may be very hard to see how the straw lottery *could* be rigged—we would not end up believing in the unknown benefactor hypothesis. Obviously, the same applies to the fine-tuning argument: if the prior probability of a multiverse is small enough then we won't accept that hypothesis even after discovering a high degree of fine-tuning in our universe. The multiverse supporter can grant this and argue that the prior probability of a multiverse is not too small. Exactly how small it can be for us still to end up accepting the multiverse hypothesis depends on both how extreme the fine-tuning is and what alternative explanations are available. If there is plenty of fine-tuning, and the only alternative explanation on the table is the design hypothesis, and if that hypothesis is assigned a much lower prior probability than the multiverse hypothesis, then the argument for the multiverse hypothesis would be vindicated. We don't need to commit ourselves to these assumptions; and in any case, different people might have different prior probabilities. What we are primarily concerned with here is to determine whether fine-tuning is in a relevant sense a *surprising* improbable event, and whether taking fine-tuning into account should substantially *increase* our credence in the multiverse hypothesis and/or the design hypothesis, not what the absolute magnitude of our credence in those hypotheses should be. Carlson and Olsson's first point is granted but it doesn't have any bite. Van Inwagen never claimed that his straw lottery example could settle the question of what the prior probabilities should be.

Carlson and Olsson's second point would be more damaging for van Inwagen, if it weren't incorrect. They claim that there is a fundamental disanalogy in that we understand at least roughly what the causal mechanisms are by which intelligent life evolved from inorganic matter, whereas no such knowledge is assumed regarding the causal chain of events that led you to draw the shortest straw. To make the lottery more closely analogous to the fine-tuning, we should therefore add to the description of the lottery example that at least the proximate causes of your drawing the shortest straw are known. Carlson and Olsson then note that:

> In such a straw lottery, our intuitive reluctance to accept the single-drawing-plus-chance hypothesis is, we think, considerably diminished. Suppose that we can give a detailed causal explanation of why you drew

> the shortest straw, starting from the state of the world twenty-four hours before the drawing. A crucial link in this explanation is the fact that you had exactly two pints of Guinness on the night before the lottery. . . . Would you, in light of this explanation of your drawing the shortest straw, conclude that, unless there have been a great many straw lotteries, somebody intentionally caused you to drink two pints of Guinness in order to ensure that you draw the shortest straw? . . . To us, this conclusion does not seem very reasonable. (Carlson and Olsson 1998), pp. 271–2

The objection strikes me as unfair. Obviously, if you knew that your choosing the shortest straw depended crucially and sensitively on your precise choice of beverage the night before, you would feel disinclined to accept the rigging hypothesis. That much is right. But this disinclination is fully accounted for by the fact that it is tremendously hard to see, under such circumstances, how anybody *could* have rigged the lottery. If we knew that successful rigging required predicting in detail such a long and tenuous causal chain of events, we could well conclude that the prior probability of rigging was negligible. For *that* reason, surviving the lottery would not make us believe the rigging hypothesis.

We can see that it is this—rather than our understanding of the proximate causes per se—that defeats the argument for rigging by considering the following variant of van Inwagen's example. Suppose that the straws are scattered over a vast area. Each straw has one railway track leading up to it, and all the tracks start from the same central station. When you pick the shortest straw, we now have a causal explanation that can stretch far back in time: you picked it because it was at the destination point of a long journey along a track that did not branch. How long the track was makes no difference to how willing we are to believe in the rigging hypothesis. What matters is only whether we think there is some plausibility to the idea that an unknown benefactor could have put you on the right track to begin with. So contrary to what Carlson and Olsson imply, what is relevant is not the known backward length of the causal chain, but whether that chain would have been sufficiently predictable by the hypothetical benefactor to give a large enough prior probability to the hypothesis that she rigged the lottery. Needless to say, the designer referred to in the design hypothesis is typically assumed to have superhuman epistemic capacities. It is not at all farfetched to suppose that *if* there were a cosmic designer, she would have been able to anticipate which boundary conditions of the universe were likely to lead to the evolution of life. We should therefore reject Carlson and Olsson's second objection against van Inwagen's analogy.

The third alleged point of disanalogy is somewhat subtler. Carlson and Olsson discuss it in the context of refuting certain claims by Arnold Zuboff (Zuboff 1991) and it is not clear how much weight they place on it as an objection against van Inwagen. But it's worth mentioning. The idea, as far as I can make it out, is that the reason why your existing after the straw

lottery is surprising, is related to the fact that you existed before the straw lottery. You could have antecedently contemplated your survival as one of a variety of possible outcomes. In the case of fine-tuning, by contrast, your existing (or intelligent life existing) is not an outcome which could have been contemplated prior to its obtaining.

> For conceptual reasons, it is impossible that you know in advance that your existence lottery is going to take place. Likewise, it is conceptually impossible that you make any ex ante specification of any possible outcome of this lottery. . . . The existence of a cosmos suitable for life does not seem to be a coincidence for anybody; nobody was ever able to specify this outcome of the cosmos lottery, independently of its actually being the actual outcome. (Carlson and Olsson 1998), p. 268

This might look like a token of the "annoyingly obtuse" reasoning that van Inwagen thought to refute through his straw lottery example. Nevertheless, there is a disanalogy between the two cases: nobody could have contemplated the existence of intelligent life unless intelligent life existed, whereas someone (even the person immediately involved) could have thought about drawing the shortest straw before drawing it. The question is whether this difference is relevant. Again it is useful to cook up a variant of the straw-drawing example:

> Suppose that in an otherwise lifeless universe there is a big bunch of straws and a simple (non-cognitive, non-conscious) automaton is about to randomly select one of the straws. There is also kind of "incubator" in which one person rests in an unconscious state; we can suppose she has been unconscious since the beginning of time. The automaton is set up in such a way that the person in the incubator will be woken if and only if the automaton picks the shortest straw. You wake up in the incubator. After examining your surroundings and learning about how the experiment was set up, you begin to wonder about whether there's anything surprising about the fact that the shortest straw was drawn.

This example shares with the fine-tuning case the feature that nobody would have been there to contemplate anything if the "special" outcome had failed to obtain. So what should we say about this case? In order for Carlson and Olsson's criticism to work, we would have to say that the person waking up in the incubator should not think that there is anything surprising at all about the shortest straw having been selected. Van Inwagen would, presumably, simply deny that that would be the correct attitude. For what it's worth, my intuition in this instance sides with van Inwagen,

although this case is perhaps less obvious than the original straw lottery gedanken where the subject had a life before the lottery.

It would be nice to have an independent account of what makes an event or a fact surprising. We could then apply the general account to the straw lotteries or directly to fine-tuning, and see what follows. Let us therefore briefly review what efforts have been made to develop such an account of surprisingness. (I'm indebted here to the literature-survey and discussion in (Manson 1998).) To anticipate the upshot: I will argue that these are dead ends as far as anthropic reasoning is concerned. The strategy relied on by those anthropic theorizers who base their case on an appeal to what is surprising is therefore ultimately of very limited utility: the strategy is based on intuitions that are no more obvious or secure than the thesis which they are employed to support. This may seem disappointing, but in fact it clears the path for a better understanding what is required to support anthropic reasoning.

The following remark by F. P. Ramsey is pertinent to the goal of determining what distinguishes surprising improbable events from unsurprising improbable events:

> What we mean by an event not being a coincidence, or not being due to chance, is that if we came to know it, it would make us no longer regard our system as satisfactory, although on our system the event may be no more improbable than any alternative. Thus 1,000 heads running would not be due to chance; i.e. if we observed it we should change our system of chances for that penny. (Ramsey 1990), p. 106

This looks like an auspicious beginning. It seems to fit the other example considered near the beginning of this section: one person winning three lotteries with a thousand tickets could make us suspect foul play, whereas one person winning a billion-ticket lottery would not in general have any tendency do so. Or ponder the case of a monkey typing out the sequence "Give me a banana!". This is surprising and it makes us change our belief that the monkey types out a random sequence. We would think that maybe the monkey had been trained to type that specific sequence, or maybe that the typewriter was rigged; but the chance hypothesis is disconfirmed. By contrast, if the monkey types out "r78o479024io; jl;", this is unsurprising and does not challenge our assumptions about the setup. So far so good.

What Ramsey's suggestion does not tell us is what it is about events such as the monkey's typing a meaningful sentence or the run of 1000 heads that makes us change our minds about the system of chances. And we need to know that if the suggestion is to throw light on the fine-tuning case. For the problem there is precisely that it is not immediately clear—lest the question be begged—whether we ought to change our system and

find some alternative explanation or be satisfied with letting chance pay the bill by regarding fine-tuning as a coincidence. Ramsey's suggestion is thus insufficient for the present purpose.

Paul Horwich takes the analysis a little further. He proposes the following as a necessary condition for the truth of a statement E being surprising:

> [T]he truth of E is surprising only if the supposed circumstances C, which made E seem improbable, are themselves substantially diminished in probability by the truth of E . . .and if there is some initially implausible (but not widely implausible) alternative view K about the circumstances, relative to which E would be highly probable. (Horwich 1982), p. 101

If we combine this with the condition that "our beliefs C are such as to give rise to ", we get what Horwich thinks is a necessary and sufficient condition for the truth of a statement being surprising. We can sum this up by saying that the truth of E is surprising iff the following holds:

(i) $P(E) \approx 0$

(ii) $P(C \mid E) << P(C)$

(iii) $P(E \mid K) \approx 1$

(iv) $P(K)$ is small but not too small

Several authors who think that fine-tuning cries out for explanation endorse views that are similar to Horwich's (Manson 1989). For instance, van Inwagen writes:

> Suppose there is a certain fact that has no known explanation; suppose that one can think of a possible explanation of that fact, an explanation that (if only it were true) would be a very *good* explanation; then it is wrong to say that that event stands in no more need of an explanation than an otherwise similar event for which no such explanation is available. (van Inwagen 1993), p. 135

And John Leslie:

> A chief (or the only?) reason for thinking that something stands in [special need for explanation], i.e. for justifiable reluctance to dismiss it as how things just happen to be, is that one in fact glimpses some tidy way in which it might be explained. (Leslie 1989), p. 10

D. J. Bartholomew also appears to support a similar principle (Bartholomew 1984). Horwich's analysis provides a reasonably good explication of these ideas.

George Schlesinger (Schlesinger 1991) has criticized Horwich's analysis, arguing that the availability of a tidy explanation is not necessary for an event being surprising. Schlesinger asks us to consider the case of a tornado that touches down in three different places, destroying one house in each place. We are surprised to learn that these houses belonged to the same person and that they are the only buildings that this misfortunate capitalist owned. Yet no neat explanation suggests itself. Indeed, it seems to be *because* we can see no tidy explanation (other than the chance hypothesis) that this phenomenon would be so surprising. So if we let E to be the event that the tornado destroys the only three buildings that some person owns and destroys nothing else, and C the chance hypothesis, then (ii)–(iv) are not satisfied. According to Horwich's analysis, E is not surprising—which seems wrong.

Surprise being ultimately a psychological matter, we should perhaps not expect any simple definition to perfectly capture all the cases where we would feel surprised. But maybe Horwich has provided at least a sufficient condition for when we ought to feel surprised? Let's run with this for a second and see what happens when we apply his analysis to fine-tuning.

In order to do this we need to determine the probabilities referred to in (i)–(iv). Let's grant that the prior probability of fine-tuning (E) is very small, $P(E) \approx 0$. Further, anthropic theorizers maintain that E makes the chance hypothesis substantially less probable than it would have been without conditionalizing on E, so let's suppose that $P(C|E) << P(C)$[5]. Let K be a multiverse hypothesis. In order to have $P(C|K) \approx 1$, it might be necessary to think of K as more specific than the proposition that there is some multiverse; we may have to define K as the proposition that there is a "suitable" multiverse (i.e. one such that $P(E|K) \approx 1$ is satisfied). But let us suppose that even such a strengthened multiverse hypothesis has a prior probability that is not "too small". If we make these assumptions then Horwich's four conditions are satisfied, and the truth of E would consequently be surprising. This is the result that the anthropic theorizer would welcome.

Unfortunately, we can construct a similar line of assumptions to show that any other possible universe would have been equally surprising. Let $E^{\#}$ be the proposition that α has some particular boring character. For instance, we can let $E^{\#}$ say that α is a universe which consists of nothing but such-and-such a pattern of electromagnetic radiation. We then have $P(E^{\#}) \approx 0$. We can let K be the same as before. Now, if we suppose that P

[5] This follows from Bayes' theorem if the probability that C gives to E is so tiny that $P(E|C) << P(E)$.

$(C|\, E^{\#}) << P(C)$ and $P(E^{\#}|\, K) \approx 1$ then the truth of $E^{\#}$ will be classified as surprising. This is counterintuitive. And if it were true that every possible universe would be just as surprising as any other then fine-tuning being surprising can surely not be what legitimizes the inference from fine-tuning to the multiverse hypothesis. We must therefore deny either $P(C|\, E^{\#}) << P(C)$ or $P(E^{\#}|\, K) \approx 1$ (or both). At the same time, if the truth of E is to be surprising, we must maintain that $P(C|\, E) << P(C)$ and $P(E|\, K) \approx 1$. This means that the anthropic theorizer wishing to ground her argument in an appeal to surprise must treat $E^{\#}$ differently from E as regards these conditional probabilities. It may be indeed be correct to do that. But what is the justification? Whatever is it, it cannot be that the truth of E is surprising whereas the truth of $E^{\#}$ is not. For although that might be true, to simply assume it would be to make the argument circular.

The appeal to the surprisingness of E is therefore quite ineffective. In order to give the appeal any force, it needs to be backed up by some argument for the claim that: $P(C|\, E) << P(C)$, $P(E|\, K) \approx 1$ but not both $P(C|\, E^{\#}) << P(C)$ and $P(E^{\#}|\, K) \approx 1$. But suppose we had such an argument. We could then sidestep considerations about surprisingness altogether! For it follows already from $P(E|\, K) \approx 1$, $P(E) \approx 0$, and P(K) being "not too small", that $P(K|\, E) \approx 1$, i.e. that fine-tuning is strong evidence for the multiverse hypothesis. (To see this, simply plug the values into Bayes' formula, $P(K|\, E) = P(E|\, K)\, P(K)\, /P(E)$.)

To make progress beyond this point, I think we need to abandon vague talk of what makes events surprising and focus explicitly on the core issue, which is to determine the conditional probability of the multiverse hypothesis/chance hypothesis/design hypothesis given the evidence we have. If we figure out how to think about these conditional probabilities then we can hopefully use this insight to sort out the quandary about whether fine-tuning should be regarded as surprising. At any rate, that quandary becomes much less important if we have a direct route to assigning probabilities to the relevant hypotheses that skips the detour through the dark netherworld of amazement and surprise. This is what we shall now do.

MODELING OBSERVATION SELECTION EFFECTS: THE ANGEL PARABLE

I submit that the only way to get a plausible model of how to reason from fine-tuning is by explicitly taking observation selection effects into account. This section will outline parts of such a theory. Later chapters will expand and support themes that are merely alluded to here. A theory of observation selection effects has applications in many domains. In this section we focus on applications in cosmology.

As before, let "α" rigidly denote our universe. We know some things K about α (it's life-permitting; it contains the Eiffel tower; it's quite big etc.). Let h_M be the multiverse hypothesis; let h_D be the design hypothesis; and let h_C be the chance hypothesis. In order to determine what values to

assign to the conditional probabilities $P(h_M | K)$, $P(h_D | K)$, and $P(h_C | K)$, we need to take account of the observation selection effects through which our evidence about the world has been filtered.

How should we model these observation selection effects? Suppose that you are an angel. So far nothing physical exists, but six days ago God told you that he was going away for a week to create a cosmos. He might create either a single universe or a multiverse, and let's say your prior probabilities for these two hypotheses are about 50%. Now a messenger arrives and informs you that God's work is completed. The messenger tells you that universe α exists but does not say whether there are other universes in addition. Should you think that God created a multiverse or only α? To answer this, we need to know something more about the situation. Consider two possible stories of what happened:

> *Case 1.* The messenger decided to travel to realm of physical existence and look at the universe or one of the universes that God had created. This universe was α, and this is what he reports to you.
>
> *Case 2.* The messenger decided to find out whether God created α. So he travels to the realm of physical existence and looks until he finds α, and reports this back to you.

In Case 1, the messenger's tidings do not in general give you any reason to believe h_M. He was bound to bring back news about some universe, and the fact that he tells you about α rather than some other universe is not significant, *unless* α has some special feature F. (More on this proviso shortly.)

