The Design Inference

How can we identify events due to intelligent causes and distinguish them from events due to undirected natural causes? If we lack a causal theory, how can we determine whether an intelligent cause acted? This book presents a reliable method for detecting intelligent causes: the design inference.

The design inference uncovers intelligent causes by isolating the key trademark of intelligent causes: specified events of small probability. Just about anything that happens is highly improbable, but when a highly improbable event is also specified (i.e., conforms to an independently given pattern) undirected natural causes lose their explanatory power. Design inferences can be found in a range of scientific pursuits from forensic science to research into the origins of life to the search for extraterrestrial intelligence.

This challenging and provocative book shows how incomplete undirected causes are for science and breathes new life into classical design arguments. It will be read with particular interest by philosophers of science and religion, other philosophers concerned with epistemology and logic, probability and complexity theorists, and statisticians.

"As the century and with it the millennium come to an end, questions long buried have disinterred themselves and come clattering back to intellectual life, dragging their winding sheets behind them. Just what, for example, is the origin of biological complexity and how is it to be explained? We have no more idea today than Darwin did in 1859, which is to say no idea whatsoever. William Dembski's book is not apt to be the last word on the inference to design, but it will surely be the first. It is a fine contribution to analysis, clear, sober, informed, mathematically sophisticated and modest. Those who agree with its point of view will read it with pleasure, and those who do not will ignore it at their peril."

David Berlinski, Author of *The Tour of the Calculus*

Cambridge Studies in Probability, Induction, and Decision Theory

General editor: Brian Skyrms

Ellery Eells, *Probabilistic Causality*
Richard Jeffrey, *Probability and the Art of Judgment*
Robert C. Koons, *Paradoxes of Belief and Strategic Rationality*
Cristina Bicchieri and Maria Luisa Dalla Chiara (eds.), *Knowledge, Belief, and Strategic Interactions*
Patrick Maher, *Betting on Theories*
Cristina Bicchieri, *Rationality and Coordination*
J. Howard Sobel, *Taking Chances*
Jan von Plato, *Creating Modern Probability: Its Mathematics, Physics, and Philosophy in Historical Perspective*
Ellery Eells and Brian Skyrms (eds.), *Probability and Conditionals*
Cristina Bicchieri, Richard Jeffrey, and Brian Skyrms (eds.), *The Dynamics of Norms*
Patrick Suppes and Mario Zanotti, *Foundations of Probability with Applications*
Paul Weirich, *Equilibrium and Rationality*
Daniel Hausman, *Causal Asymmetries*

The Design Inference

Eliminating Chance Through Small Probabilities

William A. Dembski

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
40 West 20th Street, New York, NY 10011–4211, USA
www.cambridge.org
Information on this title: www.cambridge.org/9780521623872

First published 1998
Reprinted 1999 (twice), 2000, 2001
First paperback edition 2005

Printed in the United States of America

A catalogue record for this book is available from the British Library.

Library of Congress Cataloguing in Publication data

Dembski, William A., 1960 —
The design inference: eliminating chance through small probabilities
/ William A. Dembski.
p. cm. — (Cambrige studies in probability induction and
decision theory)
Includes bibliographical reference.
ISBN 0-521-62387-1 (hb)
1. Experimental design. 2. Probabilities. I. Title. II. Series
Q A279.D455 1998
001.4'34 — dc21 98-3020
CIP

ISBN-13 978-0-521-62387-2 hardback
ISBN-10 0-521-62387-1 hardback

ISBN-13 978-0-521-67867-4 paperback
ISBN-10 0-521-67867-6 paperback

The same Arguments which explode the Notion of Luck, may, on the other side, be useful in some Cases to establish a due comparison between Chance and Design: We may imagine Chance and Design to be, as it were, in Competition with each other, for the production of some sorts of Events, and may calculate what Probability there is, that those Events should be rather owing to one than to the other.

—Abraham de Moivre, *Doctrine of Chances*, 1718

To my parents, William J. and Ursula Dembski, Proverbs 1:8–9

Contents

Preface

Highly improbable events don't happen by chance. Just about everything that happens is highly improbable. Both claims are correct as far as they go. The aim of this monograph is to show just how far they go. In *Personal Knowledge* Michael Polanyi (1962, p. 33) considers stones placed in a garden. In one instance the stones spell "Welcome to Wales by British Railways," in the other they appear randomly strewn. In both instances the precise arrangement of stones is vastly improbable. Indeed, any given arrangement of stones is but one of an almost infinite number of possible arrangements. Nevertheless, arrangements of stones that spell coherent English sentences form but a minuscule proportion of the total possible arrangements of stones. The improbability of such arrangements is not properly referred to chance.

What is the difference between a randomly strewn arrangement and one that spells a coherent English sentence? Improbability by itself isn't decisive. In addition what's needed is conformity to a pattern. When the stones spell a coherent English sentence, they conform to a pattern. When they are randomly strewn, no pattern is evident. But herein lies a difficulty. Everything conforms to some pattern or other – even a random arrangement of stones. The crucial question, therefore, is whether an arrangement of stones conforms to the right sort of pattern to eliminate chance.

This monograph presents a full account of those patterns capable of successfully eliminating chance. Present statistical theory offers only a partial account of such patterns. To eliminate chance the statistician sets up a *rejection region* prior to an experiment. If the outcome of the experiment falls within the rejection region, chance is eliminated. Rejection regions are patterns given prior to an event. Although such patterns successfully eliminate chance, they are by no means the only ones. Detectives, for instance, routinely uncover patterns after the fact – patterns identified only after a crime has been committed, and which effectively preclude attributing the crime to chance (cf. Dornstein, 1996; Evans, 1996).

Although improbability is not a sufficient condition for eliminating chance, it is a necessary condition. Four heads in a row with a fair coin is sufficiently probable as not to raise an eyebrow; four hundred heads in a row is a different story. But where is the cutoff? How small a probability is small enough to eliminate chance? The answer depends on the relevant number of opportunities for patterns and events to coincide – or what I call the relevant *probabilistic resources.* A toy universe with only 10 elementary particles has far fewer probabilistic resources than our own universe with 10^{80}. What is highly improbable and not properly attributed to chance within the toy universe may be quite probable and reasonably attributed to chance within our own universe.

Eliminating chance is closely connected with design and intelligent agency. To eliminate chance because a sufficiently improbable event conforms to the right sort of pattern is frequently the first step in identifying an intelligent agent. It makes sense, therefore, to define design as "patterned improbability," and the design inference as the logic by which "patterned improbability" is detected and demonstrated. So defined, the design inference stops short of delivering a causal story for how an intelligent agent acted. But by precluding chance and implicating intelligent agency, the design inference does the next best thing.

Who will want to read this monograph? Certainly anyone interested in the logic of probabilistic inferences. This includes logicians, epistemologists, philosophers of science, probabilists, statisticians, and computational complexity theorists. Nevertheless, a much broader audience has a vital stake in the results of this monograph. Indeed, anyone who employs small-probability chance-elimination arguments for a living will want to know the results of this monograph. The broader audience of this work therefore includes forensic scientists, SETI researchers, insurance fraud investigators, debunkers of psychic phenomena, origins-of-life researchers, intellectual property attorneys, investigators of data falsification, cryptographers, parapsychology researchers, and programmers of (pseudo-) random number generators.

Although this is a research monograph, my aim throughout has been to write an interesting book, and one that as little as possible duplicates existing texts in probability and statistics. Even the most technical portions of this monograph will be interlaced with examples accessible to nontechnical readers. I therefore encourage nontechnical readers to read this monograph from start to finish, skipping the technical portions. Readers with a background in probability theory, on the other hand, I encourage to read this monograph thoroughly from start to finish. Small-probability arguments are widely abused and misunderstood. In the analysis of small-probability arguments the devil is in the details. Because this monograph constitutes a sustained argument, a sustained reading would be optimal. Nevertheless, for the reader with limited time, I suggest reading Sections 1.1–1.2, 2.1–2.2, 5.1–5.4, and all of Chapter 6 in order, referring to Chapters 3 and 4 as needed.

This monograph is organized as follows. Chapter 1 is an examples chapter showing just how prevalent design inferences are. Chapter 2 in turn provides an overview of the logical structure of the design inference. Chapters 1 and 2 are the least technical and introduce the design inference. Chapters 3 and 4 present the twin pillars on which the design inference is based, namely, probability theory and complexity theory. With the technical apparatus of Chapters 3 and 4 in place, what remains is to elucidate the dual notions of specification and probabilistic resources. Chapter 5 treats specification (i.e., the type of pattern needed to eliminate chance). Chapter 6 treats probabilistic resources (i.e., the degree of improbability needed to eliminate chance). Together Chapters 5 and 6 yield a precise formulation of the design inference.

Acknowledgment

The beginnings of this monograph can be traced back to an interdisciplinary conference on randomness in the spring of 1988 at Ohio State University, at a time when I was just finishing my doctoral work in mathematics at the University of Chicago. Persi Diaconis and Harvey Friedman were the conveners of this conference. I shall always be grateful to Persi for bringing this conference to my attention and for urging me to attend it. Indeed, most of my subsequent thinking on the topics of randomness, probability, complexity, and design has germinated from the seeds planted at this conference.

The main mathematical tools employed in this monograph are probability theory and a generalization of computational complexity theory. Probability theory was the subject of my doctoral dissertation in mathematics at the University of Chicago (1988). This dissertation was completed under the direction of Patrick Billingsley and Leo Kadanoff. As for computational complexity theory, I was introduced to it during the academic year 1987–88, a year devoted to cryptography at the computer science department of the University of Chicago. Jeff Shallit, Adi Shamir, and Claus Schnorr were present that year and helped me gain my footing. Subsequently I had the good fortune of receiving a postdoctoral fellowship in mathematics from the National Science Foundation (DMS-8807259). This grant allowed me to pursue complexity theory further at the computer science department of Princeton University (Fall 1990). My sponsor, Andrew Yao, made my stay at Princeton not only possible but also congenial.

The monograph itself is a revised version of my philosophy dissertation from the University of Illinois at Chicago (1996). This dissertation was completed under the direction of Charles Chastain and Dorothy Grover. I want to commend Charles and Dorothy as well as

Michael Friedman and Walter Edelberg for the effort they invested in my philosophical development. This monograph would be much poorer without their insights and criticisms. I also wish to thank individual members of the philosophy departments at the University of Chicago and Northwestern University. On the south side, David Malament and Bill Wimsatt stand out. Their willingness to read my work, encourage me, and help me in such practical matters as getting a job went well beyond the call of duty. On the north side, I have profited enormously from seminars and discussions with Arthur Fine, David Hull, and Tom Ryckman. Tom Ryckman deserves special mention here, not only for his generous spirit, but also for his example of philosophical integrity.