In Case 2 on the other hand, the fact that the messenger tells you that α exists is evidence for h_M. If the messenger selected α randomly from the class of all possible universes, or from some sizeable subclass thereof (for example only big bang universes with the same laws of nature as in our universe, or only universes which contain more good than evil), then the finding that God created α suggests that God created many universes.

Our actual epistemic situation is not analogous to the angel's in Case 2. It is not as if we first randomly selected α from a class containing both actual and non-actual possible universes and then discovered that—lo and behold!—α actually exists. The fact that we know whether α exists surely has everything to do with it actually existing and we being among its inhabitants. There is an observation selection effect amounting to the following: direct observation occurs only of universes that actually exist. Case 1 comes closer to modeling our epistemic situation in this respect, since it mirrors this selection effect.

However, Case 1 is still an inadequate model because it overlooks

another observational effect. The messenger could have retrieved information about any of the actual universes, and the angel could have found out about some universe β that doesn't contain any observers. If there are no angels, gods or heavenly messengers, however, then universes that don't contain observers are not observed. Assuming the absence of extramundane observers, the selection effect restricts what is observed not only to the extent that non-actual universes are not observed but actual universes that don't contain any observers are also not observed. This needs to be reflected in our model. If we want to continue to use the creation story, we therefore need to modify it as follows:

> *Case 3.* The messenger decided to travel to the realm of physical existence and look for some universe that contains observers. He found α, and reports this back to you.

Does this provide you with any evidence for h_M? It depends. If you knew (call this *Case 3a*) that God had set out to create at least one observer-containing universe, then the tiding that α is actual does not give any support to h_M (unless you know that α has some special feature). Because then you were guaranteed to learn about the existence of some observer-containing universe or other and learning that it is α does not give any more evidence for h_M than if you had learnt about some other universe instead. The messenger's tidings T contain no relevant new information. The probably you assign to h_M remains unchanged. In Case 3a, therefore, $P(h_M | T) = P(h_M)$.

But there is second way of specifying Case 3. Suppose (*Case 3b*) that God did not set out especially to create at least one observer-containing universe, and that for any universe that He created there was only a fairly small chance that it would be observer-containing. In this case, when the messenger reports that God created the observer-containing universe α, you get evidence that favors h_M. For it is more probable on h_M than it is on $\neg h_M$ that one or more observer-containing universes should exist (one of which the messenger was then bound to bring you news about). Here, we therefore have $P(h_M | T) > P(h_M)$.

What is grounding T's support for h_M? I think it is best answered by saying not that T makes it more probable that α should exist, but rather that T makes it more probable that at least one observer-containing universe should exist. It is nonetheless true that h_M makes it more probable that α should exist. But this is not by itself the reason why h_M is to be preferred given our knowledge of the existence of α. If it were, then since the same reason operates in Case 3a, we would have to have concluded that h_M were favored in that case as well. For even though it was guaranteed in Case 3a that some observer-containing universe would exist, it was not guaranteed that it would be α. In Case 3a as well as in Case 3b, the exis-

tence of α was made more likely by h_M than by $\neg h_M$. If this should not lead us to favor h_M in Case 3a then the fact that the existence of is made more likely by h_M cannot be the whole story about why h_M is to be preferred in Case 3b.

So what is the whole story about this? This will become clearer as we proceed, but we can give at least the outlines now. In subsequent chapters we shall fill in important details and see some arguments for the claims we make here.

In a nutshell: although h_M makes it more probable that α should exist, h_M also makes it more probable that there are other observer-containing universes. And the greater the number of observer-containing universes, the smaller the probability that we should observe any particular one of them. These two effects balance each other. The result is that the messenger's tidings are evidence in favor of theories on which it is probable that at least one observer-containing universe would exist; but this evidence does not favor theories on which it is probable that there are *many* observer-containing universes over theories on which it is probable that there are merely a *few* observer-containing universes.

We can get an intuitive grasp of this if we consider a two-step procedure. Suppose the messenger first tells you that some observer-containing universe x exists. This rules out all hypotheses on which there would be no such universes; it counts against hypotheses on which it would be very unlikely that there are any observer-containing universes; and it favors hypotheses on which it would be very likely or certain that there is one or more observer-containing universes. In the second step, the messenger tells you that $x = \alpha$. This should not change your beliefs as to how many observer-containing universes there are (assuming you don't think there is anything special about α). One might say that if God were equally likely to create any universe, then the probability that α should exist is proportional to the number of universes God created. True. But the full evidence you have is not only that α exists but also that the messenger told you about α. If the messenger selected the universe he reports randomly from the class of all actual observer-containing universes, then the probability that he would select α, given that α is an actual observer-containing universe, is *inversely* proportional to the number of actual observer-containing universes. The messenger's report therefore does not allow you to discriminate between general hypotheses[6] that imply that at least one observer-containing universe exists.

---

[6] By "general hypotheses" we here mean: hypotheses that don't entail anything preferentially about α. For example, a hypothesis which says "There is exactly one life-containing universe and it's not α." will obviously be refuted by the messenger's report. But the point is that there is nothing about the messenger's report that gives reason to favor hypotheses only because they imply a greater number of observer-containing universes, assuming there is nothing special about α.

In our actual situation, our knowledge is not mediated by a messenger; but the idea is that the data we get about the world is subjected to observation selection effects that mimic the reporting biases present in Case 3. (Not quite, though. A better analogy yet would be one in which (*Case 4*) the messenger selects a random observer from among the observers that God has created, thus biasing the universe-selection in favor of those universes that have relatively large populations. But more on this in a later chapter. To keep things simple here, we can imagine all the observer-containing universes as having the same number of observers.)

When stating that the finding that α exists does not give us reason to think that there are many rather than few observer-containing universes, we have kept inserting the proviso that α not be "special". This is an essential qualification, for there clearly are some features F such that if we knew that α has them then finding that α exists *would* give support to the claim that there are a vast number of observer-containing universes. For instance, if you know that α is a universe in which a message is inscribed in every rock, in the distribution of fixed stars seen from any life-bearing planet, and in the microstructure of common crystal lattices, spelling: "God created this universe. He also created many other universes."—then the fact that the messenger tells you that α exists can obviously give you some reason to think that there are many universes. In our actual universe, if we were to find inscriptions that we were convinced could only have been created by a divine being then this would count as support for whatever these inscriptions asserted (the degree of support being qualified by the strength of our conviction that the deity was being honest). Leaving aside such theological scenarios, there are much more humdrum features our universe might have that could make it special in the sense intended here. It may be, for example, that the physics of our universe is such as to suggest a physical theory (because it's the simplest, most elegant theory that fits the facts) that entails the existence of vast numbers of observer-containing universes.

Fine-tuning may well be a "special" feature. This is so because fine-tuning seems to indicate that there is no simple, elegant theory which entails (or gives a high probability to) the existence our universe alone but not to the existence of other universes. If it were to turn out, present appearances notwithstanding, that there is such a theory, then our universe is not special. But in that case there would be little reason to think that our universe really is fine-tuned. For if a simple theory entails that precisely this universe should exist, then one could plausibly assert that no other boundary conditions than those implied by that theory are physically possible; and hence that physical constants and initial conditions could not have been different than they are—thus no fine-tuning. However, assuming that every theory fitting the facts and entailing that there is only one universe is a very ad hoc one and involving many free parameters—as fine-tuning advocates

argue—then the fine-tuning of our universe is a special feature that gives support to the hypothesis that there are many universes. There is nothing mysterious about this. Preferring simple theories that fit the facts to complicated ad hoc ones is just standard scientific practice, and cosmologists who work with multiverse theories are presumably pursuing that inquiry because they think that multiverse theories represent a promising route forward to neat theories that are empirically adequate.

We can now answer the questions asked at the beginning of this chapter: Does fine-tuning cry out for explanation? Does it give support to the multiverse hypothesis? Beginning with the latter question, we should say: Yes, to the extent that multiverse theories are simpler, more elegant (and therefore claming a higher prior probability) than any rival theories that are compatible with what we observe. In order to be more precise about the magnitude of support, we need to determine the conditional probability that a multiverse theory gives to the observations we make. We have said something about how such conditional probabilities are determined: the conditional probability is greater—ceteris paribus—the greater the probability that the multiverse theory gives to the existence of a universe exactly like ours; it is smaller—ceteris paribus—the greater the number of observer-containing universes it entails. These two factors balance each other to the effect that if we are comparing various multiverse theories, what matters, generally speaking, is the likelihood they assign to at least some observer-containing universe existing. If two multiverse theories both do that, then there is no general reason to favor or disfavor the one that entails the larger number of observer-containing universes. All this will become clearer in subsequent chapters where the current hand-waving will be replaced by mathematically precise models.

The answer to the question whether fine-tuning cries out for explanation follows from this. If something's "crying out for explanation" means that it would be unsatisfactory to leave it unexplained or to dismiss it as a chance event, then fine-tuning cries out for explanation at least to the extent that we have reason to believe in some theory that would explain it. At present, multiverse theories may look like reasonably promising candidates. For the theologically inclined, the Creator-hypothesis is also a candidate. And there remains the possibility that fine-tuning could turn out to be an illusion—if some neat single-universe theory that fits the data were to be discovered in the future.[7]

Finally, we may also ask whether there is anything surprising about our observation of fine-tuning. Let's assume, as the question presupposes, that the universe really is fine-tuned, in the sense that there is no neat single-

[7] If there is a sense of "explanation" in which a multiverse theory would not explain why we observe a fine-tuned universe, then the prospect of a multiverse theory would not add to the need for explanation in that sense.

universe theory that fits the data (but not in a sense that excludes our universe being one in an ensemble that is itself not fine-tuned). Is such fine-tuning surprising on the chance-hypothesis? It is, per assumption, a low-probability event if the chance-hypothesis is true; and it would tend to disconfirm the chance-hypothesis if there is some other hypothesis with reasonably high prior probability that assigns a high conditional probability to fine-tuning. For it to be a surprising event then (invoking Horwich's analysis) there has to be some alternative to the chance-hypothesis that meets conditions (iii) and (iv). Some would hold that the design hypothesis satisfies these criteria. But if we bracket the design hypothesis, does the multiverse hypothesis fit the bill? We can suppose, for the sake of the argument at least, that the prior probability of the multiverse hypothesis is not too low, so that (iv) is satisfied. The sticky point is condition (iii), which requires that $P(E|h_{\mathrm{M}}) \approx 1$. According to the discussion above, the conditional probability of us observing a fine-tuned universe is greater given a suitable multiverse than given the existence of a single random universe.

[8] The meaning of "representative" is *not* equivalent here to "most numerous type of universe in the multiverse" but rather "the type of universe with the greatest expected fraction of all observers".

[9] One can easily imagine multiverse theories on which this would not necessarily be the case. A multiverse theory could for example include a physics that allowed for two distinct regions in the space of possible boundary conditions to be life-containing. One of these regions could be very broad so that most universes in that region would not be fine-tuned—they would still have contained life even if the values of their physical constants had been slightly different. The other region could be very narrow. Universes in this region would be fine-tuned: a slight perturbation of the boundary conditions would knock a universe out of the life-containing region. If the universes in the two life-containing regions in parameter space are equivalent in other respects, this cosmos would be an instance of a multiverse where representative observer-containing universes would not be fine-tuned. If a multiverse theory assigns a high probability to the multiverse being of this kind, then on the hypothesis that that theory is true, representative observer-containing universes would not be fine-tuned.

[10] It may intuitively seem as if our observing a fine-tuned universe would be even *more* surprising if the only multiverse theory on the table implied that representative observer-containing universes were *not* fine-tuned, because it would then be even more improbable that we should live in a fine-tune universe. This intuition most likely derives from our not accepting the assumptions we made. For instance, the design hypothesis (which we ruled out by fiat) might be able to fit the four criteria and thus account for why we would find the fine-tuning surprising even in this case. Alternatively, we might think it implausible that we would be sufficiently convinced that the only available multiverse hypotheses would be ones in which representative universes would not be fine-tuned. So this represents a rather artificial case where our intuitions could easily go astray. I discuss it only in order to round out the argument and to more fully illustrate how the reasoning works. The point is not important in itself.

[11] It's not clear whether there is an alternative that would work here. There would be if, for instance, one assigned a high prior probability to a design hypothesis on which the designer was highly likely to create only one universe and to make it fine-tuned.

If the multiverse hypothesis is of a suitable kind—such that it entails (or makes it highly likely) that at least one observer-containing universe exists—then the conditional probability, given that hypothesis, of us observing an observer-containing universe should be set equal (or very close) to one. It then comes down to whether on this hypothesis representative[8] observer-containing universes would be fine-tuned.[9] If they would, then it follows that this multiverse hypothesis should be taken to give a very high likelihood to our observing a fine-tuned universe; so Horwich's condition (iii) would be satisfied, and our observing fine-tuning would count as a surprising event. If, on the other hand, representative observer-containing universes in the multiverse would not be fine-tuned, then condition (iii) would not be satisfied, and the fine-tuning would not qualify as surprising.[10]

Note that in answering the question whether fine-tuning was surprising, we focused on E' (the statement that there is a fine-tuned universe) rather than E (the statement that α is fine-tuned). I suggest that what is primarily surprising is E', and E is surprising only in the indirect sense of implying E'. If E is independently surprising, then on Horwich's analysis, it has to be so owing to some other alternative[11] to the chance-hypothesis than the multiverse hypothesis, since it is not the case that $P(E \mid h_M) \approx 1$. But I find it quite intuitive that what would be surprising on the chance-hypothesis is not that *this* universe (understood rigidly) should be fine-tuned but rather that there should be a fine-tuned universe at all if there is only one universe in total and fine-tuning was highly improbable.

PRELIMINARY CONCLUSIONS

It may be useful to summarize our main findings of this chapter. We set out to investigate whether fine-tuning needs explaining and whether it gives support to the multiverse hypothesis. We found:

- There is an easy part of the answer: Leaving fine-tuning unexplained is epistemically unsatisfactory to the extent that it involves accepting complicated, inelegant theories with many free parameters. If a neater theory can account for available data it is to be preferred. This is just an instance of the general methodological principle that one should prefer simpler theories, and it has nothing to do with fine-tuning as such (i.e. this point is unrelated to the fact that observers would not have existed if boundary conditions had been slightly different).

- Ian Hacking's argument that multiverse theories such as Wheeler's oscillating universe model cannot receive any support from fine-tuning data, while multiverse theories such as the one Hacking ascribes to Brandon Carter can receive such support, is flawed. So

are the more recent arguments by Roger White and Phil Dowe purporting to show that multiverse theories *tout court* would not be supported by fine-tuning.

• Those who think fine-tuning gives some support to the multiverse hypothesis have typically tried to argue for this by appealing to the surprisingness of fine-tuning. We examined van Inwagen's straw lottery example, refuted some objections by Carlson and Olsson, and suggested a variant of van Inwagen's example that is more closely analogous to our epistemic situation regarding fine-tuning. In this variant the verdict seems to favor the multiverse advocates, although there appears to be room for opposing intuitions. In order to give the idea that an appeal to the surprisingness of fine-tuning could settle the issue a full run for its money, we considered Paul Horwich's analysis of what makes the truth of a statement surprising. This analysis may provide the best available explication of what multiverse advocates mean when they talk about surprise. It was found, however, that applying Horwich's analysis to the fine-tuning situation didn't settle the issue of whether fine-tuning is surprising. We concluded that in order to determine whether fine-tuning cries out for explanation or gives support for the multiverse hypothesis, it is not enough to appeal to the surprisingness or amazingness of fine-tuning. One has to dig deeper.

• What is needed is a way of determining the conditional probability $P(E \mid h_M)$. I suggested that in order to get this right, it is essential to take into account observation selection effects. We created an informal model of how to think about such effects in the context of fine-tuning. Some of the consequences of this model are as follows:

• Suppose there exists a universe-generating mechanism such that each universe it produces has an equal probability of being observer-containing. Then fine-tuning favors (other things equal) theories on which the mechanism has operated enough times to make it probable that at least one observer-containing universe would result.

• However, if two competing general theories with equal prior probability each implies that the mechanism operated sufficiently many times to (nearly) guarantee that at least one observer-containing universe would be produced, then our observing an observer-containing universe is (nearly) no ground for favoring the theory which entails the greater number of observer-containing universes. Nor does it matter how many observerless universes the theories say exist.

• If two competing general theories with equal prior probability, T_1 and T_2, each entails the same number of observer-containing universes (and we assume that each observer-containing universe contains the same number of observers), but T_1 makes it more likely than does T_2 that a large fraction of all the observers live in universes that have those properties that we have observed that our universe has (e.g. the same values of physical constants), then our observations favor T_1 over T_2.