Many people have shaped and clarified my thinking about design inferences over the past seven years. In this respect I want to single out my coconspirators in design, Stephen Meyer and Paul Nelson. As the coauthors of a planned volume for which the present monograph will serve as the theoretical underpinnings, they have exercised an unmatched influence in shaping my thoughts about design inferences. Their continual prodding and testing of my ideas have proved a constant source of refreshment to my own research. I cannot imagine the present monograph without them. This monograph, as well as the joint work with Meyer and Nelson, has been supported by a grant from the Pascal Centre at Redeemer College, Ancaster, Ontario, Canada, and by a grant from the Center for the Renewal of Science and Culture at the Discovery Institute, Seattle.

Others who have contributed significantly to this monograph include Diogenes Allen, Douglas Axe, Stephen Barr, Michael Behe, Walter Bradley, Jon Buell, John Angus Campbell, Lynda Cockroft, Robin Collins, Pat Detwiler, Frank Döring, Herman Eckelmann, Fieldstead & Co., Hugh Gauch, Bruce Gordon, Laurens Gunnarsen, Mark Hartwig, Stanley Jaki, Jon Jarrett, Richard Jeffrey, Phillip Johnson, Bob Kaita, Dean Kenyon, Saul Kripke, Robert Mann, John Warwick Montgomery, J. P. Moreland, Robert Newman, James Parker III, Alvin Plantinga, Philip Quinn, Walter J. ReMine, Hugh Ross, Siegfried Scherer, Brian Skyrms, Paul Teller, Charlie Thaxton, Jitse van der Meer, J. Wentzel van Huyssteen, Howard van Till, Jonathan Wells, C. Davis Weyerhaeuser, John Wiester, A. E. Wilder-Smith, Leland Wilkinson, Mark Wilson, Kurt Wise, and Tom Woodward.

Finally, I wish to commend three members of my family: my parents William J. and Ursula Dembski, to whom this monograph is dedicated, and without whose patience, kindness, and support this work would never have gotten off the ground; and my wife Jana, who entered my life only after this work was largely complete, but whose presence through the final revisions continually renews my spirit.

1

Introduction

1.1 HISTORICAL OVERVIEW

Eliminating chance through small probabilities has a long history. In his dialogue on the nature of the gods, Cicero (46 BC, p. 213) remarked, "If a countless number of copies of the one-and-twenty letters of the alphabet, made of gold or what you will, were thrown together into some receptacle and then shaken out on to the ground, [would it] be possible that they should produce the *Annals* of Ennius? ... I doubt whether chance could possibly succeed in producing even a single verse!"

Eighteen centuries later the Marquis Pierre Simon de Laplace (1814, p. 1307) would question whether Cicero's method of randomly shaking out letters could produce even a single word: "On a table we see letters arranged in this order, *Constantinople*, and we judge that this arrangement is not the result of chance, not because it is less possible than the others, for if this word were not employed in any language we should not suspect it came from any particular cause, but this word being in use among us, it is incomparably more probable that some person has thus arranged the aforesaid letters than that this arrangement is due to chance." A whole book, a single verse, nay, even a long word are so unlikely that we attribute their occurrence to something other than chance.

To show the absurdity of maintaining chance in the face of small probabilities Thomas Reid (1780, p. 52) asked: "If a man throws dies and both turn up aces, if he should throw 400 times, would chance throw up 400 aces? Colors thrown carelessly upon a canvas may come up to appear as a human face, but would they form a picture beautiful as the pagan Venus? A hog grubbing in the earth with his snout may turn up something like the letter A, but would he turn up the words of a complete sentence?" The answer to each question obviously is "no."

In the preface to his classic treatise on gambling Abraham de Moivre (1718, p. v) opposes chance to a mode of explanation he calls "design":

> The same Arguments which explode the Notion of Luck, may, on the other side, be useful in some Cases to establish a due comparison between Chance and Design: We may imagine Chance and Design to be, as it were, in Competition with each other, for the production of some sorts of Events, and may calculate what Probability there is, that those Events should be rather owing to one than to the other. To give a familiar Instance of this, Let us suppose that two Packs of Piquet-Cards being sent for, it should be perceived that there is, from Top to Bottom, the same Disposition of the Cards in both packs; let us likewise suppose that, some doubt arising about this Disposition of the Cards, it should be questioned whether it ought to be attributed to Chance, or to the Maker's Design: In this Case the Doctrine of Combinations decides the Question; since it may be proved by its Rules, that there are the odds of above 263130830000 Millions of Millions of Millions of Millions to One, that the Cards were designedly set in the Order in which they were found.

Nor has eliminating chance through small probabilities diminished in our own day. When Ronald Fisher charged Gregor Mendel's gardening assistant with data falsification because Mendel's data matched Mendel's theory too closely, Fisher was eliminating chance through small probabilities (see Fisher, 1965, p. 53). When Richard Dawkins takes creationists to task for misunderstanding Darwin's theory, it is for failing to appreciate that Darwin's theory renders life's emergence and development sufficiently probable to obviate the need for a supernatural designer. As a matter of general principle, however, Dawkins concedes that eliminating chance through small probabilities constitutes a valid mode of reasoning – it's just that in the case of life's emergence and development the probabilities aren't small enough (see Dawkins, 1987, pp. 139–144). Present-day examples of chance-elimination arguments based on small probabilities are easily multiplied (cf. the later sections in this chapter).

What underlies our unwillingness to attribute highly improbable events to chance? According to the French mathematician Emile Borel this unwillingness proceeds from a regulative principle governing small probabilities. Borel (1962, p. 1) referred to this principle as *the Single Law of Chance*, and formulated it as follows: "Phenomena with very small probabilities do not occur." Though Borel's probabilistic intuitions were excellent, and though his work on small probabilities has led the field since the 1930s, his statement of the Single Law of Chance is inadequate. Two objections can be raised: (1) Borel never adequately distinguished those highly improbable events properly attributed to chance from those properly attributed to something else; and (2) Borel never clarified what concrete numerical values correspond to small probabilities.

The first objection is evident: plenty of highly improbable events happen by chance all the time. The precise sequence of heads and tails in a long sequence of coin tosses, the precise configuration of darts from throwing darts at a dart board, and the precise seating arrangement of people at a cinema are all highly improbable events that, apart from any further information, can properly be explained by appeal to chance. It is only when the precise sequence of heads and tails has been recorded in advance, when the precise configuration of the darts locates all the darts at the center of the target, and when the precise seating arrangement at the cinema corresponds to seats people have been assigned on their ticket stubs that we begin to doubt whether these events occurred by chance. In other words, it is not just the sheer improbability of an event, but also the conformity of the event to a *pattern*, that leads us to look beyond chance to explain the event.

"Eliminating chance through small probabilities" must therefore be interpreted as an elliptical expression. Sheer improbability by itself is not enough to eliminate chance. Rather, to eliminate chance, we need also to know whether an event conforms to a pattern. Unfortunately, in formulating his Single Law of Chance, Borel never made this conformity of event to pattern explicit. This is not to say that Borel dismissed an extra-probabilistic factor like "conformity to pattern" in eliminating chance. But in never making the role of patterns explicit, or analyzing it, Borel's treatment of how small probabilities eliminate chance became unnecessarily restrictive. Essentially, to eliminate chance through small probabilities Borel was forced to designate a pattern prior to an event and then eliminate chance just in case the event conformed to that pattern. This method of eliminating chance is of course extremely common in statistics, where it is known as setting a *rejection region* prior to an experiment. In statistics, if the outcome of an experiment (= event) falls within the rejection region (= pattern), the chance hypothesis supposedly responsible for the outcome is rejected (i.e., chance is eliminated).

A little reflection, however, makes clear that a pattern need not be given prior to an event to warrant eliminating chance. Consider, for instance, Alice and Bob on the occasion of their fiftieth wedding anniversary. Their six children show up bearing gifts. Each gift is part of a matching set of china. There is no duplication of gifts, and together the gifts form a complete set of china. Suppose Alice and Bob were satisfied with their old set of china, and had no inkling prior to opening their gifts that they might expect a new set of china. Alice and Bob are therefore without a relevant pattern whither to refer their gifts prior to actually receiving them from their children. Nevertheless, Alice and Bob will not attribute the gifts to random acts of kindness (i.e., to chance). Rather, Alice and Bob will attribute the new set of china to the collusion of their children (i.e., to design). Granted, Alice and Bob have been given no pattern prior to receiving the gifts. Yet on receiving the gifts, Alice and Bob discern a pattern that – though discerned after the fact – cannot be reasonably explained apart from the collusion of their children.

In the presence of small probabilities, patterns given prior to events always eliminate chance. In the presence of small probabilities, patterns identified after events may or may not eliminate chance. Thus, Alice and Bob were able to eliminate chance after the fact. But suppose I flip a coin a thousand times and subsequently record the sequence of coin tosses on paper. The sequence I flipped (= event) conforms to the sequence recorded on paper (= pattern). Moreover, the sequence I flipped is vastly improbable (the probability is approximately 10^{-300}). Nevertheless, it's clear that the pattern to which these coin flips conform was artificially concocted and, as it stands, cannot legitimately warrant eliminating chance – the pattern was simply read off the event.

Patterns may therefore be divided into two types, those that in the presence of small probabilities warrant the elimination of chance and those that despite the presence of small probabilities do not warrant the elimination of chance. The first type of pattern will be called a *specification*, the second a *fabrication*.[1] Borel never drew this distinction, and as a result only saw how to eliminate chance in the simplest case where a pattern is given prior to an event. Borel's Single Law of Chance never came to terms with specification. In place of his Single Law of Chance, I therefore propose a new regulative principle governing small probabilities, one that makes explicit reference to specification. This replacement for Borel's Single Law of Chance I call the *Law of Small Probability* (LSP). According to this law, *specified events of small probability do not occur by chance*.

The other objection to Borel's Single Law of Chance is this. Besides failing to distinguish specifications from fabrications, Borel never clarified what concrete numerical values correspond to small probabilities. In any practical application, to use small probabilities to eliminate chance we need a probability bound ω according to which any probability p less than ω is small. The need for such a probability bound now raises an obvious question: How small is small enough? This question demands a concrete numerical answer, for without such concrete numbers, eliminating chance through small probabilities becomes subjective and vague.

This is not to say Borel never offered such concrete numbers. He did. In 1930 he proposed 10^{-1000} as a bound below which probabilities could be neglected universally (i.e., neglected across the entire universe). Later, in 1939, he proposed a less stringent universal probability bound of 10^{-50} (see Knobloch, 1990, p. 228). Unfortunately, Borel never convincingly justified these probability bounds. Take 10^{-50}, the probability bound on which Borel ultimately settled. Borel (1962, p. 28) justified this bound as follows:

> If we turn our attention, not to the terrestrial globe, but to the portion of the universe accessible to our astronomical and physical instruments, we are led to define the negligible probabilities on the cosmic scale. Some astronomical laws, such as Newton's law of universal gravitation and certain physical

[1] This distinction addresses the worry that it's always possible to find any pattern one likes in a data set so long as one looks hard enough. Although there may be no limit to the patterns one can invent and afterward impose on data, there are strict limits to the patterns with the right probabilistic and complexity–theoretic properties for eliminating chance (cf. Chapter 5). Distinguishing patterns according to their ability to underwrite particular forms of inference is not new. Nelson Goodman (1983), for instance, distinguishes the patterns that lead to successful inductive inferences from those that do not, referring to the former as projectable predicates (cf. "all emeralds are green") and to the latter as nonprojectable predicates (cf. "all emeralds are grue" where grue means green before the year 2000, blue thereafter).

laws relative to the propagation of light waves, are verified by innumerable observations of all the visible celestial bodies. The probability that a new observation would contradict all these concordant observations is extremely small. We may be led to set at 10^{-50} the value of negligible probabilities on the cosmic scale. When the probability of an event is below this limit, the opposite event may be expected to occur with certainty, whatever the number of occasions presenting themselves in the entire universe. The number of observable stars is of the order of magnitude of a billion, or 10^9, and the number of observations which the inhabitants of the earth could make of these stars, even if all were observing, is certainly less than 10^{20}. [An event] with a probability of 10^{-50} will therefore never occur, or at least never be observed.

There are three problems with Borel's case for 10^{-50} as a universal probability bound. First, Borel does not adequately distinguish the occurrence of an event from the observation of an event. There is a difference between an event never occurring and never being observed. Is it that (specified) events of small probability are occurring all the time, but that we're just not observing them? Or are they not occurring at all? Borel doesn't say.

Second, Borel never makes clear how the number of opportunities for an event to occur covaries with his universal probability bound. That there's a connection is clear. If, for instance, there are 10^{50} opportunities for an event of probability 10^{-50} to occur, then there is a better than even chance that the event will occur. With 10^{50} opportunities, an event of probability 10^{-50} is therefore likely to occur. But suppose Borel is right, and that the universe is such that no event has anywhere near 10^{50} opportunities to occur. Suppose no event has more than 10^{30} opportunities to occur. An event of probability 10^{-50} with 10^{30} opportunities to occur therefore has probability around 10^{-20} of occurring. True, this last probability strikes us as absurdly small. But if all we've done is substitute one small probability (i.e., 10^{-20}) for another (i.e., 10^{-50}), then we've hardly explained what constitutes a small probability. There is a regress here, and Borel does nothing to point the way out.

Third and last, in fixing his small probability bound, Borel neglects an inquirer's interests and context. In most contexts 10^{-50} is far too stringent. When, for instance, Ronald Fisher charged Mendel's gardening assistant with data falsification, what elicited this charge was a specified event whose probability was no more extreme than one in a hundred thousand (see Freedman, Pisani, and Purves, 1978, pp. 426–7). What counts as a small probability depends on an inquirer's interests and context for eliminating chance. Social scientists who set their alpha level at .05 or .01 (= small probability bound) are less stringent than criminal courts that establish guilt to a moral certainty and beyond reasonable doubt, and these in turn are less stringent than inflationary cosmologists as they account for the relative flatness of spacetime. Borel admits that what counts as a small probability on "the human scale" differs from what counts as a small probability on the "cosmic scale" (Borel, 1962, pp. 26–8). But he never clarifies how small probabilities covary with context.

In each case the difficulty is not that Borel went wrong, but that Borel did not go far enough. The Law of Small Probability corrects Borel's Single Law of Chance, and thereby elucidates the pattern of inference from which the title of this monograph takes its name – *the design inference*. When the Law of Small Probability eliminates chance, it is always a specific chance hypothesis that gets eliminated. By itself, a given application of the Law of Small Probability therefore falls under that branch of statistics known as hypothesis testing. In hypothesis testing, when a given chance hypothesis gets eliminated, it is typically because an alternate chance hypothesis has displaced it – essentially chance gets replaced by chance (cf. Hacking, 1965, p. 89). By contrast, a successful design inference sweeps the field clear of chance hypotheses. The design inference, in inferring design, eliminates chance entirely, whereas statistical hypothesis testing, in eliminating one chance hypothesis, opens the door to others.

To appreciate the difference between statistical hypothesis testing and the design inference, imagine a die thrown six million times. Statistical hypothesis testing considers two hypotheses: H_0, the null hypothesis, which asserts that the die is fair (i.e., each face has probability $1/6$); and H_1, the alternate hypothesis, which asserts that the die is in some way loaded or skewed. Suppose now that the die is thrown six million times and that each face appears *exactly* one million times. Even if the die is fair, something is fishy about getting *exactly* one million appearances of each face. Yet the standard statistical method for testing whether the die is fair, namely, a chi-square goodness of fit test, has no possibility of rejecting the null hypothesis (cf. Freedman et al., 1978, ch. 28). Indeed, statistical hypothesis testing can do no better than advise accepting the null hypothesis H_0.

This advice, however, is clearly absurd. As with Ronald Fisher's analysis of Gregor Mendel's pea-pod experiments, the fit between data and theory is too close to be explained by chance. If the die is fair and directed by chance, our best single guess – or what is known as the *mathematical expectation* – is that each face of the die will appear a million times. But this mathematical expectation differs sharply from our practical expectation in which we expect to see each face of the die appear *roughly* a million times, but not *exactly* a million times. The probability of an exact fit with mathematical expectation is around 10^{-20}, which in any practical application constitutes a small probability. Moreover, since in any practical application the mathematical expectation will constitute a specification, the Law of Small Probability advises rejecting the null hypothesis H_0, contrary to statistical hypothesis testing. Thus, whereas statistical hypothesis testing eliminates chance because divergence from mathematical expectation is too great, the design inference eliminates chance because the fit with mathematical expectation is too close.

We may therefore think of design and chance as competing modes of explanation for which design prevails once chance is exhausted. In eliminating chance, the design inference eliminates not just a single chance hypothesis, but all relevant chance hypotheses. How do we explain an event once a design inference has swept the field clear of relevant chance hypotheses? Although a design inference is often the occasion for inferring an intelligent agent (cf. the examples in the following sections), as a pattern of inference the design inference is not tied to any doctrine of intelligent agency. The design inference focuses on features of an event that bar it from being attributed to chance, not on the causal story underlying the event. To be sure, there is a connection between the design inference and intelligent agency (see Section 2.4). This connection, however, is not part of the logical structure of the design inference. Certain events are properly attributed to chance, certain events are not. The design inference marks the difference, yet without prejudging the underlying causal story.

If the design inference at best implicates an intelligent agent without necessarily delivering one, why use the word *design* at all, especially since the word so readily connotes intelligent agency? The reference to design reflects the logical structure of the design inference, depending as it does on a coincidence between patterns and events. Taken in its most fundamental sense, the word *design* denotes a *pattern* or *blueprint*. Often the reason an event conforms to a pattern is because an intelligent agent has acted deliberately to conform the event to the pattern. There is no logical necessity, however, for turning this connection between event and pattern into a metaphysical principle. We can determine whether an event conforms to a pattern without having to explain why the conformity exists. Thus, even though a design inference is frequently the first step toward identifying an intelligent agent, design as inferred from the design inference does not logically entail an intelligent agent. The design that emerges from the design inference must not be conflated with intelligent agency. Though they are frequently linked, the two are separate. Whether an event conforms to a pattern is a separate question from what caused an event to conform to a pattern.

The effect of a design inference is to limit our explanatory options, not to identify a cause. To identify a cause we need to investigate the particulars of the situation in which design is inferred. Simply put, we need more details. As a mode of explanation, design is not in the business of telling causal stories. Rather, design signifies the output of a certain pattern of inference, to wit, the design inference. Design therefore constitutes a logical rather than causal category. So construed design falls not within teleology, but within the logical foundations of probability theory. The design inference constitutes the most exciting application of the Law of Small Probability. All the remaining sections of this chapter illustrate design inferences.

1.2 THE MAN WITH THE GOLDEN ARM

Even if we can't ascertain the precise causal story underlying an event, we often have probabilistic information that enables us to rule out ways of explaining the event. This ruling out of explanatory options is what the design inference is all about. The design inference does not by itself deliver an intelligent agent. But as a logical apparatus for sifting our explanatory options, the design inference rules out explanations incompatible with intelligent agency (such as chance). The design inference appears widely, and is memorably illustrated in the following example (*New York Times*, 23 July 1985, p. B1):

> TRENTON, July 22 – The New Jersey Supreme Court today caught up with the "man with the golden arm," Nicholas Caputo, the Essex County Clerk and a Democrat who has conducted drawings for decades that have given Democrats the top ballot line in the county 40 out of 41 times.

Mary V. Mochary, the Republican Senate candidate, and county Republican officials filed a suit after Mr. Caputo pulled the Democrat's name again last year.

The election is over – Mrs. Mochary lost – and the point is moot. But the court noted that the chances of picking the same name 40 out of 41 times were less than 1 in 50 billion. It said that "confronted with these odds, few persons of reason will accept the explanation of blind chance."

And, while the court said it was not accusing Mr. Caputo of anything, it said it believed that election officials have a duty to strengthen public confidence in the election process after such a string of "coincidences."

The court suggested – but did not order – changes in the way Mr. Caputo conducts the drawings to stem "further loss of public confidence in the integrity of the electoral process."

. . .

Justice Robert L. Clifford, while concurring with the 6-to-0 ruling, said the guidelines should have been ordered instead of suggested.

Nicholas Caputo was brought before the New Jersey Supreme Court because the Republican party filed suit against him, claiming Caputo had consistently rigged the ballot lines in the New Jersey county where he was county clerk. It is common knowledge that first position on a ballot increases one's chances of winning an election (other things being equal, voters are more likely to vote for the first person on a ballot than the rest). Since in every instance but one Caputo positioned the Democrats first on the ballot line, the Republicans argued that in selecting the order of ballots Caputo had intentionally favored his own Democratic party. In short, the Republicans claimed Caputo cheated.

The question, then, before the New Jersey Supreme Court was, Did Caputo actually rig the order of ballots, or was it without malice and forethought that Caputo assigned the Democrats first place forty out of forty-one times? Since Caputo denied wrongdoing, and since he conducted the drawing of ballots so that witnesses were unable to observe how he actually did draw the ballots (this was brought out in a portion of the article omitted in the preceding quote), determining whether Caputo did in fact rig the order of ballots becomes a matter of evaluating the circumstantial evidence connected with this case. How, then, is this evidence to be evaluated?