• Although $P(E|h_M)$ may be much closer to zero than to one, this conditional probability could nonetheless easily be large enough (taking observation selection effects into account) for E to favor the multiverse hypothesis.

• Here is the answer to the "tricky part" of the question about whether fine-tuning needs explanation or supports the multiverse hypothesis: Yes, there is something about fine-tuning as such that adds to the need for explanation and to the support for the multiverse hypothesis over and above what is accounted for by the general principle that simplicity is epistemically attractive. The ground for this is twofold: first, the availability of a potential rival explanation for why the universe is observer-containing. The design hypothesis, presumably, can more plausibly be invoked to explain a world that contains observers than one that doesn't. Second (theology apart), the capacity of the multiverse hypothesis to give a high conditional probability to E (and thereby in some sense to explain E), and to gain support from E, depends essentially on observation selection effects. Fine-tuning is therefore *not* just like any other way in which a theory may require a delicate setting of various free parameters to fit the data. The presumption that observers would not be so likely to exist if the universe were not fine-tuned is crucial. For that presumption entails that if a multiverse theory implies that there is an ensemble of universes, only a few of which are fine-tuned, then what the theory predicts that we should observe is still one of those exceptional universes that are fine-tuned. The observation selection effect enables the theory to give our observing a fine-tuned universe a high conditional probability even though such a universe may be very atypical of cosmos as a whole. If there were no observation selection effect restricting our observation to an atypical proper part of the cosmos, then postulating a bigger cosmos would not in general give a higher greater conditional probability of us observing some particular feature. (It may make it more probable that that feature should be instantiated somewhere or other, but it would also make it less probable that we should happen to be at any particular place where it was instanti-

ated.) Fine-tuning, therefore, involves issues additional to the ones common to all forms of scientific inference and explanation.

• On Horwich's analysis of what makes the truth of a statement surprising, it would be surprising against the background of the chance-hypothesis that only one universe existed and it happened to be fine-tuned. By contrast, that *this* universe should be fine-tuned would not contain any additional surprise factor (unless the design hypothesis could furnish an explanation for this datum satisfying Horwich's condition (iii) and (iv)).

CHAPTER 3

Anthropic Principles

The Motley Family

We have seen how observation selection effects are relevant in assessing the implications of cosmological fine-tuning, and we have outlined a model for how they modulate the conditional probability of us making certain observations given certain hypotheses about the large-scale structure of the cosmos. The general idea that observation selection effects need to be taken into account in cosmological theorizing has been recognized by several authors, and there have been many attempts to express this idea in the form of an "anthropic principle". None of these attempts quite hits the mark, however. Some seem not even to know what they are aiming at.

The first section of this chapter reviews some of the more helpful formulations of the anthropic principle found in the literature and considers how far these can take us. Section two briefly discusses a set of very different "anthropic principles" and explains why they are misguided or at least irrelevant for present purposes. A thicket of confusion surrounds the anthropic principle and its epistemological status. We shall need to clear that up. Since a main thrust of this book is that anthropic reasoning merits serious attention, I shall want to explicitly disown some associated ideas that I don't accept. The third section continues where the first section left off. It argues that formulations found in the literature are inadequate. A forth section proposes a new methodological principle to replace them. This principle will form the core of the theory of observation selection effects that we will develop in the subsequent chapters.

THE ANTHROPIC PRINCIPLE AS EXPRESSING AN OBSERVATION SELECTION EFFECT

The term "anthropic principle" was coined by Brandon Carter in a paper of 1974, where he defined it thus:

> . . . what we can expect to observe must be restricted by the conditions necessary for our presence as observers. (Carter 1974), p. 126

Carter's notion of the anthropic principle, as evidenced by the uses to which he put it, is appropriate and productive, yet his definitions and explanations of it are rather vague. While Carter himself was never in doubt about how to understand and apply the principle, he did not explain it a philosophically transparent enough manner to enable all his readers to do the same.

The trouble starts with the name. Anthropic reasoning has nothing in particular to do with Homo sapiens. Calling the principle "anthropic" is therefore misleading and has indeed misled some authors (e.g. (Gale 1981; Gould 1985; Worrall 1996). Carter regrets not using a different name (Carter 1983). He suggests that maybe "the psychocentric principle", "the cognizability principle" or "the observer self-selection principle" would have been better. The time for terminological reform has probably passed, but emphasizing that the anthropic principle concerns intelligent observers in general and not specifically human observers should help to prevent misunderstandings.

Carter introduced two versions of the anthropic principle, a strong version (SAP) and a weak (WAP). WAP states that:

> . . . we must be prepared to take account of the fact that our location in the universe is *necessarily* privileged to the extent of being compatible with our existence as observers. (p. 127)

And SAP that:

> . . . the Universe (and hence the fundamental parameters on which it depends) must be such as to admit the creation of observers within it at some stage. (p. 129)

Carter's formulations have been attacked alternatively for being mere tautologies (and therefore incapable of doing any interesting explanatory work whatever) and for being widely speculative (and lacking any empirical support). Often WAP is accused of the former and SAP of the latter. I think we have to admit that both these readings are possible, since the definitions of WAP and SAP are very vague. WAP says that we have to "be prepared to take into account" the fact that our location is privileged, but it does not say *how* we are to take account of that fact. SAP says that the universe "must" admit the creation of observers, but we get very different meanings depending how we interpret the "must". Does it serve merely to underscore an implication of available data ("the universe must be life-

admitting—present evidence about our existence implies that!")? Or is the "must" instead to be understood in some stronger sense, for example as alleging some kind of prior metaphysical or theological necessity? On the former alternative, the principle is indisputably true; but then the difficulty is to explain how this trivial statement can be useful or important. On the second alternative, we can see how it could be contentful (provided we can make sense of the intended notion of necessity), the difficulty now being to provide some reason for why we should believe it.

John Leslie (Leslie 1989) argues that AP, WAP and SAP can all be understood as tautologies and that the difference between them is often purely verbal. In Leslie's explication, AP simply says that:

> *Any intelligent living beings that there are can find themselves only where intelligent life is possible.* (Leslie 1989), p. 128

WAP then says that, within a universe, observers find themselves only at spatiotemporal locations where observers are possible. SAP states that observers find themselves only in universes that allow observers to exist. "Universes" means roughly: huge spacetime regions that might be more or less causally disconnected from other spacetime regions. Since the definition of a universe is not sharp, neither is the distinction between WAP and SAP. WAP talks about where within a life-permitting universe we should expect to find ourselves, while SAP talks about in what kind of universe in an ensemble of universes we should expect to find ourselves. On this interpretation the two principles are fundamentally similar, differing in scope only.

For completeness, we may also mention Leslie's (Leslie 1989) "Superweak Anthropic Principle", which states that:

> *If intelligent life's emergence, NO MATTER HOW HOSPITABLE THE ENVIRONMENT, always involves very improbable happenings, then any intelligent living beings that there are evolved where such improbable happenings happened."* (Leslie 1989), p. 132; emphasis and capitals as in the original.

The implication, as Michael Hart (Hart 1982) has stressed, is that we shouldn't assume that the evolution of life on an earth-like planet might not well be extremely improbable. Provided there are enough Earth-like planets, as there almost certainly are in an infinite universe, then even a

[1] The figure 1 in $10^{3,000}$ is Hart's most optimistic estimate of how likely it is that the right molecules would just happen to bump into each other to form a short DNA string capable of self-replication. As Hart himself recognizes, it is possible that there exists some as yet unknown abiotic process bridging the gap between amino acids (which we know can form spontaneously in suitable environments) and DNA-based self-replicating organisms. Such a bridge

chance lower than 1 in $10^{3,000}$ would be enough to ensure (i.e. give an arbitrarily great probability to the proposition) that life would evolve somewhere[1]. Naturally, what we would observe would be one of the rare planets where such an improbable chance-event had occurred. The Superweak AP can be seen as special case of WAP. It doesn't add anything to what is already contained in Carter's principles.

The question that immediately arises is: Has not Leslie trivialized anthropic reasoning with this definition of AP? Not necessarily. Whereas the principles he defines are tautologies, the invocation of them to do explanatory work is dependent on nontrivial assumptions about the world. Rather than the truth of AP being problematic, its *applicability* is problematic. That is, it is problematic whether the world is such that AP can play a role in interesting explanations and predictions. For example, the anthropic explanation of fine-tuning requires the existence of an ensemble of universes differing in a wide range of parameters and boundary conditions. Without the assumption that such an ensemble actually exists, the explanation doesn't get off the ground. SAP, as Leslie defines it, would be true even if there were no other universe than our own, but it would then be unable to help explain the fine-tuning. Writes Leslie:

> It is often complained that the anthropic principle is a tautology, so can explain nothing. The answer to this is that while tautologies cannot by themselves explain anything, they can *enter into* explanations. The tautology that three fours make twelve can help explaining why it is risky to visit the wood when three sets of four lions entered it and only eleven exited. (Leslie 1996), pp. 170–1

I would add that there is a lot more to anthropic reasoning than the anthropic principle. We discussed some of the non-trivial issues in anthropic reasoning in chapter 2, and in later chapters we shall encounter even greater conundrums. Anyhow, I shall argue shortly that the above anthropic principles are too weak to do the job they are supposed to do. They are best seen as special cases of a more general principle, the Self-Sampling Assumption, which itself seems to have the status of methodological and epistemological prescription rather than that of a tautology pure and simple.

ANTHROPIC HODGEPODGE

dramatically improve the odds of life evolving. Some suggestions have been given for what it could be: self-replicating clay structures, perhaps, or maybe something isomorphic to Stuart Kaufmann's autocatalytic sets. But we are still very much in the dark about how life got started on Earth or what the odds are of it happening on a random Earth-like planet.

There are multitudinous "anthropic principles"—I have counted over thirty different ones in the literature. They can be divided into three categories: those that express a purported observation selection effect; those that state some speculative empirical hypothesis; and those that are too muddled or ambiguous to make any clear sense at all. The principles discussed in the previous section are in the first category. Here we will briefly review some members of the other two categories.

Among the better-known definitions are those of physicists John Barrow and Frank Tipler, whose influential 700-page monograph of 1986 has served to introduce anthropic reasoning to a wide audience. Their formulation of WAP is as follows:

> ($WAP_{B\&T}$) The observed values of all physical and cosmological quantities are not equally probable but they take on values restricted by the requirement that there exist sites where carbon-based life can evolve and by the requirement that the Universe be old enough for it to have already done so. (Barrow and Tipler 1986), p. 16[2]

The reference to "carbon-based life" does not appear in Carter's original definition. Indeed, Carter has explicitly stated that he intended the principle to be applicable "not only by our human civilization, but also by any extraterrestrial (or non-human future-terrestrial) civilization that may exist" (Carter 1989, p. 18). It is infelicitous to introduce a restriction to carbon-based life, and misleading to give the resulting formulation the same name as Carter's.

Restricting the principle to carbon-based life forms is a particularly bad idea for Barrow and Tipler, because it robs the principle of its tautological status, thereby rendering their position inconsistent, since they claim that WAP is a tautology. To see that WAP as defined by Barrow and Tipler is not a tautology, it is suffices to note that it is not a tautology that all observers are carbon-based. It is no contradiction to suppose that there are observers who are implemented with other chemical elements, and thus that there may be observed values of physical and cosmological constants that are not restricted by the requirement that carbon-based life evolves.[3]

---

[2] A similar definition was given by Barrow in 1983:

> [The] observed values of physical variables are not arbitrary but take values $V(x,t)$ restricted by the spatial requirement that $x \in L$, where L is the set of sites able to sustain life; and by the temporal constraint that t is bound by time scales for biological and cosmological evolution of living organisms and life-supporting environments. (Barrow 1983), p. 147

[3] There is also no contradiction involved in supposing that we might discover that *we* are not carbon-based.

Realizing that the anthropic principle must not be restricted to carbon-based creatures is not a mere logical nicety. It is paramount if we want to apply anthropic reasoning to hypotheses about other possible life forms that may exist or come to exist in the cosmos. For example, when we discuss the Doomsday argument in chapter 6, this becomes crucial.

Limiting the principle to carbon-based life also has the side effect of encouraging a common type of misunderstanding of what anthropic reasoning is all about. It makes it look as if it were part of a project to restitute Homo sapiens into the glorious role of Pivot of Creation. For example, Stephen Jay Gould's criticism (Gould 1985) of the anthropic principle is based on this misconception. Isn't it ironic that anthropic reasoning should have been attacked from this angle! Anthropic reasoning could rather be said to be *anti*-theological and *anti*-teleological, since it holds up the prospect of an alternative explanation for the appearance of fine-tuning—the puzzlement that forms the basis for the modern version of the teleological argument for the existence of a creator.

Barrow and Tipler also provide a new formulation of SAP:

> ($SAP_{B\&T}$) The Universe must have those properties which allow life to develop within it at some stage in its history. (Barrow and Tipler 1986), p. 21

On the face of it, this is rather similar to Carter's SAP. The two definitions differ in one obvious but minor respect. Barrow and Tipler's formulation refers to the development of *life*. Leslie's version improves this to *intelligent life*. But Carter's definition speaks of *observers*. "Observers" and "intelligent life" are not the same concept. It seems possible that there could be (and might come to be in the future) intelligent, conscious observers who are not part of what we call life—for example by lacking such properties as being self-replicating or having a metabolism, etc. For reasons that will become clear later, Carter's formulation is superior in this respect. Not *being alive*, but *being an (intelligent) observer* is what matters for the purposes of anthropic reasoning.

Barrow and Tipler have each provided their own personal formulations of SAP. These definitions turn out to be quite different from $SAP_{B\&T}$:

> Tipler: . . . intelligent life must evolve somewhere in any physically realistic universe. (Tipler 1982), p. 37

> Barrow: The Universe must contain life. (Barrow 1983), p. 149

These definitions state that life must exist, which implies that life exists. The other formulations of SAP we looked at, by Carter, Barrow & Tipler,

and Leslie, all stated that the universe must *allow* or *admit* the creation of life (or observers). This is most naturally read as saying only that the laws and parameters of the universe must be *compatible* with life—which does not imply that life exists. The propositions are not equivalent.

We are also faced with the problem of how to understand the "must". What is its modal force? Is it logical, metaphysical, epistemological or nomological? Or even theological or ethical? The definitions remain highly ambiguous until this is specified.

Barrow and Tipler list three possible interpretations of $SAP_{B\&T}$ in their monograph:

> (A) There exists one possible Universe 'designed' with the goal of generating and sustaining 'observers'.
> (B) Observers are necessary to bring the Universe into being.
> (C) An ensemble of other different universes is necessary for the existence of our Universe.

---

[4] (A) points to the teleological idea that the universe was designed with the goal of generating observers (spiced up with the added requirement that the "designed" universe be the only possible one). Yet, anthropic reasoning is counter-teleological in the sense described above; taking it into account *diminishes* the probability that a teleological explanation of the nature of the universe is correct. And it is hard to know what to make of the requirement that the universe be the only possible one. This is definitely not part of anything that follows from Carter's original exposition.

(B) is identical to what John Wheeler had earlier branded the *Participatory Anthropic Principle* (PAP) (Wheeler 1975; Wheeler 1977). It echoes Berkelian idealism, but Barrow and Tipler want to invest it with physical significance by considering it in the context of quantum mechanics. Operating within the framework of quantum cosmology and the many-worlds interpretation of quantum physics, they state that, at least in its version (B), SAP imposes a boundary condition on the universal wave function. For example, all branches of the universal wave function have zero amplitude if they represent closed universes that suffer a big crunch before life has had a chance to evolve, from which they conclude that such short-lived universes do not exist. "SAP requires a universe branch which does not contain intelligent life to be non-existent; that is, branches without intelligent life cannot appear in the Universal wave function." ((Barrow and Tipler 1986), p. 503). As far as I can see, this speculation is totally unrelated to anything Carter had in mind when he introduced the anthropic principle, and PAP is irrelevant to the issues we discuss in this book. (For a critical discussion of PAP, see e.g. (Earman 1987).

Barrow and Tipler think that statement (C) receives support from the many-worlds interpretation and the sum-over-histories approach to quantum gravity "because they must unavoidably recognize the existence of a whole class of *real* 'other worlds' from which ours is selected by an optimizing principle." ((Barrow and Tipler 1986), p. 22). (Notice, by the way, that what Barrow and Tipler say about (B) and (C) indicates that the necessity to which these formulations refer should be understood as nomological: physical necessity.) Again, this seems to have little do to with observation selection effects. It is true that there is a connection between SAP and the existence of multiple worlds. From the standpoint of Leslie's explication, this connection can be stated as follows: SAP is applicable (non-vacuously) only if there is a suitable world ensemble; only then can SAP be involved in doing explanatory work. But in no way does anthropic reasoning presuppose that our universe could not have existed in the absence of whatever other universes there might be.

Since none of these is directly related to idea of about observation selection effects, I shall not discuss them further (except for some brief remarks relegated to this footnote[4]).

A "Final Anthropic Principle" (FAP) has been defined by Tipler (Tipler 1982) Barrow (Barrow 1983) and Barrow & Tipler (Barrow and Tipler 1986) as follows:

> Intelligent information-processing must come into existence in the universe, and, once it comes into existence, it will never die out.

Martin Gardner charges that FAP is more accurately named CRAP, the *Completely Ridiculous Anthropic Principle* (Gardner 1986). The spirit of FAP is antithetic to Carter's anthropic principle (Leslie 1985; Carter 1989). FAP has no claim on any special methodological status; it is pure speculation. The appearance to the contrary, created by affording it the honorary title of a "Principle", is what prompts Gardner's mockery.

It may be possible to interpret FAP simply as a scientific hypothesis, and that is indeed what Barrow and Tipler set out to do. In a later book (Tipler 1994), Tipler considers the implications of FAP in more detail. He proposes what he calls the "Omega Point Theory". This theory assumes that our universe is closed, so that at some point in the future it will recollapse in a big crunch. Tipler tries to show that it is physically possible to perform an infinite number of computations during this big crunch by using the shear energy of the collapsing universe, and that the speed of a computer in the final moments can be made to diverge to infinity. Thus there could be an infinity of subjective time for beings that were running as simulations on such a computer. This idea can be empirically tested, and if present data suggesting that our universe is open or flat are confirmed, then the Omega Point Theory will indeed have been falsified (as Tipler himself acknowledges).[5] The point to emphasize here is that FAP is not in any way an application or a consequence of anthropic reasoning (although, of course, anthropic reasoning may have a bearing on how hypotheses such as FAP should be evaluated).

If one does want to treat FAP as an empirical hypothesis, it helps if one charitably deletes the first part of the definition, the part that says that intelligent information processing *must* come into existence. If one does this, one gets what Milan C. Ćirković? and I have dubbed the *Final Anthropic Hypothesis* (FAH). It simply says that intelligent information processing will never cease, making no pretenses to being anything other than an inter-

[5] For further critique of Tipler's theory, see (Sklar 1989).

[6] A non-zero cosmological constant has been considered desirable from several points of view in recent years, because it would be capable of solving the cosmological age problem and because it would arise naturally from quantum field processes (see e.g. (Klapdor and Grotz 1986; Singh 1995; Martel, Shapiro et al. 1998). A universe with a cosmological density parameter $\Omega \approx 1$ and a cosmological constant of about the suggested magnitude $\Lambda \approx 0.7$ would allow the formation of galaxies (Weinberg 1987; Efstathiou 1995) and would last long enough for life to have a chance to develop.

esting empirical question that one may ask. We find (Ćirković and Bostrom 2000) that current balance of evidence seems to tip towards a negative answer. For instance, that recent evidence for a large cosmological constant[6] (Perlmutter, Aldering et al. 1998; Reiss 1998) only makes things worse for FAH. There are, however, some other possible ways in which FAH may be true which cannot be ruled out at the present time, involving poorly understood mechanisms in quantum cosmology.

FREAK OBSERVERS AND WHY EARLIER FORMULATIONS ARE INADEQUATE

The relevant anthropic principles for our purposes are those that describe observation selection effects. The formulations mentioned in the first section of this chapter are all in that category, yet they are insufficient. They cover only a small fraction of the cases that we would want to have covered. Crucially, in all likelihood they don't even cover the actual case: they cannot be used to make interesting inferences about the world we are living in. This section explains why that is so, and why it constitutes serious gap in earlier accounts of anthropic methodology and a fortiori in scientific reasoning generally.

Space is very, *very* big. On the currently most favored cosmological theories we are living in an infinite world, a world that contains an infinite number of planets, stars, galaxies and black holes. This is an implication of most "multiverse theories", according to which our universe is just one in a vast ensemble of physically real universes. But even if our universe is the only one there is, we would still have reason to think that we are prob-

[7] A widespread misconception is that the open universe in the standard Big Bang model becomes spatially infinite only in the temporal limit. The *observable* universe is finite, but only a small part of the whole is observable (by us). One fallacious intuition that might be responsible for this misconception is that the universe came into existence at some spatial point in the Big Bang. A better way of picturing things is to imagine space as an infinite rubber sheet, and gravitationally bound groupings (such as stars and galaxies) as buttons glued on. As we move forward in time, the sheet is stretched in all directions so that the separation between the buttons increases. Going backwards in time, we imagine the buttons coming closer together until, at "time zero", the density of the (still spatially infinite) universe becomes infinite everywhere. See e.g. (Martin 1995).

Until recently, it appeared that the mass density of the universe fell far short of the critical density and thus that the universe is open. Recent evidence, however, suggests that the missing mass might have be the form of vacuum energy (a cosmological constant). This is supported by studies of supernovae and the microwave background radiation. If this is confirmed, it would bring the actual density very close to the critical density, and it may thus be hard to tell whether the universe is open, flat, or closed.

Some additional backing for the infinite-universe hypothesis can be garnered if we consider models of eternal inflation, in which an infinite number of galaxies are produced over time.

[8] I.e. that space is singly connected. There is a recent spate of interest in the possibility that our universe might be multiply connected, in which case it could be both finite and hyperbolic. A multiply connected space could lead to a telltale pattern consisting of a superposition of multiple images of the night sky seen at varying distances from Earth (roughly, one image for each lap around the universe which the light has traveled). Such a pattern has not been found, although the search continues. For an introduction to multiply connected topolo-

ably living in an infinite world. In standard Big Bang cosmology, the universe is (at every point in time) spatially infinite[7] and hence presumably contains infinitely many planets etc., provided we assume the simplest topology[8].

Most modern philosophical investigation relating to the vastness of the cosmos have focused on the fine-tuning of our universe. As we saw in chapter 2, something of a philosophical cottage industry has sprung up around controversies over issues such as whether fine-tuning is in some sense "improbable", whether it should be regarded as surprising, whether it calls out for explanation and if so whether a multiverse theory could explain it, whether it suggests ways in which current physics is incomplete, or whether it is evidence for the hypothesis that our universe resulted from design.

Here we shall turn our attention to a more fundamental problem: How can vast-world cosmologies have *any* observational consequences *at all*? I will show that these cosmologies imply (or give a very high probability to) the proposition that every possible observation is in fact made. This creates a challenge: if a theory is such that for any possible human observation that we specify, the theory says that that observation will be made, then how do we test the theory? I call this a "challenge" because cosmologists are constantly modifying and refining theories in light of empirical findings, and they are surely not irrational in doing so. The challenge is explain how that is possible, i.e. to find the missing methodological link that enables a reliable connection to be established between cosmological theories and astronomic observation.

Consider a random phenomenon, for example Hawking radiation. When black holes evaporate, they do so in a random manner[9] such that for any given physical object there is a finite (although, typically, astronomically small) probability that it will be emitted by any given black hole in a given time interval. Such things as boots, computers, or ecosystems have

---

cosmology, see (Lachièze-Rey and Luminet 1995). There is an obvious methodological catch in trying to gain high confidence about the global topology of spacetime—if it is so big that we observe but a tiny, tiny speck of it, then how can we be sure that the whole resembles this particular part that we are in? A large sphere, for example, appears flat if you look at a small patch of it.

[9] Admittedly, a complete understanding of black holes probably requires new physics. For example, the so-called information loss paradox is a challenge for the view that black hole evaporation is totally random (see e.g. (Belot, Earman et al. 1999) for an overview). But even pseudo-randomness, like that of the trajectories of molecules in gases in a deterministic universe, would be sufficient for the present argument.

[10] See e.g. (Hawking and Israel 1979): "[I]t is possible for a black hole to emit a television set or Charles Darwin" (p. 19). (To avoid making a controversial claim about personal identity, Hawking and Israel ought perhaps to have weakened this to ". . . an exact replica of Charles Darwin".) See also (Garriga and Vilenkin 2001).

some finite probability of popping out from a black hole. The same holds true, of course, for human bodies, or human brains in particular states.[10] Assuming that mental states supervene on brain states, there is thus a finite probability that a black hole will produce a brain in a state of making any given observation. Some of the observations made by such brains will be illusory, and some will be veridical. For example, some brains produced by black holes will have the illusory of experience of reading a measurement device that does not exist. Other brains, with the same experiences, will be making veridical observations —a measurement device may materialize together with the brain and may have caused the brain to make the observation. But the point that matters here is that any observation we could make has a finite probability of being produced by any given black hole.

The probability of *anything* macroscopic and organized appearing from a black hole is, of course, minuscule. The probability of a given conscious brain-state being created is even tinier. Yet even a low-probability outcome has a high probability of occurring if the random process is repeated often enough. And that is precisely what happens in our world, if the cosmos is very vast. In the limiting case where the cosmos contains an infinite number of black holes, the probability of any given observation being made is one.[11]

There are good grounds for believing that our universe is infinite and contains an infinite number of black holes. Therefore, we have reason to think that any possible human observation is in fact instantiated in the actual world.[12] Evidence for the existence of a multiverse would only add further support to this proposition.

It is not necessary to invoke black holes to make this point. Any random physical phenomenon would do. It seems we don't even have to limit the argument to quantum fluctuations. Classical thermal fluctuations could, presumably, in principle lead to the molecules in a cloud of gas, which contains the right elements, to spontaneously bump into each other so as to form a biological structure such as a human brain.

The problem is that it seems impossible to get any empirical evidence that could distinguish between different Big World theories. For any obser-

---

[11] In fact, there is a probability of unity that infinitely many such observers will appear. But one observer will suffice for our purposes.

[12] I restrict the assertion to *human* observations in order to avoid questions as to whether there may be other kinds of possible observations that perhaps could have infinite complexity or be of some alien or divine nature that does not supervene on stuff that is emitted from black holes—such stuff is physical and of finite size and energy.

[13] Some cosmologists are recently becoming aware of the problematic that this section describes (e.g. (Linde and Mezhlumian 1996; Vilenkin 1998). See also (Leslie 1992).

vation we make, *all* such theories assign a probability of one to the hypothesis that that observation be made. That means that the fact that the observation is made gives us no reason whatever for preferring one of these theories to the others. Experimental results appear totally irrelevant.[13]

We can see this formally as follows. Let B be the proposition that we are in a Big World, defined as one that is big enough and random enough to make it highly probable that every possible human observation is made. Let T be some theory that is compatible with B, and let E be some proposition asserting that some specific observation is made. Let P be an epistemic probability function. Bayes' theorem states that

$$P(T|E\&B) = P(E|T\&B)P(T|B) / P(E|B).$$

In order to determine whether E makes a difference to the probability of T (relative to the background assumption B), we need to compute the difference $P(T|E\&B) - P(T|B)$. By some simple algebra, it is easy to see that

$$P(T|E\&B) - P(T|B) \approx 0 \text{ if and only if } P(E|T\&B) \approx P(E|B).$$

This means that E will fail to give empirical support to E (modulo B) if E is about equally probable given $T\&B$ as it is given B. We saw above that $P(E|T\&B) \approx P(E|B) \approx 1$. Consequently, whether E is true or false is irrelevant for whether we should believe in T, given that we know that B.

Let T_2 be some perverse permutation of an astrophysical theory T_1 that we actually accept. T_2 differs from the T_1 by assigning a different value to some physical constant. To be specific, let us suppose that T_1 says that the current temperature of the cosmic microwave background radiation is about 2.7 degrees Kelvin (which is the observed value) whereas T_2 says it is, say, 3.1 K. Suppose furthermore that both T_1 and T_2 say that we are living in a Big World. One would have thought that our experimental evidence favors T_1 over T_2. Yet, the above argument seems to show that this view is mistaken. Our observational evidence supports T_2 just as much as T_1. We really have no reason to think that the background radiation is 2.7 K rather than 3.1 K.

At first blush, it could seem as if this simply rehashes the lesson, familiar from Duhem and Quine, that it is always possible to rescue a theory from falsification by modifying some auxiliary assumption, so that strictly speaking no scientific theory ever implies any observational consequences. The above argument would then merely have provided an illustration of how this general result applies to cosmological theories. But that would totally miss the point.

If the argument given above is correct, it establishes a much more radical conclusion. It purports to show that all Big World theories are not only

logically compatible with any observational evidence, but they are also *perfectly probabilistically compatible*. They all give the same conditional probability (namely one) to every observation statement E defined as above. This entails that no such observation statement can have *any* bearing, whether logical or probabilistic, on whether the theory is true. If that were the case, it would not be worthwhile to make astronomical observations if what we are interested in is determining which Big World theory to accept. The only reasons we could have for choosing between such theories would be either a priori ones (simplicity, elegance etc.) or pragmatic ones (such as ease of calculation).

Nor is the argument making the ancient statement that human epistemic faculties are fallible, that we can never be certain that we are not dreaming or that we are not brains in a vat. No, the point here is not that such illusions *could* occur, but rather that we have reason to believe that they *do* occur, not just some of them but all possible ones. In other words, we can be fairly confident that the observations we make, along with all possible observations we could make in the future, are being made by brains in vats and by humans that have spontaneously materialized from black holes or from thermal fluctuations. The argument would entail that this abundance of observations makes it impossible to derive distinguishing observational consequences from contemporary cosmological theories.

Most readers will find this conclusion unacceptable. Or so, at least, I hope. Cosmologists certainly appear to be doing experimental work and to modify their theories in light of new empirical findings. The COBE satellite, the Hubble Space Telescope, and other devices are these days showering us with a wealth of new and exciting data, causing a minor renaissance in the world of astrophysics. Yet the argument described above would show that the empirical import of this information could never go beyond the limited role of providing support for the hypothesis that we are living in a Big World, for instance by showing that the universe is open. Nothing apart from this one fact could be learnt from such observations. Once we have established that the universe is open and infinite, then any further work in observational astronomy would be a waste of time and money.

Worse still, the leaky connection between theory and observation in cosmology spills over into other domains. Since nothing hinges on how we defined T in the derivation above, the argument can easily be extended to prove that observation does not have a bearing on *any* empirical scientific question so long as we assume that we are living in a Big World.

This consequence is absurd, so we should look for a way to fix the methodological pipeline and restore the flow of testable observational consequences from Big World theories. How can we do that?

Taking into account the selection effects expressed by SAP, much less those expressed by WAP or the Super-weak AP, will not help us. It isn't true that we couldn't have observed a universe that wasn't fine-tuned for

life. For even "uninhabitable" universes can contain the odd, spontaneously materialized "freak observer", and if they are big enough or if there are sufficiently many such universes, then it is indeed highly likely that they contain infinitely many freak observers making all possible human observations. It's even logically consistent with all our evidence that *we are* such freak observers.

It may appear as if this is a fairly superficial problem. It is based on the technical point that some infrequent freak observers will appear even in non-tuned universes. Couldn't it be thought that this shouldn't really matter because it is still true that the overwhelming majority of all observers are regular observer, not freak observers? We can't interpret "the majority" in the straightforward cardinal sense, since the class of freak observers may well be of the same cardinality as the class of regular observers; but nonetheless, in some natural sense, "almost all" observers in a multiverse live in the fine-tuned parts and have evolved via ordinary processes. So if we modify SAP slightly, to allow for a small proportion of observers living in non-tuned universes, maybe we could repair the methodological pipeline and make the anthropic fine-tuning explanation (among other useful results) go through?

I think that this is precisely the right approach! The presence of the odd observer in a non-tuned universe changes nothing essential. SAP should be modified or strengthened to make this clear. Let's set aside the aside for the moment the complication of infinite numbers of observers and assume that the total number is finite. Then the idea is that so long as the vast majority of observers are in fine-tuned universes, and the ones in non-tuned universes are a small minority, then what the multiverse theory predicts is that we should *with overwhelming probability* find ourselves in one of the fine-tuned universes. That we observe such a universe is thus what such a multiverse theory predicts, and our observations would therefore tend to confirm it to some degree. A multiverse theory of the right kind, coupled with this ramified version of the anthropic principle, can potentially account for the apparent fine-tuning of our universe and explain how our scientific theories are testable even when conjoined with Big World hypotheses. (In chapter 5 we shall explain how this works in more detail.)