In trying to explain the remarkable coincidence of Nicholas Caputo selecting the Democrats forty out of forty-one times to head the ballot line, the court faced three explanatory options:

Regularity: Unknown to Caputo, he was not employing a reliable random process to determine ballot order. Caputo was like someone who thinks a fair coin is being flipped when in fact it's a double-headed coin. Just as flipping a double-headed coin is going to yield a long string of heads, so Caputo, using his faulty method for ballot selection, generated a long string of Democrats coming out on top. An unknown regularity controlled Caputo's ballot line selections.

Chance: In selecting the order of political parties on the state ballot, Caputo employed a reliable random process that did not favor one political party over another. The fact that the Democrats came out on top forty out of forty-one times was simply a fluke. It occurred by chance.

Agency: Caputo, acting as a fully conscious intelligent agent and intending to aid his own political party, purposely rigged the ballot line selections to keep the Democrats coming out on top. In short, Caputo cheated.

The first option – that Caputo chose poorly his procedure for selecting ballot lines, so that instead of genuinely randomizing the ballot order, it just kept putting the Democrats on top – was not taken seriously by the court. The court could dismiss this option outright because Caputo claimed to be using an urn model to select ballot lines. Thus, in a portion of the *New York Times* article not quoted, Caputo claimed to have placed capsules designating the various political parties running in New Jersey into a container, and then swished them around. Since urn models are among the most reliable randomization techniques available, there was no reason for the court to suspect that Caputo's randomization procedure was at fault. The key question, therefore, was whether Caputo actually put this procedure into practice when he made the ballot line selections, or whether he purposely circumvented this procedure to keep the Democrats coming out on top. And since Caputo's actual drawing of the capsules was obscured to witnesses, it was this question the court had to answer.

With the regularity explanation at least for the moment bracketed, the court next decided to dispense with the chance explanation. Having noted that the chance of picking the same political party 40 out of 41 times was less than 1 in 50 billion, the court concluded that "confronted with these odds, few persons of reason will accept the explanation of blind chance." Now this certainly seems right. Nevertheless, a bit more needs to be said. As we saw in Section 1.1, exceeding improbability is by itself not enough to preclude an event from happening by chance. Whenever I am dealt a bridge hand, I participate in an exceedingly improbable event. Whenever I play darts, the precise position where the darts land represents an exceedingly improbable configuration. In fact, just about anything that happens is exceedingly improbable once we factor in all the other ways what actually happened might have happened. The problem, then, does not reside simply in an event being improbable.

All the same, in the absence of a causal story detailing what happened, improbability remains a crucial ingredient in eliminating chance. For suppose that Caputo actually was cheating right from the beginning of his career as Essex County clerk. Suppose further that the one exception in Caputo's career as "the man with the golden arm" – that is, the one case where Caputo placed the Democrats second on the ballot line – did not occur till after his third time selecting ballot lines. Thus, for the first three ballot line selections of Caputo's career the Democrats all came out on top, and they came out on top precisely because Caputo intended it that way. Simply on the basis of three ballot line selections, and without direct evidence of Caputo's cheating, an outside observer would be in no position to decide whether Caputo was cheating or selecting the ballots honestly.

With only three ballot line selections, the probabilities are too large to reliably eliminate chance. The probability of randomly selecting the Democrats to come out on top given that their only competition is the Republicans is in this case 1 in 8 (here p equals 0.125; compare this with the p-value computed by the court, which equals 0.00000000002). Because three Democrats in a row could easily happen by chance, we would be acting in bad faith if we did not give Caputo the benefit of the doubt in the face of such large probabilities. Small probabilities are therefore a necessary condition for eliminating chance, even though they are not a sufficient condition.

What, then, besides small probabilities do we need for evidence that Caputo cheated? As we saw in Section 1.1, the event in question needs to conform to a pattern. Not just any pattern will do, however. Some patterns successfully eliminate chance while others do not.

Consider the case of an archer. Suppose an archer stands fifty meters from a large wall with bow and arrow in hand. The wall, let us say, is sufficiently large that the archer cannot help but hit it. Now suppose every time the archer shoots an arrow at the wall, she paints a target around the arrow, so that the arrow is positioned squarely in the bull's-eye. What can be concluded from this scenario? Absolutely nothing about the archer's ability as an archer. The fact that the archer is in each instance squarely hitting the bull's-eye is utterly bogus. Yes, she is matching a pattern; but it is a pattern she fixes only after the arrow has been shot and its position located. The pattern is thus purely ad hoc.

But suppose instead that the archer paints a fixed target on the wall and then shoots at it. Suppose she shoots 100 arrows, and each time hits a perfect bull's-eye. What can be concluded from this second scenario? In the words of the New Jersey Supreme Court, "confronted with these odds, few persons of reason will accept the explanation of blind chance." Indeed, confronted with this second scenario we infer that here is a world-class archer.

The difference between the first and the second scenario is that the pattern in the first is purely ad hoc, whereas the pattern in the second is not. Thus, only in the second scenario are we warranted eliminating chance. Let me emphasize that for now I am only spotlighting a distinction without explicating it. I shall in due course explicate the distinction between "good" and "bad" patterns – those that respectively do and don't permit us to eliminate chance (see Chapter 5). But for now I am simply trying to make the distinction between good and bad patterns appear plausible. In Section 1.1 we called the good patterns *specifications* and the bad patterns *fabrications*. Specifications are the non-ad hoc patterns that can legitimately be used to eliminate chance and warrant a design inference. Fabrications are the ad hoc patterns that cannot legitimately be used to eliminate chance.

Thus, when the archer first paints a fixed target and thereafter shoots at it, she *specifies* hitting a bull's-eye. When in fact she repeatedly hits the bull's-eye, we are warranted attributing her success not to beginner's luck, but to her skill as an archer. On the other hand, when the archer paints a target around the arrow only after each shot, squarely positioning each arrow in the bull's-eye, she *fabricates* hitting the bull's-eye. Thus, even though she repeatedly hits the bull's-eye, we are not warranted attributing her "success" in hitting the bull's-eye to anything other than luck. In the latter scenario, her skill as an archer thus remains an open question.[2]

How do these considerations apply to Nicholas Caputo? By selecting the Democrats to head the ballot forty out of forty-one times, Caputo appears to have participated in an event of probability less than 1 in 50 billion ($p = 0.00000000002$). Yet as we have noted, events of exceedingly small probability happen all the time. Hence by itself Caputo's participation in an event of probability less than 1 in 50 billion is no cause for alarm. The crucial question is whether this event is also specified – does this event follow a non-ad hoc pattern so that we can legitimately eliminate chance?

Now there is a very simple way to avoid ad hoc patterns and generate specifications, and that is by designating an event prior to its occurrence – C. S. Peirce (1883 [1955], pp. 207–10) referred to this type of specification as a *predesignation*. In the archer example, by painting the bull's-eye before taking aim, the archer specifies in advance where the arrows are to land. Because the pattern is set prior to the event, the objection of ad-hocness or fabrication is effectively blocked.

In the Caputo case, however, the pattern is discovered after the event: only after we witness an extended series of ballot line selections do we notice a suspicious pattern. Though discovered after the fact, this pattern is not a fabrication. Patterns given prior to an event, or Peirce's predesignations, constitute but a proper subset of the patterns that legitimately eliminate chance. The important thing about a pattern is not when it was identified, but whether in a certain well-defined sense it is independent of an event. We refer to this relation of independence as *detachability*, and say that a pattern is *detachable* just in case it satisfies this relation.

[2] The archer example introduces a tripartite distinction that will be implicit throughout our study of chance elimination arguments: a reference class of possible events (e.g., the arrow hitting the wall at some unspecified place); a pattern that restricts the reference class of possible events (e.g., a target on the wall); and the precise event that has occurred (e.g., the arrow hitting the wall at some precise location). In a chance elimination argument, the reference class, the pattern, and the event are always inseparably linked, with the pattern mediating between the event and the reference class, helping to decide whether the event really is due to chance. Throughout this monograph we shall refer to patterns and events as such, but refer to reference classes by way of the chance hypotheses that characterize them (cf. Section 5.2).

Detachability distinguishes specifications from fabrications. Although a precise account of detachability will have to wait until Chapter 5, the basic intuition underlying detachability is this: Given an event, would we be able to formulate a pattern describing it if we had no knowledge which event occurred? Here is the idea. An event has occurred. A pattern describing the event is given. The event is one from a range of possible events. If all we knew was the range of possible events without any specifics about which event actually occurred, could we still formulate the pattern describing the event? If so, the pattern is detachable from the event.

To illustrate detachability in the Caputo case, consider two possible courses Nicholas Caputo's career as Essex County clerk might have taken (for simplicity assume no third-party candidates were ever involved, so that all elections were between Democrats and Republicans). In the one case – and for the sake of argument let us suppose this is what actually happened – Caputo chose the Democrats over the Republicans forty out of forty-one times in the following order:

(A) DDDDDDDDDDDDDDDDDDDDDDRDDDDDDDDDDDDDDDDDD.

Thus, the initial twenty-two times Caputo chose the Democrats to head the ballot line, then the twenty-third time he chose the Republicans, after which for the remaining times he chose the Democrats.

In the second possible course Caputo's career as county clerk might have taken, suppose Caputo once again had forty-one occasions on which to select the order of ballots, but this time that he chose both Democrats and Republicans to head the ballot pretty evenly, let us say in the following order:

(B) DRRDRDRRDDDRDRDDRDRRDRRDRRRDRRRDRDDDRDRDD.

In this instance the Democrats came out on top only twenty times, and the Republicans twenty-one times.

Sequences (A) and (B) are both patterns and describe possible ways Caputo might have selected ballot orders in his years as Essex County clerk. (A) and (B) are therefore patterns describing possible events. Now the question detachability asks is whether (A) and (B) could have been formulated without our knowing which event occurred. For (A) the answer is yes, but for (B) the answer is no. (A) is therefore detachable whereas (B) is not.

How is this distinction justified? To formulate (B) I just one moment ago flipped a coin forty-one times, recording "D" for Democrat whenever I observed heads and "R" for Republican whenever I observed tails. On the other hand, to formulate (A) I simply recorded "D" forty times and then interspersed a single "R." Now consider a human subject S confronted with sequences (A) and (B). S comes to these sequences with considerable background knowledge which, we may suppose, includes the following:

(1) Nicholas Caputo is a Democrat.
(2) Nicholas Caputo would like to see the Democrats appear first on the ballot since having the first place on the ballot line significantly boosts one's chances of winning an election.
(3) Nicholas Caputo, as election commissioner of Essex County, has full control over who appears first on the ballots in Essex County.
(4) Election commissioners in the past have been guilty of all manner of fraud, including unfair assignments of ballot lines.
(5) If Caputo were assigning ballot lines fairly, then both Democrats and Republicans should receive priority roughly the same number of times.