How to formulate the requisite kind of anthropic principle? Astrophysicist Richard Gott III has taken one step in the right direction with his "Copernican anthropic principle":

> [T]he location of your birth in space and time in the Universe is privileged (or special) only to the extent implied by the fact that you are an intelligent observer, that your location among intelligent observers is not special but rather picked at random from the set of all intelligent observers (past, present and future) any one of whom you could have been. (Gott 1993), p. 316

This definition comes closer than any of the others we have examined to giving an adequate expression of the basic idea behind anthropic reasoning. It introduces a notion of randomness that can be applied to the Big World theories that we are examining. Yes, you could have lived in a non-tuned universe, but if the vast majority of observers live in fine-tuned universes then the multiverse theory predicts that you should (very probably) find yourself in a fine-tuned universe.

One drawback with Gott's definition is that it makes some problematic claims which may not be essential to anthropic reasoning. It says your location was "picked at random". But who or what did the picking? Maybe that is too naïve a reading. Yet the expression does suggest that there is some kind of physical randomization mechanism at work, which, so to speak, picks out a birthplace for you. We can imagine a possible world where this would be a good description of what was going on. Suppose God, after having created a multiverse, posts a world-map on the door to His celestial abode. He takes a few steps back and starts throwing darts at the map, creating bodies wherever they hit, and sends down souls to inhabit the bodies. Alternatively, maybe one could imagine some sort of physical apparatus, involving a time travel machine, that could putter about in spacetime and distribute observers in a truly random fashion. But what evidence is there that any such randomization mechanism exists? None, as far as I can see. Perhaps some less farfetched story could be spun that would lead to the same result, but anthropic reasoning would be tenuous indeed had it to rely on such suppositions—which, thankfully, it doesn't.

Also, the assertion that "you could have been" any of these intelligent observers who will ever have existed is problematic. Ultimately, we *may* have to confront this problem but it would be nicer to have a definition that doesn't preempt that debate.

Both these points are relatively minor quibbles. I think one could reasonably explicate Gott's definition so that it comes out right in these regards.[14] There is, however, a much more serious problem with Gott's approach which we shall discuss during the course of our examination of the Doomsday argument in chapter 6. We will therefore work with a different principle which sidesteps these difficulties.

The Self-Sampling Assumption

The preferred explication of the anthropic principle that we shall use as a starting point for subsequent investigations is the following, which we call the *Self-Sampling Assumption*:

---

[14] In his work on inflationary cosmology, Alexander Vilenkin has proposed a "Principle of Mediocrity" (Vilenkin 1995), which is similar to Gott's principle.

> (SSA) One should reason as if one were a random sample from the set of all observers in one's reference class.

This is a *preliminary* formulation. Anthropic reasoning is about taking observation selection effects into account, which creep in when we are trying to evaluate evidence that has an indexical component. In chapter 10 we shall replace SSA with another principle that takes more indexical information into account. That principle will show that only under certain special conditions is SSA a permissible simplification. However, in order to get to the point where we can understand and appreciate the more general principle, it is pedagogically necessary to first thoroughly examine SSA—both the reasons for accepting it, and the consequences that flow from its use. Wittgenstein's famous ladder, which one must first climb and then kick away, is a perfect metaphor for how we should view SSA. Thus, rather than inserting qualifications everywhere, we'll just state here once that we will revisit and reassess SSA when we reach chapter 10.

SSA as stated leaves open what the appropriate reference class might be and what sampling density should be imposed over this reference class. Those are crucial issues that will need very careful studying, an enterprise that we shall embark on in the next chapter.

The other observational selection principles discussed above are special cases of SSA. Take first WAP (in Carter and Leslie's rendition). If a theory *T* says that there is only one universe and some regions of it contain no observers, then WAP says that *T* predicts that we don't observe one of those observerless regions. (That is, that we don't observe them "from the inside". If the region is observable from a region where there are observers, then obviously it could be observable by those observers.) SSA yields the same result, since if there is no observer in a region, then there is zero probability that a sample taken from the set of all observers will be in that region, and hence zero probability that you should observe that region given the truth of *T*.

Similarly, if *T* says there are multiple universes, only some of which contain observers, then SAP (again in Carter and Leslie's sense) says that *T* predicts that what you should observe is one of the universes that contain observers. SSA says the same, since it assigns zero sampling density to being an observer in an observerless universe.

The meaning, significance, and use of SSA will be made clearer as we proceed. We can already state, however, that SSA and its strengthenings and specifications are to be understood as *methodological prescriptions*. They state how reasonable epistemic agents ought to assign credence in certain situations and how to make certain kinds of probabilistic inferences. As will appear from subsequent discussion, SSA is not (in any straightforward way at least) a restricted version of the principle of indif-

CHAPTER 4

Thought Experiments Supporting the Self-Sampling Assumption

This chapter and the next argue that we should accept SSA. In the process, we also elaborate on the principle's intended meaning and we begin to develop a theory of how SSA can be used in concrete scientific contexts to guide us through the thorny issues of anthropic biases.

The case for accepting SSA has two separable parts. One part focuses on its applications. We will continue the argument begun in the last chapter, that a new methodological rule is needed in order to explain how observational consequences can be derived from contemporary cosmological and other scientific theories. I will try to show how SSA can do this for us. This part will be considered in the next chapter, where we'll also look at how SSA underwrites some types of inferences in thermodynamics, evolutionary biology, and traffic analysis.

This chapter will deal with the other part of the case for SSA. It consists of a series of thought experiments designed to demonstrate that it is rational to reason in accordance with SSA in a rather wide range of circumstances. While the application-part can be likened to field observations, the thought experiments we shall conduct in this chapter are more like laboratory research. We here have full control over all relevant variables and can stipulate away inessential complications in order to hopefully get a more accurate measurement of our intuitions and epistemic convictions regarding SSA.

The Dungeon gedanken

Our first gedanken is *Dungeon*:

> The world consists of a dungeon that has one hundred cells. In each cell there is one prisoner. Ninety of the cells are painted blue on the outside and the other ten are painted red. Each prisoner is

> asked to guess whether he is in a blue or a red cell. (And everybody knows all this.) You find yourself in one of these cells. What color should you think it is?—Answer: Blue, with 90% probability.

Since 90% of all observers are in blue cells, and you don't have any other relevant information, it seems you should set your credence of being in a blue cell to 90%. Most people I've talked to agree that this is the correct answer. Since the example does not depend on the exact numbers involved, we have the more general principle that in cases like this, your credence of having property *P* should be equal to the fraction of observers who have *P*, in accordance with SSA.[1] Some of our subsequent investigations in this chapter will consider arguments for extending this class in various ways.

While many accept without further argument that SSA is applicable to the *Dungeon* gedanken, let's consider how one might seek to defend this view if challenged to do so.

One argument we may advance is the following. Suppose everyone accepts SSA and everyone has to bet on whether they are in a blue or a red cell. Then 90% of all prisoners will win their bets; only 10% will lose. Suppose, on the other hand, that SSA is rejected and the prisoners think that one is no more likely to be in a blue cell; so they bet by flipping a coin. Then, on average, 50% of the prisoners will win and 50% will lose. It seems better that SSA be accepted.

This argument is incomplete as it stands. Just because one pattern *A* of betting leads more people to win their bets than another pattern *B*, we shouldn't think that it is rational for anybody to bet in accordance with pattern *A* rather than *B*. In *Dungeon*, consider the betting pattern *A* which specifies that "If you are Harry Smith, bet you are in a red cell; if you are Geraldine Truman, bet that you are in a blue cell; . . ."—such that for each person in the experiment, *A* gives the advice that will lead him or her to be right. Adopting rule *A* will lead to more people winning their bets (100%) than any other rule. In particular, it outperforms SSA which has a mere 90% success rate.

Intuitively it is clear that rules like *A* are cheating. This is maybe best seen if we put *A* in the context of its rival permutations *A′*, *A″*, *A‴* etc., which map the captives' names to recommendations about betting red or

[1] This does not rule out that there could be other principles of assigning probabilities that would also provide plausible guidance in *Dungeon*, provided their advice coincides with that of SSA. For example, a relatively innocuous version of the Principle of Indifference, formulated as "*Assign the same credence to any two hypotheses if you don't have any reason to prefer one to the other*", would also do the trick in *Dungeon*. But subsequent thought experiments impose additional constraints, and for reasons that will become clear, it doesn't seem that any straightforward principle of indifference would suffice to express the needed methodological rule.

blue in other ways than does *A*. Most of these permutations do rather badly. On average they give no better advice than flipping a coin, which we saw was inferior to accepting SSA. Only if the people in the cells could pick the right *A*-permutation would they benefit. In *Dungeon,* they don't have any information enabling them to do this. If they picked *A* and consequently benefited, it would be pure luck.

What allows the people in *Dungeon* to do better than chance is that they have a relevant piece of empirical information regarding the distribution of observers over the two types of cells. They have been informed that 90% of them are in blue cells, and it would be irrational of them not to take this information into account. We can imagine a series of thought experiments where an increasingly large fraction of observers are in blue cells—91%, 92%, . . . , 99%. The situation gradually degenerates into the 100%-case where they are told, "You are all in blue cells", from which each can deductively infer that she is in a blue cell. As the situation approaches this limiting case, it is plausible to require that the strength of participants' beliefs about being in a blue cell should gradually approach probability 1. SSA has this property.

One may notice that while it is true that if the detainees adopt SSA, 90% of them wouldwin their bets; yet there are even simpler methods that produce the same result. For instance: "Set your probability of being in a blue cell equal to 1 if most people are in blue cells; and to 0 otherwise." Using this epistemic rule will also result in 90% of the people winning their bets. Such a rule would not be attractive however. First, when the participants step out of their cells, some of them will find that they were in red cells. Yet if their prior probability of that were zero, they could never learn that by Bayesian belief updating. The second and more generic point is that when we consider rational *betting quotients*, rules like this are revealed to be inferior. A person whose probability for finding herself in a blue cell was 1 would be willing to bet on that hypothesis at any odds.[2] The people following this simplified rule would thus risk losing arbitrarily great sums of money for an arbitrarily small and uncertain gain—an uninviting strategy. Moreover, collectively, they would be *guaranteed* to lose an arbitrarily large sum.

Suppose we agree that all the participants should assign the same probability to being in a blue cell (which is quite plausible since their evidence does not differ in any relevant way). It is then easy to show that out of all possible probabilities they could assign to finding themselves in blue cells, a probability of 90% is the only one which would make it impossible to bet against them in such a way that they were collectively guaranteed to lose money. And in general, if we vary the numbers of the example, their

[2] Setting aside, as is customary in contexts like this, any risk aversion or aversion against gambling, or computational limitations that the person might have.

degree of belief would in each case have to be what SSA prescribes in order to save them from being a *collective sucker*.

On an individual level, if we imagine the experiment repeated many times, the only way a given participant could avoid having a negative expected outcome when betting repeatedly against a shrewd outsider would be by setting her odds in accordance with SSA.

All these considerations support what seems to be most persons' initial intuition about *Dungeon*: that it is a situation where one should reason in accordance with SSA. Any plausible principle of the epistemology of information that has an indexical component would have to agree with SSA's verdicts in this particular case.

One thing that should be noticed about *Dungeon* is that we didn't specify how the prisoners arrived in their cells. The prisoners' ontogenesis is irrelevant so long as they don't know anything about it that gives them clues about the color of their abodes. For example, they may have been allocated to their respective cells by some objectively random process such as drawing tickets from a lottery urn, after which they were blindfolded and led to their designated locations. Or they may have been allowed to choose cells for themselves, and a fortune wheel subsequently spun to determine which cells should be painted blue and which red. But the gedanken doesn't depend on there being a well-defined randomization mechanism. One may just as well imagine that prisoners have been in their cells since the time of their birth or indeed since the beginning of the universe. If there is a possible world where the laws of nature dictate which individuals are to appear in which cells, without any appeal to initial conditions, then the inmates would still be rational to follow SSA, provided only that they did not have knowledge of the laws or were incapable of deducing what the laws implied about their own situation. Objective chance, therefore, is not an essential part of the thought experiment; it runs on low-octane subjective uncertainty.

Two thought experiments by John Leslie

We shall now look at an argument for extending the range of cases where SSA can be applied. We shall see that the synchronous nature of *Dungeon* is an inessential feature: you can in some contexts legitimately reason as if you were a random sample from a reference class that includes observers who exist at different times. Also, we will find that one and the same reference class can contain observers who differ in many respects, including their genes and gender. To this effect, consider an example due to John Leslie, which we shall refer to as *Emeralds*:

> Imagine an experiment planned as follows. At some point in time, three humans would each be given an emerald. Several centuries afterwards, when a completely different set of humans was alive, five thousand

> humans would each be given an emerald. Imagine next that you have yourself been given an emerald in the experiment. You have no knowledge, however, of whether your century is the earlier century in which just three people were to be in this situation, or in the later century in which five thousand were to be in it. . . .
>
> Suppose you in fact betted that you lived [in the earlier century]. If every emerald-getter in the experiment betted in this way, there would be five thousand losers and only three winners. The sensible bet, therefore, is that yours is instead the later century of the two. (Leslie 1996), p. 20

The arguments that were made for SSA in *Dungeon* can be recycled in *Emeralds*. Leslie makes the point about more people being right if everyone bets that they are in the later of the two centuries. As we saw in the previous section, this point needs to be supplemented by additional arguments before it yields support for SSA. (Leslie gives the emeralds example as a response to one objection against the Doomsday argument. He never formulates SSA, but parts of his arguments in defense of the Doomsday argument and parts of his account of anthropic reasoning in cosmology are relevant to evaluating SSA.)

As Leslie notes, we can learn a second lesson if we consider a variant of the emeralds example (*Two Batches*):

> A firm plan was formed to rear humans in two batches: the first batch to be of three humans of one sex, the second of five thousand of the other sex. The plan called for rearing the first batch in one century. Many centuries later, the five thousand humans of the other sex would be reared. Imagine that you learn you're one of the humans in question. You don't know which centuries the plan specified, but you are aware of being female. You very reasonably conclude that the large batch was to be female, almost certainly. If adopted by every human in the experiment, the policy of betting that the large batch was of the same sex as oneself would yield only three failures and five thousand successes. . . . [Y]ou mustn't say: 'My *genes* are female, so I have to observe myself to be female, no matter whether the female batch was to be small or large. Hence I can have no special reason for believing it was to be large.' (Ibid. pp. 222–3)

If we accept this, we can conclude that members of both genders can be in the same reference class. In a similar vein, one can argue for the irrelevance of short or tall, black or white, rich or poor, famous or obscure, fierce or meek, etc. If analogous arguments with two batches of people with any of these property pairs are accepted, then we have quite a broad reference class already. We shall return in a moment to consider what limits there might be to how wide the reference class can be, but first we want

to look at another dimension in which one may seek to extend the applicability of SSA.

THE *INCUBATOR* GEDANKEN

All the examples so far have been of situations where all the competing hypotheses entail the same number of observers in existence. A key new element is introduced in cases where the total number of observers is different depending on which hypothesis is true. Here is a simple case where this happens.

Incubator, version I

Stage (a): In an otherwise empty world, a machine called "the incubator"[3] kicks into action. It starts by tossing a fair coin. If the coin falls tails then it creates one room and a man with a black beard inside it. If the coin falls heads then it creates two rooms, one with a black-bearded man and one with a white-bearded man. As the rooms are completely dark, nobody knows his beard color. Everybody who's been created is informed about all the above. You find yourself in one of the rooms. *Question:* What should be your credence that the coin fell tails?

Stage (b): A little later, the lights are switched on, and you discover that you have a black beard. *Question:* What should your credence in tails be now?

Consider the following three models of how you should reason:

Model 1 (Naïve)

Neither at stage (*a*) nor at stage (*b*) do you have any relevant information as to how the coin (which you know to be fair) landed. Thus in both instances, your credence of tails should be 1/2.

Answer: At stage (a) your credence of tails should be 1/2 and at stage (b) it should be 1/2.

Model 2 (SSA)

If you had had a white beard, you could have inferred that there were two rooms, which entails heads. Knowing that you have a

[3] We suppose the incubator to be a mindless automaton that doesn't count as an observer.

black beard does not allow you to rule out either possibility but it is still relevant information. This can be seen by the following argument. The prior probability of Heads is one half, since the coin was fair. If the coin fell heads, then the only observer in existence has a black beard; hence by SSA, the conditional probability of having a black beard given heads is one. If the coin fell tails, then one out of two observers has a black beard; hence, also by SSA, the conditional probability of a black beard given tails is one half. That is, we have

P(Heads) = P(¬Heads) = ½

P(Black | Heads) = ½

P(Black | ¬Heads) = 1

By $$= \frac{P(\text{Black} \mid \text{Heads})P(\text{Heads})}{P(\text{Black} \mid \text{Heads})P(\text{Heads}) + P(\text{Black} \mid \neg\text{Heads})P(\neg\text{Heads})} = 1/3.$$

Bayes' theorem, the posterior probability of heads, after conditionalizing on Black, is

P(Heads | Black)

Answer: At stage (a) your credence of tails should be ½ and at stage (b) it should be ⅔.

Model 3 (SSA & SIA)

It is twice as likely that you should exist if two observers exist than if only one observer exists. This follows if we make the Self-*Indication* Assumption (SIA), to be explained shortly. The prior probability of heads should therefore be ⅔, and of tails, ⅓. As in Model 2, the conditional probability of a black beard given heads is 1 and the conditional probability of black beard given tails is ½.

P(Heads) = ⅔

P(¬Heads) = ⅓

P(Black | Heads) = ½

P(Black | ¬Heads) = 1

By Bayes' theorem, we get

P (Heads | Black) = ½.

Answer: At stage (a) your credence of tails should be ⅓ and at stage (b) it should be ½.

The last model uses something that we have dubbed the Self-Indication Assumption, according to which you should conclude from the fact that you came into existence that probably quite a few observers did:

> (SIA) Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist.

SIA may seem *prima facie* implausible, and we shall argue in chapter 7 that it is no less implausible *ultimo facie*. Yet some of the more profound criticisms of specific anthropic inferences rely implicitly on SIA. In particular, adopting SIA annihilates the Doomsday argument. It is therefore good to put it on the table so we can consider what reasons there are for accepting or rejecting it. To give SIA the best chance it can get, we will postpone this evaluation until we have discussed the Doomsday argument and have seen why a range of more straightforward objections against the Doomsday argument fail. The fact that SIA could seem to be the only coherent way (but later we'll show that it only seems that way!) of resisting the Doomsday argument is possibly the strongest argument that can be made in its favor.

For the time being, we put SIA to one side (i.e. we assume that it is false) and focus on comparing Model 1 and Model 2. The difference between these models is that Model 2 uses SSA and Model 1 doesn't. By determining which of these models is correct, we get a test of whether SSA should be applied in epistemic situations where hypotheses implying different numbers of observers are entertained. If we find that Model 2 (or, for that matter, Model 3) is correct, we have extended the applicability of SSA beyond what was established in the previous sections, where the number of observers did not vary between the hypotheses under consideration.

In Model 1 we are told to consider the objective chance of 50% of the coin falling heads. Since you know about this chance, you should according to Model 1 set your subjective credence equal to it.

The step from knowing about the objective chance to setting your credence equal to it follows from the *Principal Principle*[4]. This is not the place to delve into the details of the debates surrounding this principle and the connection between chance and credence (see Skyrms 1980; Kyburg, Jr.

[4] David Lewis (Lewis 1986; Lewis 1994). A similar principle had earlier been formulated by Hugh Mellor (Mellor 1971).

1981; Bigelow, Collins et al. 1993; Hall 1994; Halpin 1994; Thau 1994; Strevens 1995; Hoefer 1997; Black 1998; Sturgeon 1998; Vranas 1998; Bostrom 1999; Hoefer 1999). Suffice it to point out that the Principal Principle does not say that you should always set your credence equal to the corresponding objective chance if you know it. Instead, it says that you should do this *unless* you have other relevant information that should be taken into account. There is some controversy about how to specify which types of such additional information will modify reasonable credence when the objective chance is known, and which types of additional information will leave the identity intact. But there is general agreement that the proviso is needed. For example, no matter how objectively chancy a process is, and no matter how well you know the chance, if you have actually seen what the outcome was, your credence in that observed outcome should of course be one (or extremely close to one) and your credence in any other outcome the process could have had should be (very close to) zero. This is so quite independently of what the objective chance was. None of this is controversial.

Now the point is that in *Incubator* you have such extra relevant information that you need to take into account, and Model 1 fails to do that. The extra information is that you have a black beard. This information is relevant because it bears probabilistically on whether the coin fell heads or tails. We can see this as follows. Suppose you are in a room but you don't know what color your beard is. You are just about to look in the mirror. If the information that you have a black beard were not probabilistically relevant to how the coin fell, there would be no need for you to change your credence about the outcome after looking in the mirror. But this is an incoherent position. For there are two things you may find when looking in the mirror: that you have a black beard or that you have a white beard. Before the light comes on and you peek in the mirror, you know that if you find that you have a white beard then you will have conclusively refuted the hypothesis that the coin fell tails. So the mirror *might* give you information that would increase your credence of Heads (to 1). But that entails that making the other possible finding (that you have a black beard) must *decrease* your credence in heads. In other words, your conditional credence of Heads given black beard must be less than your unconditional credence of Heads.

If your conditional probability of heads given a black beard were not lower than the probability you assign to heads, while also your conditional probability of heads given a white beard, is one, then you would be incoherent. This is easily shown by a standard Dutch book argument, or more simply as follows:

$$P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e)}$$

$$P(\neg h \mid e) = \frac{P(e \mid \neg h)P(\neg h)}{P(e)}.$$

$$\frac{P(h \mid e)}{P(\neg h \mid e)} = \frac{P(e \mid h)\Pr(h)}{P(\neg h)} < \frac{P(h)}{P(\neg h)}.$$

Write h for the hypothesis that the coin fell heads, and e for the evidence that you have a black beard. We can assume that $P(e \mid h) < 1$. Then we have

and

.

Dividing these two equations and using , we get

So the quotients between the probabilities of h and $\neg h$ is less *after* e is known than *before*. In other words, learning e decreases the probability of h and increases the probability of $\neg h$.

So the observation that you have a black beard gives you relevant information that you need to take into account and it should lower your credence of Tails to below your unconditional credence of Tails, which (provided we reject SIA) is 50%. Model 1, which fails to do this, is therefore wrong.

Model 2 does take the information about your beard color into account and sets your posterior credence of heads to ⅓, lower than it would have been had you not seen your beard. This is a consequence of SSA. The exact figure depends on the assumption that your conditional probability of a black beard equals that of a white beard, *given* heads. If you knew that the coin landed heads but you hadn't yet looked in the mirror, you would know that there was one man with a white beard and one with black. Provided these men were sufficiently similar in other respects (so that from your present position of ignorance about your beard color you didn't have any evidence as to which one of them you are), these conditional credences should both be 50% according to SSA.

If we agree that Model 2 is the correct one for *Incubator* then we have seen how SSA can be applied to problems where the total number of observers in existence is not known. In chapter 10, we will reexamine *Incubator* and argue for adoption of a fourth model, which conflicts with Model 2 in subtle but important ways. The motivation for doing this, however, will become clear only after detailed investigations into the conse-

quences of accepting Model 2. So for the time being, we will adopt Model 2 as our working assumption in order to explore the implications of the way of thinking it embodies.

If we combine this with the lessons of the previous thought experiments, we now have a very wide class of problems where SSA can be applied. In particular, we can apply it to reference classes that contain observers who live at different times; that are different in many substantial ways including genes and gender; and that may be of different sizes depending on which hypothesis under consideration is true.

One may wonder if there are any limits at all to how much we can include in the reference class. *There are.* We shall now see why.

THE REFERENCE CLASS PROBLEM

The reference class in the SSA is the class of entities such that one should reason as if one were randomly selected from it. We have seen examples of things that must be included in the reference class. In order to complete the specification of the reference class, we also have to determine what things must be excluded.

In many cases, where the total number of observers is the same on any of the hypotheses assigned non-zero probability, the problem of the reference class appears irrelevant. For instance, take *Dungeon* and suppose that in ten of the blue cells there is a polar bear instead of a human observer. Now, whether the polar bears count as observers who are members of the reference class makes no difference. Whether they do or not, you know you are not one of them. Thus you know that you are not in one of the ten cells they occupy. You therefore recalculate the probability of being in a blue cell to be $^{8}/_{9}$, since 80 out of the 90 observers whom you—for all you know—might be, are in blue cells. Here you have simply eliminated the ten polar-bear cells from the calculation. But this does not rely on the assumption that polar bears aren't included in the reference class. The calculation would come out the same if the bears were replaced with human observers who were very much like yourself, provided you knew you were not one of them. Maybe you are told that ten people who have a birthmark on their right calves are in blue cells. After verifying that you yourself don't have such a birthmark, you adjust your probability of being in a blue cell to $^{8}/_{9}$. This is in agreement with SSA. According to SSA (given that the people with the birthmarks are in the reference class), P(Blue cell | Setup) = $^{9}/_{100}$. But also by SSA, P(Blue cell | Setup & Ten of the people in blue cells have birth marks of a type you don't have) = $^{8}/_{9}$.

Where the definition of the reference class becomes an issue is when the total number of observers is unknown and is correlated with the hypotheses under consideration. Consider the following schema for producing *Incubator*-type experiments: There are two rooms. Whichever way

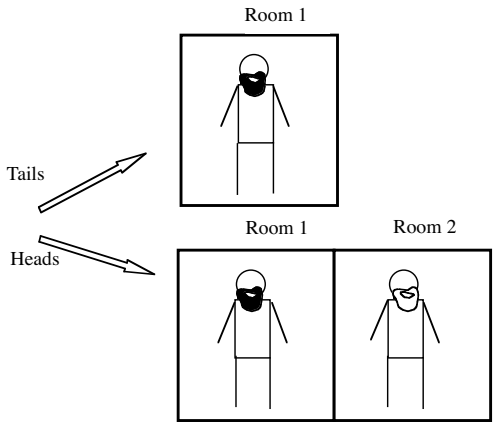

Figure 1: *Incubator*, version I

the coin falls, a person with a black beard is created in Room 1. If and only if it falls heads, then one other thing x is created in Room 2. You find yourself in one of the rooms and you are informed that it is Room 1. We can now ask, for various choices of x, what your credence should be that the coin fell heads.

The original version of *Incubator* was one where x is a man with white

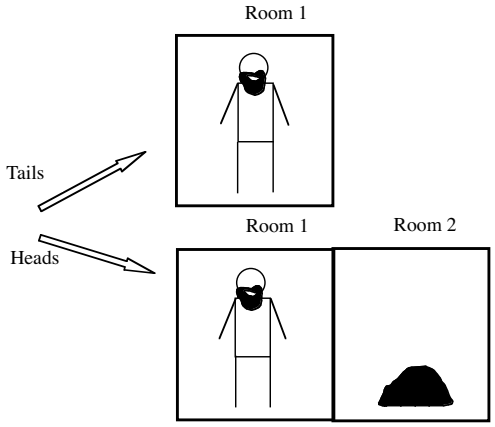

Figure 2: *Incubator*, version II

beard:

As we saw above, on Model 2 ("SSA and not SIA"), your credence of Heads is ⅓. But now consider a second case (version II) where we let x be a rock:

In version II, when you find that you are the man in Room 1, it is evident that your credence of Heads should be ½. The conditional probability of you observing what you are observing (i.e. your being the man in Room 1) is unity on both Heads and Tails, because with this setup you couldn't possibly have found yourself observing being in Room 2. (We assume, of course, that the rock does not have a soul or a mind.) Notice that the arguments used to argue for SSA in the previous examples cannot be used in version II. A rock cannot bet and cannot be wrong, so the fraction of observers who are right or would win their bets is not improved here by including rocks in the reference class. Moreover, it seems impossible to conceive of a situation where you are ignorant as to whether you are the man in Room 1 or the rock in Room 2.

If this is right then the probability you should assign to heads depends on what you know would be in Room 2 if the coin fell heads, even though you know that you are in Room 1. The reference class problem can be relevant in cases like this, where the size of the population depends on which hypothesis is true. What you should believe depends on whether the object x that would be in Room 2 would be in the reference class or not; it makes a difference to your rational credence whether x is rock or an observer like yourself.

Rocks, consequently, are not in the reference class. In a similar vein we can rule out armchairs, planets, books, plants, bacteria and other such non-observer entities. It gets trickier when we consider possible borderline cases such as a gifted chimpanzee, a Neanderthal or a mentally disabled

[5] An additional problem with the principle of indifference is that it balances precariously between vacuity and inconsistency. Starting from the generic formulation suggested earlier, "*Assign equal credence to any two hypotheses if you don't have any reason to prefer one to the other*", one can make it go either way depending on how a strong an interpretation one gives of "reason". If reasons can include any subjective inclination, the principle loses most if not all of its content. But if having a reason requires one to have objectively significant statistical data, then the principle can be shown to be inconsistent.

human. It is not immediately obvious whether the earlier arguments for including things in the reference class could be used to argue that these entities should be admitted. Can a severely mentally disabled person bet? Could you have found yourself as such a person? (Although anybody could of course in one sense become severely mentally disabled, it could be argued that the being that results from such a process would not in any real sense still be "you" if the damage is sufficiently severe.)

That these questions arise seems to suggest that something beyond a plain version of the principle of indifference is involved. The principle of indifference is primarily about what your credence should be when you are ignorant of certain facts (Castell 1998; Strevens 1998). SSA purports to determine conditional probabilities of the form P("*I'm* an observer with such and such properties" | "The world is such and such"), and it applies even when you were never ignorant of who you are and what properties you have.[5]

Intellectual insufficiency might not be the only source of vagueness or indeterminacy of the reference class. Here is a list of possible borderlines:

- *Intellectual limitations* (e.g. chimpanzees; brain-damaged persons; Neanderthals; persons who can't understand SSA and the probabilistic reasoning involved in using it in the application in question)

- *Insufficient information* (e.g. persons who don't know about the experimental setup)

- *Lack of some occurrent thoughts* (e.g. persons who, as it happens, don't think of applying SSA to a given situation although they have the capacity)

- *Exotic mentality* (e.g. angels; superintelligent computers; posthumans)

No claim is made that all of these dimensions are such that one can exit the reference class by going to a sufficiently extreme position along them. For instance, maybe an intellect cannot by disqualified for being too smart. The purpose of the list is merely to illustrate that the exact way of delimiting the reference class has not been settled by the preceding discussion and that in order to so one would have to address at least these four points.

We will return to the reference class problem in the next chapter, where we'll see that an attempted solution by John Leslie fails, and yet again in chapters 10 and 11.

For many purposes, however, the details of the definition of the refer-

CHAPTER 5

The Self-Sampling Assumption in Science

We turn to the second strand of arguments for SSA. Here we shall show that many important scientific fields implicitly rely on SSA and that it (or something much like it) constitutes an indispensable part of scientific methodology.

SSA IN COSMOLOGY

Recall our earlier hunch that the trouble in deriving observational consequences from theories that were coupled to some Big World hypothesis might originate in the somewhat "technical" point that while in a large enough cosmos, every observation will be made by *some* observers here and there, it is notwithstanding true that those observers are exceedingly rare and far between. For every observation made by a freak observer spontaneously materializing from Hawking radiation or thermal fluctuations, there are trillions and trillions of observations made by regular observers who have evolved on planets like our own, and who make veridical observations of the universe they are living in. Maybe we can solve the problem, then, by saying that although all these freak observers exist and are suffering from various illusions, it is highly unlikely that *we* are among their numbers? In this case we should think, rather, that we are very probably one of the regular observers whose observations reflect reality. We could safely ignore the freak observers and their illusions in most contexts when doing science. Because the freak observers are in such a tiny minority, their observations can usually be disregarded. It is *possible* that we are freak observers. We should assign to that hypothesis some finite probability—but such a tiny one that it doesn't make any practical difference.

To see how SSA enables us to cash in on this idea, it is first of all crucial that we construe our evidence differently than we did when originally stating the conundrum. If our evidence is simply "Such and such an

observation is made" then the evidence has probability one given any Big World theory—and we ram our heads straight into the problem that all Big World theories become impotent. But if we construe our evidence in the more specific form "*We* are making such and such observations." then we have a way out. For we can then say that although Big World theories make it probable ($P \approx 1$) that some such observations be made, they need not make it probable that we should be the ones making them.

Let us therefore define:

E' := "Such and such observations are made by us."

E' contains an indexical component that the original evidence-statement we considered, *E*, did not. *E'* is logically stronger than *E*. Since the rationality requirement that one should take all relevant evidence into account dictates that in case *E'* leads to different conclusions than does *E*, it is *E'* that determines what we ought to believe.

A question that now arises is how to determine the evidential bearing that statements of the form of *E'* have on cosmological theories. Using Bayes' theorem, we can turn the question around and ask, how do we evaluate P(*E'*| *T*&*B*), the conditional probability that a Big World theory gives to us making certain observations? The argument in chapter 3 showed that if we hope to be able to derive any empirical implications from Big World theories, then P(*E'*| *T*&*B*) should not generally be set to unity or close to unity. P(*E'*| *T*&*B*) must take on values that depend on the particular theory and the particular evidence that we are we are considering. Some theories *T* are supported by some evidence *E'*; for these choices P(*E'*| *T*&*B*) is relatively large. For other choices of *E'* and *T*, the conditional probability will be relatively small.