Given this background knowledge S is in a position to formulate various "cheating patterns" by which Caputo might attempt to give the Democrats first place on the ballot. The most blatant cheat is of course to assign the Democrats first place all the time. Next most blatant is to assign the Republicans first place just once (as in (A) – there are 41 ways to assign the Republicans first place just once). Slightly less blatant – though still blatant – is to assign the Republicans first place exactly two times (there are 820 ways to assign the Republicans first place exactly two times). This line of reasoning can be extended by throwing the Republicans a few additional sops. The point is, given S's background knowledge, S is easily able (possibly with the aid of a personal computer) to formulate ways Caputo could cheat, one of which would surely include (A).

Contrast this now with (B). Since (B) was generated by a sequence of coin tosses, (B) represents one of two trillion or so possible ways Caputo might legitimately have chosen ballot orders. True, in this respect probabilities do not distinguish (A) from (B) since all such sequences of Ds and Rs of length 41 have the same small probability of occurring by chance, namely 1 in 2^{41}, or approximately 1 in two trillion. But S is a finite agent whose background knowledge enables S to formulate only a tiny fraction of all the possible sequences of Ds and Rs of length 41. Unlike (A), (B) is not among them. Confronted with (B), S will scrutinize it, try to discover a pattern that isn't ad hoc, and thus seek to uncover evidence that (B) resulted from something other than chance. But given S's background knowledge, nothing about (B) suggests an explanation other than chance. Indeed, since the relative frequency of Democrats to Republicans actually favors Republicans (twenty-one Rs versus twenty Ds), the Nicholas Caputo responsible for (B) is hardly "the man with the golden arm." Thus, while (A) is detachable, (B) is not.

But can one be absolutely certain (B) is not detachable? No, one cannot. There is a fundamental asymmetry between detachability and its negation, call it *nondetachability*. In practice one can decisively demonstrate that a pattern is detachable from an event, but not that a pattern is incapable of being detached from an event. A failure to establish detachability always leaves open the possibility that detachability might still be demonstrated at some later date.

To illustrate this point, suppose I walk down a dirt road and find some stones lying about. The configuration of stones says nothing to me. Given my background knowledge I can discover no pattern in this configuration that I could have formulated on my own without actually seeing the stones lying about as they do. I cannot detach the pattern of stones from the configuration they assume. I therefore have no reason to attribute the configuration to anything other than chance. But suppose next an astronomer travels this same road and looks at the same stones only to find that the configuration precisely matches some highly complex constellation. Given the astronomer's background knowledge, this pattern now becomes detachable. The astronomer will therefore have grounds for thinking that the stones were intentionally arranged to match the constellation.

Detachability must always be relativized to a subject and a subject's background knowledge. Whether one can detach a pattern from an event depends on one's background knowledge coming to the event. Often one's background knowledge is insufficient to detach a pattern from an event. Consider, for instance, the case of cryptographers trying to break a cryptosystem. Until they break the cryptosystem, the strings of characters they record from listening to their enemy's communications will seem random, and for all the cryptographers know might just be gibberish. Only after the cryptographers have broken the cryptosystem and discovered the key for decrypting their enemy's communications will they discern the detachable pattern present in the communications they have been monitoring (cf. Section 1.6).

Is it, then, strictly because our background knowledge and abilities are limited that some patterns fail to be detachable? Would, for instance, an infinitely powerful computational device be capable of detaching any pattern whatsoever? Regardless whether some super-being possesses an unlimited capacity to detach patterns, as a practical matter we humans find ourselves with plenty of patterns we cannot detach. Whether all patterns are detachable in some grand metaphysical sense, therefore, has no bearing on the practical problem whether a certain pattern is detachable given certain limited background knowledge. Finite rational agents like ourselves can formulate only a very few detachable patterns. For instance, of all the possible ways we might flip a coin a thousand times, we can make explicit only a minuscule proportion. It follows that a human subject will be unable to specify any but a very tiny fraction of these possible coin flips. In general, the patterns we can know to be detachable are quite limited.[3]

Let us now wrap up the Caputo example. Confronted with Nicholas Caputo assigning the Democrats the top ballot line forty out of forty-one times, the New Jersey Supreme Court first rejected the regularity explanation, and then rejected the chance explanation ("confronted with these odds, few persons of reason will accept the explanation of blind chance"). Left with no other option, the court therefore accepted the agency explanation, inferred Caputo was cheating, and threw him in prison.

Well, not quite. The court did refuse to attribute Caputo's golden arm to either regularity or chance. Yet when it came to giving a positive explanation of Caputo's golden arm, the court waffled. To be sure, the court knew something was amiss. For the Democrats to get the top ballot line in Caputo's county forty out of forty-one times, especially

[3] This conclusion is consistent with algorithmic information theory, which regards a sequence of numbers as nonrandom to the degree that it is compressible. Since compressibility within algorithmic information theory constitutes but a special case of detachability, and since most sequences are incompressible, the detachable sequences are indeed quite limited. See Kolmogorov (1965), Chaitin (1966), and van Lambalgen (1989). See also Section 1.7.

with Caputo solely responsible for ballot line selections, something had to be fishy. Nevertheless, the New Jersey Supreme Court was unwilling explicitly to charge Caputo with corruption. Of the six judges, Justice Robert L. Clifford was the most suspicious of Caputo, wanting to *order* Caputo to institute new guidelines for selecting ballot lines. The actual ruling, however, simply *suggested* that Caputo institute new guidelines in the interest of "public confidence in the integrity of the electoral process." The court therefore stopped short of charging Caputo with dishonesty.

Did Caputo cheat? Certainly this is the best explanation of Caputo's golden arm. Nonetheless, the court stopped short of convicting Caputo. Why? The court had no clear mandate for dealing with highly improbable ballot line selections. Such mandates exist in other legal settings, as with discrimination laws that prevent employers from attributing to the luck of the draw their failure to hire sufficiently many women or minorities. But in the absence of such a mandate the court needed an exact causal story of how Caputo cheated if the suit against him was to succeed. And since Caputo managed to obscure how he selected the ballot lines, no such causal story was forthcoming. The court therefore went as far as it could.

Implicit throughout the court's deliberations was the design inference. The court wanted to determine whether Caputo cheated. Lacking a causal story of how Caputo selected the ballot lines, the court was left with circumstantial evidence. Given this evidence, the court immediately ruled out regularity. What's more, from the specified improbability of selecting the Democrats forty out of forty-one times, the court also ruled out chance.

These two moves – ruling out regularity, and then ruling out chance – constitute the design inference. The conception of design that emerges from the design inference is therefore eliminative, asserting of an event what it is not, not what it is. To attribute an event to design is to say that regularity and chance have been ruled out. Referring Caputo's ballot line selections to design is therefore not identical with referring it to agency. To be sure, design renders agency plausible. But as the negation of regularity and chance, design is a mode of explanation logically preliminary to agency. Certainly agency (in this case cheating) best explains Caputo's ballot line selections. But no one was privy to Caputo's ballot line selections. In the absence of an exact causal story, the New Jersey Supreme Court therefore went as far as it could in the Caputo case.[4]

1.3 INTELLECTUAL PROPERTY PROTECTION

If the courts are at times less than decisive in connecting design with agency, the same cannot be said for the many professions whose livelihood depends on drawing design inferences and using them to attribute agency. Money, property, and even human lives depend on the events we attribute to agency. Often the design inference is the only way to distinguish agency from other causes, so that for practical purposes design (i.e., the elimination of regularity and chance) and agency (i.e., the intentional activity of an intelligent cause or agent) become identified, if not conflated. Thus, we find entire industries dedicated to drawing design inferences, and therewith immediately attributing agency. These industries include patent offices, copyright offices, insurance companies, actuarial firms, statistical consultants, cryptographers, forensic scientists, and detectives to name but a few. For the remainder of this chapter I intend simply to sketch how these "design industries" infer design and thereby attribute agency. In each case the key that turns the lock is a specified event of small probability.

Consider first intellectual property protection. There are two primary industries that protect intellectual property – patent and copyright offices. Patents protect inventions, copyrights protect text. People avail themselves of patent and copyright laws to assert the priority of their work, and thus to keep copycats from obtaining a share in the market. The laws are such that if person A creates some artifact X and files X with the relevant patent or copyright office at time t_1, and person B claims to have created the same artifact X at any time t_2 subsequent to time t_1, person B is liable to penalties. Is this fair? What if B created X independently of A? Chance can never exclude this possibility. But since the probability of B creating X independently

[4] Legal scholars continue to debate the proper application of probabilistic reasoning to legal problems. Larry Tribe (1971), for instance, views the application of Bayes's theorem within the context of a trial as fundamentally unsound. Michael Finkelstein takes the opposite view (see Finkelstein, 1978, p. 288 ff.). Still, there appears no getting rid of the design inference in the law. Cases of bid-rigging (Finkelstein and Levin, 1990, p. 64), price-fixing (Finkelstein and Levenbach, 1986, pp. 79–106), and collusion often cannot be detected save by means of a design inference.

of A is in most cases minuscule, the presumption is that B copied X from A rather than that B came up with X independently of A.

Sometimes manufacturers assist patent and copyright offices by introducing "traps" into their artifacts so that anyone who copies the artifacts gets caught red-handed. Copycats, after all, typically introduce variations into the things they copy so that the match between original and copy is not exact. Thus, after A has produced X, the copycat B, instead of reproducing X exactly, will produce a variation on X, call it X′, which B hopes will be sufficiently different from X to circumvent the protection offered by patent and copyright offices. Traps block this move.

Consider, for instance, a trap formerly employed by Encyclopedia Britannica to monitor the copying of its encyclopedia by rival encyclopedia companies: because rival encyclopedia companies were not directly copying the articles in *Encyclopedia Britannica*, but rather seemed to be rewriting or paraphrasing its articles, Encyclopedia Britannica decided to include among its articles bogus biographies, i.e., biographies of people who were made up by Encyclopedia Britannica and thus never existed except in its pages. Thus, whenever the biography of some "famous" artist who had never existed reappeared in a rival encyclopedia, Encyclopedia Britannica could with complete confidence assert that the rival encyclopedia company had plagiarized. Bogus biographies thus served to trap plagiarizing encyclopedia companies.

A variation on this theme occurs when two parties, say A and B, have the power to produce exactly the same artifact X, but where producing X requires so much effort that it is easier to copy X once X has already been produced than to produce X from scratch. For instance, before the advent of computers logarithmic tables had to be computed by hand. Although there is nothing esoteric about calculating logarithms, the process is tedious and time-consuming if done by hand. Once the calculation has been accurately performed, however, there is no need to repeat it.