To be concrete, consider the two rival theories T_1 and T_2 about the temperature of the cosmic microwave background radiation. (T_1 was the theory that says that the temperature of the cosmic microwave background radiation is about 2.7 degrees K (the observed value); T_2 says it is 3.1 K.) Let *E'* be the proposition that we have made those observations that cosmologists innocently take to support T_1. *E'* includes readouts from radio telescopes, etc. Intuitively, we want $P(E'|T_1\&B) > P(E'|T_2\&B)$. That inequality must be the reason why cosmologists believe that the background radiation is in accordance with T_1 rather than T_2, since a priori there is no ground for assigning T_1 a substantially greater probability than T_2.

A natural way in which we can achieve this result is by postulating that we should think of ourselves as being in some sense "random" observers. Here we use the idea that the essential difference between T_1 and T_2 is that the *fraction* of observers who would be making observations in agreement with *E'* is enormously greater on T_1 than on T_2. If we reason as if we

were randomly selected samples from the set of all observers, or from some suitable subset thereof, then we can explicate the conditional probability P(*E'*| *T*&*B*) in terms of the expected fraction of all observers in the reference class that the conjunction of *T* and *B* says would be making the kind of observations that *E'* says that we are making. This will enable us to conclude that P(*E'*| T_1&*B*) > P(*E'*| T_2&*B*).

In order to spotlight basic principles, we can make some simplifying assumptions. In the present application, we can think of the reference class as consisting of all observers who will ever have existed. We can also assume a uniform sampling density over this reference class. Moreover, it simplifies things if we set aside complications arising from assigning probabilities over infinite domains by assuming that *B* entails that the number of observers is finite, albeit such a large finite number that the problems described earlier obtain.

Here is how SSA supplies the missing link needed to connect theories like T_1 and T_2 to observation. On T_2, the only observers who observe an apparent temperature of the cosmic microwave background CBM ≈ 2.7 K are those that have various sorts of rare illusions, for example because their brains have been generated by black holes and are therefore not attuned to the world they are living in. On T_1, by contrast, every observer who makes the appropriate astronomical measurements and is not deluded will observe CBM ≈ 2.7 K. A much greater fraction of the observers in the reference class observe CBM≈2.7 K if T_1 is true than if T_2 is true. By SSA, we consider ourselves as random observers; it follows that on T_1 we would be more likely to find ourselves as one of those observers who observe CBM ≈ 2.7 K than we would on T_2. Therefore P(*E'*| T_1&*B*) >> P(*E'*| T_2&*B*). Supposing that the prior probabilities of T_1 and T_2 are roughly the same, P(T_1) ≈ P(T_2), it is then trivial to derive via Bayes' theorem that P(T_1 |*E'*&*B*) > P(T_2 |*E'*&*B*). This vindicates the intuitive view that we do have empirical evidence that favors T_1 over T_2.

The job that SSA is doing in this derivation is to enable the step from a proposition about fractions of observers to propositions about corresponding probabilities. We get the propositions about fractions of observers by analyzing T_1 and T_2 and combining them with relevant background information *B*. From this, we conclude that there would be an extremely small fraction of observers observing CBM ≈ 2.7 K given T_2 and a much larger fraction given T_1. We then consider the evidence *E'*, which is that *we* are observing CBM ≈ 2.7 K. SSA authorizes us to think of the "we" as a kind of random variable ranging over the class of actual observers. From this it then follows that *E'* is more probable given T_1 than given T_2. But without assuming SSA, all we can say is that a greater fraction of observers observe CBM ≈ 2.7 K if T_1 is true; at that point the argument would grind to a halt. We could not reach the conclusion that T_1 is supported over T_2. Therefore,

SSA, or something like it, must be adopted as a methodological principle.

SSA IN THERMODYNAMICS

Here we'll examine Ludwig Boltzmann's famous attempt to explain why entropy is increasing in the forward time-direction. We'll show that a popular and intuitively very plausible objection against Boltzmann relies on an implicit appeal to SSA.

The outlines of Boltzmann's[1] explanation can be sketched roughly as follows. The direction of time's arrow appears to be connected to the fact that entropy increases in the forward time-direction. Now, if one assumes, as is commonly done, that low entropy corresponds in some sense to low probability, then one can see that if a system starts out in a low-entropy state then it will probably evolve over time into a higher entropy state, which, after all, is a more probable state of the system. The problem of explaining why entropy is increasing is thus reduced to the problem of explaining why entropy is currently so low. This would appear to be a priori improbable. Boltzmann points out, however, that in a sufficiently large system (and the universe may well be such a system) there are (with high probability) local regions of the system—let's call them "subsystems"—which are in low-entropy states even if the system as a whole is in a high-entropy state. Think of it like this: In a sufficiently large container of gas, there will be some places where all the gas molecules in that local region are lumped together in a small cube or some other neat pattern. That is probabilistically guaranteed by the random motion of the gas molecules together with the fact that there are so many of them. Thus, Boltzmann argued, in a large-enough universe, there will be some places and some times at which just by chance the entropy happens to be exceptionally low. Since life can only exist in a region if it has very low entropy, we would naturally find that in our part of the universe entropy is very low. And since low-entropy subsystems are overwhelmingly likely to evolve towards higher-entropy states, we thus have an explanation of why entropy is currently low here and increasing. An observation selection effect guarantees that we observe a region where that is the case, even though such regions are enormously sparse in the bigger picture.

Lawrence Sklar has remarked about Boltzmann's explanation that it has been "credited by many as one of the most ingenious proposals in the history of science, and disparaged by others as the last, patently desperate, ad hoc attempt to save an obviously failed theory" ((Sklar 1993), p. 44). I think that the ingenuity of Boltzmann's contribution should be fully granted (especially considering that when writing this in 1895, he was nearly seventy years ahead of his time in directly considering observation selection effects when reasoning about the large-scale structure of the world), but

[1] Boltzmann attributes the idea to his assistant, Dr. Schuetz. Thank heaven for postdocs.

that nonetheless the idea is flawed.

The standard objection is that Boltzmann's datum—that the observable universe is a low-entropy subsystem—turns out on a closer look to be in conflict with his explanation. It is noted that very large low-entropy regions, such as the one we observe, are very sparsely distributed if the universe as a whole is in a high-entropy state. A much smaller low-entropy region would have sufficed to permit intelligent life to exist. Boltzmann's theory fails to account for why the observed low-entropy region is so large and so grossly out of equilibrium.

This plausible objection can be fleshed out with the help of SSA. Let us follow Boltzmann and suppose that we are living in a very vast (perhaps infinite) universe which is in thermal equilibrium and that observers can exist only in low-entropy regions. Let *T* be the theory that asserts this. According to SSA, what *T* predicts we should observe depends on where *T* says that the bulk of observers tend to be. Since *T* is a theory of thermodynamic fluctuations, it implies that smaller fluctuations (i.e. low-entropy regions) are *vastly* more frequent than larger fluctuations, and hence that most observers will find themselves in rather small fluctuations. This is so because the infrequency of larger fluctuations increases rapidly enough to make sure that even though a given large fluctuation will typically contain more observers than a given small fluctuation, the previous sentence nonetheless holds true. By SSA, *T* assigns a probability to us observing what we actually observe that is proportional to the fraction of all observers it says would make that kind of observations. Since an extremely small fraction of all observers will observe a low entropy region as large as ours if *T* is true, it follows that *T* gives an extremely small probability to the hypothesis that we should observe such a large low-entropy region. Hence *T* is heavily disfavored by our empirical evidence and should be rejected unless its a priori probability was so extremely high as to compensate for its empirical implausibility. For instance, if we compare *T* with a rival theory *T**, which asserts that the average entropy in the universe as a whole is about the same as the entropy of the region we observe, then in light of the preceding argument we have to acknowledge that *T** is much more likely to be true, unless our prior probability function was severely biased towards *T*. (The bias would have to be truly extreme. It would not suffice, for example, if one's prior probabilities where P(*T*) = 99.999999% and P(*T**) = 0.000001%.) This validates the objection against Boltzmann. His anthropic explanation is refuted—probabilistically but with extremely high probability—by a more careful application of the anthropic principle. His account should therefore be modified or given up in favor of some other explanation.

Sklar, however, thinks that the Boltzmannian has a "reasonable reply" (ibid. p. 299) to this objection, namely that in Boltzmann's picture there will be *some* large regions where entropy is low, so our observations are not really incompatible with his proposal. However, while there is no log-

ical incompatibility, the *probabilistic incompatibility* is of a very high degree. This can for all practical purposes be just as decisive as a logical deduction of a falsified empirical consequence, making it totally unreasonable to accept this reply.

Sklar then goes on to state what he seems to see as the real problem for Boltzmannians:

> The major contemporary objection to Boltzmann's account is its apparent failure to do justice to the observational facts . . . as far as we can tell, the parallel direction of entropic increase of systems toward what we intuitively take to be the future time direction that we encounter in our local world seems to hold throughout the universe." (Ibid. p. 300)

It is easy to see that this is just a veiled reformulation of the objection discussed above. If there were a "reasonable reply" to the former objection, the same reply would work equally well against this reformulated version. An unreformed Boltzmannian could simply retort: "Hey, even on my theory there will be some regions and some observers in those regions for whom, as far as they can tell, entropy seems to be on the increase throughout the universe—they see only their local region of the universe after all. Hence our observations are compatible with my theory!" If we are not impressed by this reply, it is because we are willing to take probabilistic entailments seriously. Failing to do so would spell methodological disaster for any theory that postulates a sufficiently big cosmos, since according to such theories there will always be some observer somewhere who observes what we are observing, so the theories would be logically compatible with any observation we could make.[2] But that is clearly not how such theories work.

SSA IN EVOLUTIONARY BIOLOGY

Anthropic reasoning has been applied to estimate probabilistic parameters in evolutionary biology. For example, we may ask how difficult it was for intelligent life to evolve on our planet.[3] Naively, one may think that since intelligent life evolved on the only planet we have closely examined, evolution of intelligent life seems quite easy. Science popularizer Carl Sagan seems to have held this view: "the origin of life must be a highly probable circumstance; as soon as conditions permit, up it pops!" (Sagan 1995). A

[2] The only observational consequence such theories would have on that view is that we don't make observations that are logically incompatible with the laws of nature which that theory postulates. But that is much too weak to be of any use. Any finite string of sensory stimulation we could have seems to be logically compatible with the laws of nature, both in the classical mechanics framework used in Boltzmann's time and in a contemporary quantum mechanical setting.

[3] One natural way of explicating this is to think of it as asking for what fraction of all Earth-like planets actually develop intelligent life, provided they are left untouched by alien civi-

moment's reflection reveals that this inference is incorrect, since no matter how unlikely it was for intelligent life to develop on any given planet, we should still expect to have originated from a planet where such an improbable sequence of events took place. As we saw in chapter 2, the theories that are disconfirmed by the fact that intelligent life exists here are those according to which the difficulty of evolving intelligent life is so great that they give a small likelihood to there being even a single planet with intelligent life in the whole world.

Brandon Carter (Carter 1983; Carter 1989) combines this realization with some additional assumptions and argues that the chance that intelligent life will evolve on a given Earth-like planet is in fact very small. His argument is outlined in this footnote.[4]

Carter has also suggested a clever way of estimating the number of

[4] Define the three time intervals: $\bar{t}$, "the expected average time . . . which would be intrinsically most likely for the evolution of a system of 'intelligent observers', in the form of a scientific civilization such as our own" (Carter 1983), p. 353); t_e, which is the time taken by biological evolution on this planet $\approx 0.4 \times 10^{10}$ years; and τ_0, the lifetime of the main sequence of the sun $\approx 10^{10}$ years.

The argument in outline runs as follows: Since at the present stage of understanding in biochemistry and evolutionary biology we have no way of making even an approximate calculation of how likely the evolution of intelligent life is on a planet like ours, we should use a very broad prior probability distribution for this. We can partition the range of possible values of $\bar{t}$ roughly into three regions: $\bar{t} \ll \tau_0$, $\bar{t} \approx \tau_0$, or $\bar{t} \gg \tau_0$. Of these three possibilities we can, according to Carter, "rule out" the second one a priori, with fairly high probability, since it represents a very narrow segment of the total hypothesis space, and since a priori there is no reason to suppose that the expected time to evolve intelligent life should be correlated with the duration of the main sequence of stars like the sun. But we can also rule out (with great probability) the first alternative, since if the expected time to evolve intelligent life were much smaller than τ_0, then we would expect life to have evolved much earlier than it in fact did. This leaves us with $\bar{t} \gg \tau_0$, meaning that life was very unlikely to evolve as fast as it did, within the lifetime of the main sequence of the sun.

What drives this conclusion is the near coincidence between t_e and τ_0 where we would a priori have no reason to suppose that these two quantities would be within an order of magnitude (or even within a factor of about two) from each other. This fact is combined with an observation selection effect to yield the prediction that the evolution of intelligent life is very unlikely to happen on a given planet within the main sequence of its star. The contribution that the observation selection effect makes is that it prevents observations of intelligent life taking *longer* than τ_0 to evolve. Whenever intelligent life evolves on a planet we must find that it evolved before its sun went extinct. Were it not for the fact that the only evolutionary processes that are observed first-hand are those which gave rise to intelligent observers in a shorter time than τ_0, then the observation that $t_e \approx \tau_0$ would have disconfirmed the hypothesis that $\bar{t} \gg \tau_0$ just as much as it disconfirmed $\bar{t} \gg \tau_0$. But thanks to this selection effect, $t_e \approx \tau_0$ is precisely what one would expect to observe even if the evolutionary process leading to intelligent life were intrinsically very unlikely to take place in as short a time as τ_0.

Patrick Wilson (Wilson 1994) advances some objections against Carter's reasoning, but as these objections do not concern the basic anthropic methodology that Carter uses, they don't need to be addressed here.

A corollary of Carter's conclusion is that there very probably aren't any extraterrestrial civilizations anywhere near us, maybe not even in our galaxy.

improbable "critical" steps in the evolution of humans. A little story may provide the easiest way to grasp the idea: A princess is locked in a tower. Suitors have to pick five combination locks to get to her, and they can do this only through random trial and error, i.e. without memory of which combinations they have tried. A suitor gets one hour to pick all five locks. If he doesn't succeed within the allotted time, he is shot. However, the princess' charms are such that there is an endless line of hopeful suitors waiting their turn.

After the deaths of some unknown number of suitors, one of them finally passes the test and marries the princess. Suppose that the numbers of possible combinations in the locks are such that the expected time to pick each lock is .01, .1, 1, 10, and 100 hours respectively. Suppose that pick-times for the suitor who got through are (in hours) {.00583, .0934, .248, .276, .319}. By inspecting this set you could reasonably guess that .00583 hour was the pick-time for the easiest lock and .0934 hour the pick-time for the second easiest lock. However, you couldn't really tell which locks the remaining three pick-times correspond to (Hanson 1998). This is a typical result. When conditioning on success before the cut-off (in this case 1 hour), the average completion time of a step is nearly independent of its expected completion time provided the expected completion time is much longer than the cut-off. Thus, for example, even if the expected pick-time of one of the locks had been a million years, you would still find that its average pick-time *in successful runs* is closer to .2 or .3 than to 1 hour, and you wouldn't be able to tell it apart from the 1, 10, and 100 hours locks.

If we don't know the expected pick-times or the number of locks that the suitor had to break, we can obtain estimates of these parameters if we know the time it took him to reach the princess. The less surplus time left over before the cut-off, the greater the number of difficult locks he had to pick. For example, if the successful suitor took 59 minutes to get to the princess, then that would favor the hypothesis that he had to pick a fairly large number of locks. If he reached the princess in 35 minutes, that would strongly suggest that the number of difficult locks was small. The relation also works the other way around so that if we are not sure what the maximum allowed is, it can be estimated using information about the number of difficult locks and their combined pick-time in a random successful trial. Monte Carlo simulations confirming these claims can be found in (Hanson 1998), which also derives some analytical expressions.

Carter applies these mathematical ideas to evolutionary theory by noting that an upper bound on the cut-off time after which intelligent life could not have evolved on Earth is given by the duration of the main sequence of the sun—about $10*10^9$ years. It took about $4*10^9$ years for intelligent life to develop. From this (together with some other assumptions which are problematic but not in ways relevant for our purposes) Carter concludes that the number of critical steps in human evolution is likely very small—not much greater than two.

a critical
ntelligent
e precur-
difficult.