The problem, then, confronting the manufacturers of logarithmic tables was that after expending so much effort to compute logarithms, if they were to publish their results without safeguards, nothing would prevent a plagiarist from copying the logarithms directly, and then simply claiming that he or she had calculated the logarithms independently. To solve this problem, manufacturers of logarithmic tables introduced occasional – but deliberate – errors into their tables, errors which they carefully noted to themselves. Thus, in a table of logarithms that was accurate to eight decimal places, errors in the seventh and eighth decimal places would occasionally be introduced.

These errors then served to trap plagiarists, for even though plagiarists could always claim to have computed the logarithms correctly by mechanically following a certain algorithm, they could not reasonably claim to have committed the same errors. As Aristotle remarked in his *Nichomachean Ethics* (McKeon, 1941, p. 1106), "it is possible to fail in many ways, . . . while to succeed is possible only in one way." Thus, when two manufacturers of logarithmic tables record identical logarithms that are correct, both receive the benefit of the doubt that they have actually done the work of computing logarithms. But when both record the same errors, it is perfectly legitimate to conclude that whoever published last plagiarized.[5]

1.4 FORENSIC SCIENCE AND DETECTION

Forensic scientists, detectives, lawyers, and insurance fraud investigators cannot do without the design inference. Something as common as a forensic scientist placing someone at the scene of a crime by matching fingerprints requires a design inference. Indeed, there is no logical or genetic impossibility preventing two individuals from sharing the same fingerprints. Rather, our best understanding of fingerprints and the way they are distributed in the human population is that they are, with very high probability, unique to individuals. And so, whenever the fingerprints of an individual match those found at the scene of a crime, we conclude that the individual was indeed at the scene of the crime.

The forensic scientist's stock of design inferences is continually increasing. Consider the following headline: "DNA Tests Becoming Elementary in Solving Crimes." The lead article went on to describe

[5] In the same spirit there's the joke about two college students who sat next to each other while taking a final exam. In the opinion of the professor teaching the course, the one student was the brightest in the class, the other the worst. Yet when the professor got back their exams, she found both students had gotten every question but the last perfectly correct. When it came to the last question, however, the brightest student wrote, "I don't know the answer to this question," while the worst student wrote, "I don't know the answer to this question either." If there was any doubt about who was cheating, the incriminating use of the word "either" removed it.

the type of reasoning employed by forensic scientists in DNA testing. As the following excerpt makes clear, all the key features of the design inference described in Sections 1.1 and 1.2 are present in DNA testing (*The Times – Princeton-Metro*, N.J., 23 May 1994, p. A1):

> TRENTON – A state police DNA testing program is expected to be ready in the fall, and prosecutors and police are eagerly looking forward to taking full advantage of a technology that has dramatically boosted the success rate of rape prosecutions across the country.
>
> Mercer County Prosecutor Maryann Bielamowicz called the effect of DNA testing on rape cases "definitely a revolution. It's the most exciting development in my career in our ability to prosecute."
>
> She remembered a recent case of a young man arrested for a series of three sexual assaults. The suspect had little prior criminal history, but the crimes were brutal knifepoint attacks in which the assailant broke in through a window, then tied up and terrorized his victims.
>
> "Based on a DNA test in one of those assaults he pleaded guilty to all three. He got 60 years. He'll have to serve 27 1/2 before parole. That's pretty good evidence," she said.
>
> All three women identified the young man. But what really intimidated the suspect into agreeing to such a rotten deal were the enormous odds – one in several million – that someone other than he left semen containing the particular genetic markers found in the DNA test. Similar numbers are intimidating many others into foregoing trials, said the prosecutor.[6]

Not just forensic science, but the whole field of detection is inconceivable without the design inference. Indeed, the mystery genre would be dead without it.[7] When in the movie *Double Indemnity* Edward G. Robinson ("the insurance claims man") puts it together that Barbara Stanwyck's husband did not die an accidental death by falling off a train, but instead was murdered by Stanwyck to

[6] It's worth mentioning that at the time of this writing, the accuracy and usefulness of DNA testing is still a matter for debate. As a *New York Times* (23 August 1994, p. A10) article concerned with the currently ongoing O. J. Simpson case remarks, "there is wide disagreement among scientific experts about the accuracy and usefulness of DNA testing and they emphasize that only those tests performed under the best of circumstances are valuable." My interest, however, in this matter is not with the ultimate fate of DNA testing, but with the logic that underlies it, a logic that hinges on the design inference.

[7] Cf. David Lehman's (1989, p. 20) notion of "retrospective prophecy" as applied to the detective-fiction genre: "If mind-reading, backward-reasoning investigators of crimes – sleuths like Dupin or Sherlock Holmes – resemble prophets, it's in the visionary rather than the vatic sense. It's not that they see into the future; on the contrary, they're not even looking that way. But reflecting on the clues left behind by the past, they see patterns where the rest of us see only random signs. They reveal and make intelligible what otherwise would be dark." The design inference is the key that unlocks the patterns that "the rest of us see only [as] random signs."

collect on a life insurance policy, the design inference is decisive. Why hadn't Stanwyck's husband made use of his life insurance policy earlier to pay off on a previously sustained injury, for the policy did have such a provision? Why should he die just two weeks after taking out the policy? Why did he happen to die on a train, thereby requiring the insurance company to pay double the usual indemnity (hence the title of the movie)? How could he have broken his neck falling off a train when at the time of the fall, the train could not have been moving faster than 15 m.p.h.? And who would seriously consider committing suicide by jumping off a train moving only 15 m.p.h.? Too many pieces coalescing too neatly made the explanations of accidental death and suicide insupportable. Thus, at one point Edward G. Robinson exclaims, "The pieces all fit together like a watch!" Suffice it to say, in the movie Barbara Stanwyck and her accomplice/lover Fred MacMurray did indeed kill Stanwyck's husband.

Whenever there is a mystery, it is the design inference that elicits the crucial insight needed to solve the mystery. The dawning recognition that a trusted companion has all along been deceiving you (cf. *Notorious*); the suspicion that someone is alive after all, even though the most obvious indicators point to the person having died (cf. *The Third Man*); and the realization that a string of seemingly accidental deaths were carefully planned (cf. *Coma*) all derive from design inferences. At the heart of these inferences is a convergence of small probabilities and specifications, a convergence that cannot properly be explained by appealing to chance.

1.5 DATA FALSIFICATION IN SCIENCE

R. A. Fisher uncovered a classic case of data falsification by analyzing Gregor Mendel's data on peas. Fisher inferred that "Mendel's data were massaged," as one statistics text puts it, because Mendel's data matched Mendel's theory too closely.[8] Interestingly, the coincidence that elicited this charge of data falsification was a specified event whose probability was no more extreme than one in a hundred

[8] I'm basing my remarks about Mendel's data on Freedman et al. (1978, pp. 426–7) and Fisher (1965, p. 53). For a more recent reevaluation of Mendel's data, which still concludes that "the segregations are in general closer to Mendel's expectations than chance would dictate," see Edwards (1986).

thousand (a probability that is huge compared with the 1 in 50 billion probability of the Caputo example in Section 1.2). Fisher concluded his analysis of Mendel's experiment by charging Mendel's gardening assistant with deception.

In a more recent example of data falsification, an experimental psychologist intent on increasing the number of publications in his curriculum vitae decided to lift a two-by-two table of summary statistics from one of his articles and insert it – unchanged – into another article.[9] Data falsification was clearly implicated because of the vast improbability that data from two separate experiments should produce the same summary table of statistics. When forced to face a review board, the psychologist resigned his academic position rather than try to explain how this coincidence could have occurred without any fault on his part. The incriminating two-by-two table that appeared in both articles consisted of four blocks each containing a three-digit number. The odds would therefore have been roughly one in a trillion ($= 10^{12}$) that this same table might by chance have appeared twice in his research.

Why did the experimental psychologist resign rather than defend a 1 in 10^{12} improbability? Why not simply attribute the coincidence to chance? There were two reasons. First, at the review board the psychologist would have had to produce the experimental protocols for the two experiments that supposedly gave rise to the identical two-by-two tables. If he was guilty of data falsification, these protocols would have incriminated him. Second, even if the protocols were lost, the sheer improbability of producing so unlikely a match between the two papers would have been enough to impugn the researcher's honesty. Once a specification is in place (here the first article containing the two-by-two table specifies the other – cf. our discussion of plagiarism in Section 1.3) and the probabilities become too small, the burden of proof, at least within the scientific community, shifts to the experimenter suspected of data falsification.

Last, consider the debunking of parapsychology. Parapsychological experiments all follow a common pattern: producing a specified

[9] This example was aired on PBS back in the mid-eighties in a documentary concerned with dishonesty and fraud in the sciences. The name of the documentary as well as that of the experimental psychologist have long since escaped me. The main point of the case, however, has remained with me, and continues to provide a striking example of the design inference in action.

event of small probability, and then explaining it in terms of a theoretical construct called psi (i.e., the factor or faculty supposedly responsible for such events). For instance, shuffle some cards and then have a human subject guess their order. If the human subject guesses correctly, the improbability of this coincidence (= specified event of small probability) is regarded as evidence for psi. To attribute such coincidences to psi the parapsychologist must first draw a successful design inference (i.e., eliminate regularity and chance). The debunker's task, then, is to block the parapsychologist's design inference, that is, to show that the design inference was drawn incorrectly, and that chance or regularity can in fact account for the coincidence in question. In showing that parapsychologists failed to draw their design inferences correctly, debunkers typically look for outright fraud or sloven experimental method (see Reinsel, 1990, p. 194).

1.6 CRYPTOGRAPHY (AND SETI)

Cryptography is the study of secure communications. Given Alice and Bob, and their mutual enemy Carl, Alice wants to send messages to Bob without worrying about Carl reading Bob's mail. Before any messages are sent, Alice and Bob will therefore agree on a method that enables Bob to interpret Alice's transmissions, but one that makes it difficult for anyone ignorant of the method to know what Alice is communicating. In the parlance of the field, Alice *encrypts* messages and relays them to Bob; and Bob, in interpreting Alice's messages, *decrypts* them. Any such method for encrypting and decrypting messages is known as a *cryptosystem*.