One potential problem with Carter's argument is that the duration of the

main sequence of the sun only gives an upper bound on the cut-off; maybe climate change or some other type of event would have made Earth unconducive to evolution of complex organisms long before the sun becomes a red giant. Recognizing this possibility, Barrow and Tipler (Barrow and Tipler 1986) apply Carter's reasoning in the opposite direction and seek to infer the true cut-off time by directly estimating the number of critical steps.[5] In a recent paper, Robin Hanson (Hanson 1998) scrutinizes Barrow and Tipler's suggestions for what are the critical steps and argues that their model does not fit the evidence very well when considering the relative time the various proposed critical steps actually took to complete.

Our concern here is not which estimate is correct or even whether at the current state of biological science enough empirical data and theoretical understanding are available to supply the substantive premises needed to derive any specific conclusion from the sort of considerations described in this section.[6] My contention, rather, is twofold. Firstly, if one wants to argue about or make a claim regarding such things as the improbability of intelligent life evolving, or the probability of finding extraterrestrial life, or the number of critical steps in human evolution, or the planetary window of opportunity during which evolution of intelligent life is possible, then one has to make sure that one's position is coherent. The work by Carter and others reveals subtle ways in which some views on these things are probabilistically incoherent. Secondly, underlying the basic constraints

[5] For example, the step from prokaryotic to eukaryotic life is a candidate for being a critical step, since it seems to have happened only once and appears to be necessary for intelligent life to evolve. By contrast, there is evidence that the evolution of eyes from an "eye precursor" has occurred independently at least forty times, so this step does not seem to be difficult. A good introduction to some of the relevant biology is (Schopf 1992).

[6] There are complex empirical issues that would need to be confronted were one to the seriously pursue an investigation into these questions. For instance, if a step takes a very long time, that *may* suggest that the step was very difficult (perhaps requiring simultaneous muliloci mutations or other rare occurrences). But there can be other reasons for a step taking long to complete. For example, oxygen breathing took a long time to evolve, but this is not a ground for thinking that it was a difficult step. For oxygen breathing became adaptive only after there were significant levels of free oxygen in the atmosphere, and it took anaerobic organisms hundreds of millions of years to produce enough oxygen to satiate various oxygen sinks and raise the levels of atmospheric oxygen to the required levels. This process was very slow but virtually guaranteed to run to completion eventually, so it would be a mistake to infer that the evolution of oxygen breathing and the concomitant Cambrian explosion represent a hugely difficult step in human evolution.—Likewise, that a step took only a short time (as, for instance, the transition from our ape ancestors to homo sapiens) *can* be evidence suggesting it was relatively easy, but it need not be if we suspect that there was only a small window of opportunity for the step to occur (so that if it occurred at all, it would have to happen within that time-interval).

[7] In case of an infinite (or extremely large finite) cosmos, intelligent life would also evolve after the "cut-off". Normally we may feel quite confident in stating that intelligent life cannot

appealed to in Carter's reasoning (and this is quite independent of the specific empirical assumptions he needs to get any concrete results) is an application of SSA. WAP and SAP are inadequate in these applications. SSA makes its entrée when we realize that in a large universe there will be actual evolutionary histories of most any sort. On some planets, life will evolve swiftly; on others it will use up all the time available before the cut-off.[7] On some planets, difficult steps will be completed more quickly than easy steps. Without some probabilistic connection between the distribution of evolutionary histories and our own observed evolutionary past, none of the above considerations would even make sense.

SSA is not the only methodological principle that would establish such a connection. For example, we could formulate a principle stating that every *civilization* should reason as if it were a random sample from the set of all civilizations.[8] For the purposes of the above anthropic arguments in evolution theory this principle would amount to the same thing as the SSA, provided that all civilizations contained the same number of observers. However, when considering hypotheses on which certain types of evolutionary histories are correlated with the evolved civilizations containing a greater or smaller number of observers, this principle is not valid. We would then have to take recourse to the more generally valid principle given by SSA.

SSA IN TRAFFIC ANALYSIS

When driving on the motorway, have you ever wondered about (and cursed) the phenomenon that cars in the other lane appear to be getting ahead faster than you? Although one may be inclined to account for this by invoking Murphy's Law[9], a recent paper in *Nature* (Redelmeier and Tibshirani 1999), further elaborated in (Redelmeier and Tibshirani 2000) seeks a deeper explanation. According to this view, drivers suffer from systematic illusions causing them to mistakenly think they would have been better off in the next lane. Here we show that their argument fails to take into account an important observation selection effect. Cars in the next lane actually do go faster.

In their paper, Redelmeier and Tibshirani present some evidence that drivers on Canadian roadways (which don't have an organized laminar flow) think that the next lane is typically faster. The authors seek to explain

in chapter 3 can of course be extended to show that in an infinite universe there would with probability one be some red giants that enclose a region where—because of some ridiculously improbable statistical fluke—an Earth-like planet continues to exist and develop intelligent life. Strictly speaking, it is not impossible but only highly improbable that life will evolve on any given planet after its orbit has been swallowed by an expanding red giant.

[8] Such a principle would be very similar to what Alexander Vilenkin has (independently) called the "principle of mediocrity" (Vilenkin 1995).

[9] "If anything can go wrong, it will." (Discovered by Edward A. Murphy, Jr., in 1949.)

[10] Fo
Gilov
al. 19

this phenomenon by appealing to a variety of psychological factors. For example, "a driver is more likely to glance at the next lane for comparison when he is relatively idle while moving slowly"; "Differential surveillance can occur because drivers look forwards rather than backwards, so vehicles that are overtaken become invisible very quickly, whereas vehicles that overtake the index driver remain conspicuous for much longer"; and "human psychology may make being overtaken (losing) seem more salient than the corresponding gains". The authors recommend that drivers should be educated about these effects so as to discourage them from giving in to small temptations to switch lanes, thereby reducing the risk of accidents.

While all these illusions may indeed occur[10], there is a more straightforward explanation of the phenomenon. It goes as follows. One frequent cause of why a lane (or a segment of a lane) is slow is that there are too many cars in it. Even if the ultimate cause is something else (e.g. road work) there is nonetheless typically a negative correlation between the speed of a lane and how densely packed are the vehicles driving in it. That suggests (although it doesn't logically imply) that a disproportionate fraction of the average driver's time is spent in slow lanes. And by SSA, that means that there is a greater than even prior probability of that holding true about you in particular.

The last explanatory link can be tightened up further if we move to a stronger version of the SSA replaces "observer" with "observer-moment" (i.e. a time-segment of an observer). If you think of your present observation, when you are driving on the motorway, as a random sample from all observations made by drivers, then chances are that your observation will be made from the viewpoint that most observers have, which is the viewpoint of the slow-moving lane. In other words, appearances are faithful: more often than not, the "next" lane *is* faster! (We will discuss this stronger principle, which we'll denote "SSSA", in depth in chapter 10; the invocation of it here is just an aside.)

Even when two lanes have the same average speed, it can be advantageous to switch lanes. For what is relevant to a driver who wants to reach her destination as quickly as possible is not the average speed of the lane as a whole, but rather the speed of some segment extending maybe a couple of miles forwards from the driver's current position. More often than not, the next lane has a higher average speed at this scale than does the driver's present lane. On average, there is therefore a benefit to switching lanes (which of course has to be balanced against the costs of increased levels of effort and risk). Adopting a thermodynamics perspective, it is easy to see that (at least in the ideal case) increasing the "diffusion rate" (i.e. the probability of lane-switching) will speed the approach to "equilibrium" (i.e.

[10] For some relevant empirical studies, see e.g. (Feller 1966; Tversky and Kahneman 1981; Gilovich, Vallone et al. 1985; Larson 1987; Tversky and Kahneman 1991; Angrilli, Cherubini et al. 1997; Snowden, Stimpson et al. 1998; Walton and Bathurst 1998).

equal velocities in both lanes), thereby increasing the road's throughput and the number of vehicles that reach their destinations per unit time.

The mistake one must avoid is ignoring the selection effect residing in the fact that when you randomly select a driver and ask her whether she thinks the next lane is faster, more often than not you will have selected a driver in the lane which is in fact slower. And if there is no random selection of a driver, but it is just yourself wondering why you are so unlucky as to be in the slow lane, then the selection effect is an observational one.

SSA IN QUANTUM PHYSICS

One of the fundamental problems in the interpretation of quantum physics is how to understand the probability statements that the theory makes. On one kind of view, the "single-history version", quantum physics describes the "propensities" or physical chances of a range of possible outcomes, but only one series of outcomes actually occurs. On an alternative view, the "many-worlds version", all possible sequences of outcomes (or at least all that have nonzero measure) actually occur. These two kinds of views are often thought to be observationally indistinguishable (Wheeler 1957; DeWitt 1970; Omnès 1973), but, depending on how they are fleshed out, SSA may provide a method of telling them apart experimentally. What follows are some sketchy remarks about how such an observational wedge could be inserted. We're sacrificing rigor and generality in this section in order to keep things brief and simple.

The first problem faced by many-worlds theories is how to connect statements about the measure of various outcomes with statements about how probable we should think it is that we will observe a particular outcome. Consider first this simpleminded way of thinking about the many-worlds approach: When a quantum event E occurs in a quantum system in state S, and there are two possible outcomes A and B, then the wavefunction of S will after the event contain two components or "branches", one were A obtains and one where B obtains, and these two branches are in other respects equivalent. The problem with this view is that it fails to give a role to the amplitude of the wavefunction. If nothing is done with the fact that one of the branches (say A) might have a higher amplitude squared (say ⅔) than does the other branch, then we've lost an essential part of quantum theory, namely that it specifies not just what can happen but also the probabilities of the various possibilities. In fact, if there are equally many observers on the branch were A obtains as on the branch were B obtains, and if there is no other relevant difference between these branches, then by SSA the probability that you should find yourself on branch A is ½, rather than ⅔ as asserted by quantum physics. This simpleminded interpretation must therefore be rejected.

One way of trying to improve the interpretation would be to postulate that when the measurement occurs, the wavefunction splits into more than

two branches. Suppose, for example, that there are two branches where A obtains and one branch were B obtains (and that these branches are otherwise equivalent). Then, by SSA, you'd have a ⅔ probability of observing A—the correct answer. If one wanted to adopt this interpretation, one would have to stipulate that there are lots of branches. One could represent this interpretation pictorially as a tree, where a thick bundle of fibers in the trunk gradually split off into branches of varying degrees of thickness. Each fiber would represent one "world". When a quantum event occurs in one branch, the fibers it contains would divide into smaller branches, with the number of fibers going into each sub-branch being proportional to the amplitude squared of the wave function. For example, ⅔ of all the fibers on a branch where the event E occurs in system S would go into a sub-branch where A obtains, and ⅓ into the sub-branch where B obtains. In reality, if we wanted to hold on to the exact real-valued probabilities given by quantum theory, we'd have to postulate a continuum of fibers, so it wouldn't really make sense to speak of different fractions of fibers going into different branches, but something of the underlying ontological picture could possibly be retained so that we could speak of the more probable outcomes as obtaining "in more worlds" in some generalized sense of that expression.

Alternatively, a many-worlds interpretation could simply decide to take the correspondence between quantum mechanical measure and the probability of one observing the correlated outcome as a postulated primitive. It would then be assumed that, as a brute fact, you are more likely to find yourself on one of the branches of higher measure. (Maybe one could speak of such higher-measure branches as having a "higher degree of reality".)

On either of these alternatives, there are observational consequences that diverge from those one gets if one accepts the single-history interpretation. These consequences come into the light when one considers quantum events that lead to different numbers of observers. This was recently pointed out by Don N. Page (Page 1999). The point can be made most simply by considering a quantum cosmological toy model:

World 1: Observers; measure or probability 10^{-30}

World 2: No observers; measure or probability $1-10^{-30}$

The single-history version predicts with overwhelming probability ($P = 1-10^{-30}$) that World 2 would be the (only) realized world. If we exist, and consequently World 1 has been realized, this gives us strong reasons for rejecting the single-history version, given this particular toy model. By contrast, on the many-worlds version, both World 1 and World 2 exist, and since World 2 has no observers, what is predicted (by SSA) is that we

should observe World 1, notwithstanding its very low measure. In this example, if the choice is between the single-history version and the many-worlds version, we should therefore accept the latter.

Here's another toy model:

World A: 10^{10} observers; measure or probability $1\text{-}10^{-30}$

World B: 10^{50} observers; measure or probability 10^{-30}

In this model, finding that we are in World B does not logically refute the single-history version, but it does make it extremely improbable. For the single-history gives a conditional probability of 10^{-30} to us observing World B. The many-worlds version, on the other hand, gives a conditional probability of approximately 1 to us observing World B.[11] Provided, then, that our subjective prior probabilities for the single-history and the many-worlds versions are in the same (very big) ballpark, we should in this case again accept the latter. (The opposite would hold, of course, if we found that we are living in World A.)

These are toy models, sure. In practice, it will no doubt be hard to get a good grip on the measure of "worlds". A few things should be noted though. First, the "worlds" to which we need assign measures needn't be temporally unlimited; we could instead focus on smaller "world-parts" that arose from, and got their measures from, some earlier quantum event whose associated measures or probabilities we think we know. Such an event could, for instance, be a hypothetical symmetry-breaking event in an early inflationary epoch of our universe, or it could be some later occurrence which influences how many observers there will be (we'll study in depth some cases of this kind in chapter 9). Second, the requisite measures may be provided by other theories so that the conjunction of such theories with either the single-history or the many-worlds versions may be empirically testable. For example, Page performs some illustrative calculations using the Hartle-Hawking "no-boundary" proposal and some other assumptions. Third, since in many quantum cosmological models, the difference in the number of observers existing in different worlds can be quite huge, we might get results that are robust for a rather wide range of

[11]

$$P = \frac{10^{50} \cdot 10^{-30}}{10^{50} \cdot 10^{-30} + 10^{10} \cdot (1 - 10^{-30})} \approx 1$$

[12] On some related issues, see especially (Leslie 1996; Page 1996; Page 1997) but also (Albert 1989; Papineau 1995; Tegmark 1996; Papineau 1997; Schmidhuber 1997; Tegmark 1997; Olum 2002). Page has independently developed a principle he calls the "Conditional Aesthemic Principle", which is a sort of special-case version of SSSA applied to quantum physics.

plausible measures that the component worlds might have. And fourth, as far as our project is concerned, the important point is that our methodology ought to be able to make this kind of consideration intelligible and meaningful, whether or not at the present time we have enough data to put it into practice.[12]

SUMMARY OF THE CASE FOR SSA

In the last chapter, we argued through a series of thought experiments for reasoning in accordance with SSA in a wide range of cases. We showed that while the problem of the reference class is sometimes irrelevant when all hypotheses under consideration imply the same number of observers, the definition of the reference class becomes crucial when different hypotheses entail different numbers of observers. In those cases, what probabilistic conclusions we can draw depends on what sort of things are included in the reference class, even if the observer doing the reasoning knows that she is not one of the contested objects. We argued that many types of entities should be excluded from the reference class (rocks, bacteria, buildings, plants etc.). We also showed that variations in regard to many quite "deep-going" properties (such as gender, genes, social status etc.) are not sufficient grounds for discrimination when determining membership in the reference class. Observers differing in any of these respects can at least in some situations belong to the same reference class.

In this chapter, a complementary set of arguments was presented, focusing on how SSA caters to a methodological need in science by providing a way of connecting theory to observation. The scientific applications we looked at included:

- Deriving observational predictions from contemporary cosmological models.

- Evaluating a common objection against Boltzmann's proposed thermodynamic explanation of time's arrow.

- Identifying probabilistic coherence constraints in evolutionary biology. These are crucial in a number of contexts, such as when asking questions about the likelihood of intelligent life evolving on an Earth-like planet, the number of critical steps in human evolution, the existence of extraterrestrial intelligent life, and the cut-off time after which the evolution of intelligent life would no longer have been possible on Earth.

- Analyzing claims about perceptual illusions among drivers.

- Realizing a potential way of experimentally distinguishing between single-history and many-worlds versions of quantum theory.

Any proposed rival to SSA should be tested in all the above thought experiments and scientific applications. Anybody who is not convinced that something like SSA is needed is hereby challenged to propose a simpler or more plausible method of reasoning that works in all these cases. *Something* is evidently required, since (for instance) Big-World models are so central in contemporary science.

Our survey of applications is by no means exhaustive. We shall now turn to a purported application of SSA to evaluating hypotheses about humankind's prospects. Here we are entering controversial territory where it is not obvious whether or how SSA can be applied, or what conclusions to derive from it. Indeed, the ideas we begin to pursue at this point will eventually lead us (in chapter 10) to propose important revisions to SSA. But we have to take one step at a time.