It is vital that when Alice and Bob agree on a cryptosystem, they not reveal the method of decryption to Carl. Carl's overriding interest is to crack the cryptosystem. Presumably Carl will intercept some of the messages from Alice to Bob. In fact, not wanting to underestimate Carl, Alice and Bob will assume that Carl is able to intercept all transmissions from Alice to Bob. Alice, Bob, and Carl therefore have three respective tasks: Alice's task is to encrypt plaintext into ciphertext and get the ciphertext into Bob's hands. Bob's task is to decrypt ciphertext received from Alice back into plaintext. Carl's task is to intercept as many ciphertext messages from Alice to Bob as possible, steal as many translations into plaintext of these ciphertext messages as possible, and with this information try to break the cryptosystem that Alice and Bob are employing (i.e., figure out the method of encryption and decryption). Of course, if Carl can steal the actual method of encryption and decryption, so much the better for Carl – but then the game is up.[10]

Alice and Bob want their cryptosystem to be as secure as possible. The security of a cryptosystem, however, can never be perfect since Carl might by pure luck happen to guess the cryptosystem implemented by Alice and Bob. The security of a cryptosystem is therefore always a matter of probability. Hence for any cryptosystem it is necessary to ask, What is the probability that Carl will figure out the cryptosystem and thereafter be able to read Bob's mail? This probability will depend on several factors. We shall always give Carl the benefit of the doubt that Carl has full access to ciphertext transmissions. Hence the probability will depend on the number of ciphertext transmissions that have passed from Alice to Bob. It will also depend on any plaintext translations that Carl has happened to obtain of ciphertext transmissions. So too it will depend on the computational resources that Carl has available for running through possible cryptosystems and checking whether one of these possibilities is the one actually being employed (the computational power of Carl is the most important theme in current cryptographic research). In each case we give Carl the benefit of the doubt, never underestimating Carl's abilities, always assuming that short of a lucky guess Carl will adopt the most expedient course for breaking the cryptosystem.

When all is said and done, the security of a cryptosystem is encapsulated in a positive probability p – the probability that Carl at his most resourceful will break the cryptosystem. Once this measure of security is in hand, several questions immediately arise: How do Alice and Bob determine whether their cryptosystem has been compromised? Once they are confident their cryptosystem has been compromised, do they attribute its compromise to the ingenuity of Carl as a code breaker, or to the cunning of Carl as an infiltrator into Alice's and Bob's communication centers? As for Carl, when does he know that he has broken the cryptosystem of Alice and Bob? Alternatively, suppose Carl has come up with a method for decrypting ciphertext from Alice; how does Carl know that his method of

[10] See Patterson (1987) for an overview of cryptography.

decryption is the one actually being employed by Bob? This is a problem of underdetermination. Bob and Carl might devise alternate methods for decrypting the messages coming from Alice. Since Alice and Bob have agreed in advance what constitutes the proper way of decrypting Alice's ciphertext transmissions, Bob knows that his reading of Alice's transmissions is correct. But how does Carl know that his reading of Alice's transmissions is correct (i.e., that it is the same as Bob's)? Finally, suppose Carl monitors signals along a communication channel, suspecting that Alice and Bob are using it to communicate. Prior to breaking the cryptosystem, how can Carl be sure the signals are not random but meaningful? How does Carl determine the difference? How does Carl know that Alice isn't just sending Bob gibberish?

Let us now consider these questions in turn. When do Alice and Bob know their cryptosystem has been compromised? Clearly this will depend on the occurrence of one too many unhappy coincidences. Suppose the last ten times a field marshal ordered an attack, he sent his field commanders encrypted instructions, and each time the enemy took appropriate countermeasures. If the field marshal can preclude treachery within his ranks, after a while he will be convinced that the enemy is reading his mail – that is, intercepting the transmissions to his commanders and decrypting them. Eventually it will become too unlikely that the enemy happened by chance to apply just the right countermeasures every time the field marshal issued instructions.

As a historical note, the allies in World War II did break the cryptosystems of both the Japanese and the Germans. This was one of the best kept secrets of the war, and one that saved countless allied lives. It was necessary to keep it a secret, for if the Nazis had suspected that the British and Americans had broken their cryptosystem (in this case the Enigma cryptosystem), they would promptly have changed cryptosystems. The allies therefore purposely allowed some of the Nazis' plans to succeed lest the Nazis suspect their cryptosystem had been compromised. The actual practice of cryptography is very much a poker game, giving up an advantage here to obtain a bigger advantage elsewhere.

The cryptosystems employed by the Nazis in World War II were not very secure, even by the standards of the time, and certainly not by today's standards. Most of the cryptosystems of the past are easily broken given the computational resources of the present. But suppose a cryptosystem is known to be highly secure irrespective of the computational resources available now or at some future date (the one-time pad constitutes such a cryptosystem). For such a cryptosystem, the probabilistic p value for breaking it is so small that all the computational resources that might ever be available to us still do not make breaking it likely. What shall we then conclude if we become convinced such a cryptosystem has been broken? In practice we look for a fifth column. A fifth column, instead of breaking the cryptosystem fair and square, goes into our files and steals the answer key.

A compromised cryptosystem that in the absence of a fifth column is supposed to be highly secure has in all likelihood been subverted by a fifth column. Given such a compromised cryptosystem, counter-intelligence services are quick to conduct an investigation, looking for everything from negligence to outright treachery to explain how the cryptosystem was leaked to the enemy. Note that the chance hypothesis – that the enemy by adroitly searching the space of possible cryptosystems happened to figure out the right method of decryption – does not cut any ice if the cryptosystem is known to be highly secure (e.g., a one-time pad). Highly secure cryptosystems are not broken by chance, or even by ingenuity, but by cheating. Genius can take us only so far. Or as the proverb goes, the difference between genius and stupidity is that genius has its limits.

Let us consider next the question of underdetermination. Though in principle underdetermination might be a problem, in practice it turns out never to be a problem for cryptography. If a proposed decryption scheme is successful at coherently interpreting numerous ciphertext transmissions, the cryptosystem is considered broken. Breaking a cryptosystem is like finding a key that turns a lock. Once the lock turns, we're sure the door will open. True, there is always the possibility that the lock will turn without the door opening. But as a proposed decryption scheme assigns a coherent sense not only to prior transmissions, but also to incoming transmissions of ciphertext, any doubts about the correctness of the decryption scheme disappear. Note that underdetermination is never a problem for Alice and Bob: having specified in advance the encryption-decryption scheme they will be using, Bob is limited to only one way of interpreting Alice's ciphertext transmissions. Underdetermination is potentially a problem only for Carl, who in finding a way to make sense of Alice's ciphertext transmissions will want to make sure his way of doing so is the same as Bob's. In practice Carl attains this assurance simply by finding he can assign a coherent sense to the ciphertext transmissions coming his way.

Finally, suppose Carl monitors a communication channel, suspecting that Alice is sending Bob encrypted messages over this channel. Suppose, moreover, that Carl has yet to discover a way to interpret the signals crossing the channel. Carl therefore wonders, Do the signals constitute encrypted messages (and thus convey meaning) or are they merely random, meaningless signals (Alice, for instance, might be flipping a coin and signaling a sequence of coin flips)? There is only one way for Carl to find out: provisionally assume the signals moving across the communication channel encrypt meaningful messages, and try to figure out a method of encryption-decryption that makes sense out of the signals. Without such a method, Carl cannot know for certain whether a meaningful message is in fact being relayed, or whether he is just listening to noise. Short of knowing what Alice is actually doing on her end of the communications channel (e.g., flipping a coin or formulating meaningful English sentences), Carl can be sure that the channel is being used as a conduit for meaningful communication only if he can devise a decryption scheme that is able to render meaningful the transmissions moving across the channel.

Notice that there is an asymmetry here. Just from listening to the communication channel, Carl may legitimately conclude that meaningful messages are being transmitted across the channel – Carl has simply to come up with a decryption scheme that renders Alice's transmissions meaningful. But just from listening to the channel, Carl may never legitimately conclude that no meaningful communication is being transmitted – Carl's inability to discover a decryption scheme is never evidence that no such scheme is operating. This is true even if the signals crossing the channel are highly repetitive or simple from a communication engineer's point of view, for by agreeing in advance that a simple transmission signifies a complicated message, Alice and Bob can expand without limit the information content of even a simple transmission.[11]

Cryptography provides a framework for understanding the SETI research program (SETI = Search for Extraterrestrial Intelligence).

[11] As Fred Dretske (1981, p. 51) remarks, "there simply is *no limit* to what can be learned from a particular signal about another state of affairs."

Over the past several years radio observatories have been employed to monitor millions of radio channels in the hope of detecting radio transmissions from space that reliably indicate ETIs (extraterrestrial intelligences). Since unlike the ETIs on *Star Trek*, genuine ETIs are presumed not to communicate in English, or any other human language for that matter, the problem of determining when a radio transmission is the product of an intelligence falls under the cryptographic framework just described. Indeed, the SETI researcher's task is to eavesdrop on interplanetary communications, trying to determine whether a given radio signal was transmitted by an intelligent agent, and if so, the signal's meaning.

Of course, the actual cryptography employed by SETI researchers is pretty minimal. Typically it is assumed that ETIs are so intent on making their presence known that they will do something terribly obvious, like transmit a sequence of prime numbers. For the SETI program to have even a chance of being successful, the following hypotheses must hold: (1) ETIs must at some point in the history of the universe have existed; (2) ETIs have been sufficiently advanced technologically to signal their presence by means of radio signals; (3) ETIs have indeed signaled their presence by means of radio transmissions; and (4) we happen to be living at just the right time in cosmic history to receive those transmissions.

These hypotheses are sufficiently tenuous that SETI researchers avoid the further complication of asking whether ETIs are communicating enigmatically, and thereby making it difficult to discern their presence. Any "cryptosystem" the ETIs are employing is therefore assumed to be strictly minimal and unintended. The ETIs, we assume, want desperately to make their presence known. If, therefore, we need to do any cryptanalysis, it is solely because the means by which ETIs communicate are so foreign to ours.

If the SETI program ever proves successful (something it has yet to do), its success will consist in drawing a successful design inference, matching radio transmissions it has monitored with patterns it deems clear and reliable indicators of intelligence. As it monitors millions of radio channels, SETI attempts to match patterns it has specified in advance. Insofar as SETI fails to specify the patterns employed by ETIs, SETI will fail to detect them – their presence will slip past the SETI researchers' sieve. Regardless whether one thinks SETI constitutes an ill-fated research program, it raises important questions about the nature of intelligence, the possibility of detecting intelligences other than human, and the role of design inferences in detecting intelligence.

1.7 RANDOMNESS

In the 1960s, the Russian probabilist Andrei Kolmogorov investigated what makes a sequence of coin flips random. If we flip a fair coin and note the occurrences of heads and tails in order, denoting heads by 1 and tails by 0, then a sequence of 100 coin flips looks as follows:

(R) 1100001101011000110111111
1010001100011011001110111
0001100100001011110111011
0011111010010100101011110.

This is in fact a sequence I have just now obtained by flipping a penny 100 times. Alternatively, I might have obtained the following sequence:

(N) 111111111111111111111111
11111111111111111111111111
111111111111111111111111
11111111111111111111111111.

Now the problem facing Kolmogorov was this: given probability theory and its usual way of computing probabilities for coin tosses, Kolmogorov was unable to distinguish these sequences in terms of their degree of randomness. Sequences (R) and (N) have been labeled suggestively, R for "random," N for "nonrandom." Kolmogorov wanted to say that (R) was "more random" than (N). But given the usual way of computing probabilities, Kolmogorov could only say that each of these sequences had the same small probability of occurring, namely 1 in 2^{100}, or approximately 1 in 10^{30}. Indeed, every sequence of 100 coin tosses has exactly this same small probability of occurring.

To get around this difficulty Kolmogorov introduced some concepts from recursion theory, a subfield of mathematical logic concerned with computation and generally considered quite far removed from probability theory. What Kolmogorov said was that a string of 0s and 1s becomes increasingly random as the shortest computer program that generates the string increases in length (Kolmogorov, 1965). For our purposes we can think of a computer program as a shorthand description of a sequence of coin tosses. Thus, the sequence (N) is not very random because it has a very short description, namely,

```
repeat '1' a hundred times.
```

Note that we are interested in the shortest descriptions since any sequence can always be described in terms of itself. Thus, (N) has the longer description

```
copy '1111111111111111111111111
      1111111111111111111111111
      1111111111111111111111111
      1111111111111111111111111'.
```

But this description holds no interest since there is one so much shorter.

The sequence

```
(H)   1111111111111111111111111
      1111111111111111111111111
      0000000000000000000000000
      0000000000000000000000000
```

is slightly more random than (N) since it requires a longer description, for example,

```
repeat '1' fifty times, then repeat '0' fifty times.
```

So too the sequence

```
(A)   1010101010101010101010101
      0101010101010101010101010
      1010101010101010101010101
      0101010101010101010101010
```

has a short description,

```
repeat '10' fifty times.
```

The sequence (R), on the other hand, has no short and neat description (at least none that has yet been discovered). For this reason Kolmogorov would want to regard it as more random than the sequences (N), (H), and (A).

Since one can always describe a sequence in terms of itself, (R) has the description

```
copy '11000011010110001101111111
          01000110001101100111011100
      0110010000101111011101100
          1111101001010010101111011'.
```

Because (R) was constructed by flipping a coin, it is very likely that this is the shortest description of (R). It is a combinatorial fact that the vast majority of sequences of 0s and 1s have as their shortest description just the sequence itself, that is, most sequences are random in Kolmogorov's computational sense. Kolmogorov used the language of statistical mechanics to describe this fact, calling the random sequences high entropy sequences, and the nonrandom sequence low entropy sequences.[12] It follows that the collection of nonrandom sequences has small probability among the totality of sequences, so that observing a nonrandom sequence is reason to look for explanations other than chance.

To illustrate Kolmogorov's ideas, imagine someone informs you she just flipped a coin 100 times. If she hands you sequence (R), you examine it and try to discover a short description. After repeated attempts you find you cannot describe the sequence more efficiently than the sequence describes itself. Hence you conclude it is a genuinely random sequence, that is, a sequence she might well have gotten by flipping a fair coin. Of course you might be wrong – you might simply have missed some simple and short description. But until you have such a description in hand, you will suppose the sequence is random.

Next, suppose this same individual hands you the sequence (R) on a slip of paper and then disappears. A week later she reappears and says, "Guess what? Remember that sequence I handed you a week ago? Well, last night I was flipping this penny. And would you believe it, I got the same sequence as on the slip of paper." You examine the coin and observe it is a genuine U.S. government mint penny that is evenly balanced and has distinguishable sides. Moreover, she insists that each time she flipped the penny, she gave it a good jolt

[12] For the deep connection between entropy in statistical mechanics and entropy in the information theoretic sense of Kolmogorov see Yockey (1992, pp. 66–7; but note the errors in formulas 2.27 and 2.28). See also Zurek (1990).

(these were not phony flips). What do you conclude now? As before, you will not be able to find any shorter description than the sequence itself – it is a random sequence. Nevertheless, you are entirely justified rejecting her story. The problem is that the timing is all off. When she handed you the sequence a week earlier, she specified a highly improbable event. When she returned and claimed subsequently to have reproduced the sequence, she in effect claimed to prophesy an improbable chance event. Prophecy of improbable chance events is highly dubious. Indeed, anyone with this gift should be a billionaire many times over (either in Las Vegas or on Wall Street).

Finally, suppose this individual comes to you and says, "Would you believe it? I just flipped this penny 100 times, and it came up heads each time!" As before, the coin she shows you is a genuine penny and she is emphatic that hers were not phony flips. Rather than being specified in advance, this time the pattern of coin flips is specified in virtue of its low computational complexity. The sequence (N) has, in Kolmogorov's terminology, about the lowest entropy possible. There are very few sequences with descriptions as short as "repeat '1' 100 times." Once again, you would be ill-advised to trust her story. The problem is not that low-entropy sequences like (N) are highly improbable. Rather, the problem is that there are too many other sequences for which no short description can be found.

Our coin flipping friend, who claims to have flipped 100 heads in a row with a fair coin, and without phony flips, is in the same position as a lottery manager whose relatives all win the jackpot or an election commissioner whose own political party repeatedly gets the first ballot line (cf. Section 1.2). In each case public opinion rightly draws a design inference, eliminating chance and attributing fraud. Granted, the evidence is always circumstantial. What's more, our legal system has yet to think through how such evidence should be handled. Nevertheless, the inference that chance was offset by an act of intentional meddling is in each case compelling.[13]

[13] For further discussion of randomness see Dembski (1991; in press) as well as Section 5.10.

2

Overview of the design inference

2.1 THE EXPLANATORY FILTER

Whenever explaining an event, we must choose from three competing modes of explanation. These are *regularity, chance*, and *design*. To attribute an event to a regularity is to say that the event will (almost) always happen. To attribute an event to chance is to say that probabilities characterize the occurrence of the event, but are also compatible with some other event happening. To attribute an event to design is to say that it cannot reasonably be referred to either regularity or chance. Defining design as the set-theoretic complement of the disjunction regularity-or-chance guarantees that the three modes of explanation are mutually exclusive and exhaustive. It remains to show that design is significant in its own right.

The principal advantage of characterizing design as the complement of regularity and chance is that it avoids committing itself to a doctrine of intelligent agency. In practice, when we eliminate regularity and chance, we typically do end up with an intelligent agent. Thus, in practice, to infer design is typically to end up with a "designer" in the classical sense. Nevertheless, it is useful to separate design from theories of intelligence and intelligent agency. An intelligence may, after all, act to mimic regularity or chance, and thereby render its actions indistinguishable from regularity or chance (cf. the discussion of cryptography and randomness in Sections 1.6 and 1.7). Anything might have an intelligent cause. Not everything can be known to have an intelligent cause. Defining design as the negation of regularity and chance avoids prejudicing the causal stories we associate with design inferences.

When called to explain an event, we therefore have a decision to make – are we going to attribute it to regularity or chance or design? To answer this question we employ a standard operating procedure. The flowchart in Figure 2.1 summarizes this procedure. It will be called the *Explanatory Filter*, or simply the *filter*. To use the

The Explanatory Filter

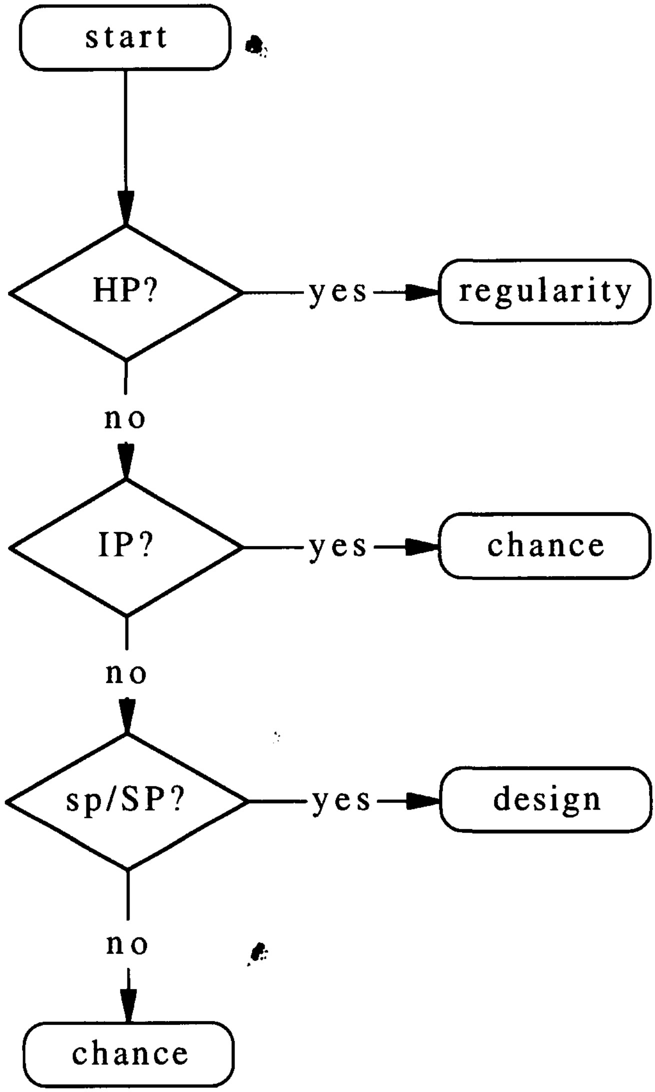


Figure 2.1.

Explanatory Filter we begin with an event E. The filter consists of two types of nodes, initial and terminal nodes represented by ovals, and decision nodes represented by diamonds. We therefore start E off at the node labeled "start." From "start" E moves to the first decision node. This node asks whether E is highly probable (hence the label HP).

To say that E is highly probable is to say that given the relevant antecedent circumstances, E will for all practical purposes always happen. HP events characterize the deterministic and nondeterministic regularities of nature (or what we frequently call natural laws), conveniently situating them within a single probabilistic framework. For instance, the event of a bullet firing when a gun's trigger is pulled and the event of getting at least one head when a fair coin is tossed a hundred times are both HP events. Generally speaking, we regard the first event as nonprobabilistic, the second as probabilistic. It is convenient to think of all such regularities as probabilistic, assimilating the nonprobabilistic case to the probabilistic case in which probabilities collapse to 0 and 1.

Thus, if E happens to be an HP event, we stop and attribute E to a regularity. Regularities are always the first line of defense. If we can explain by means of a regularity, chance and design are automatically precluded. Similarly, chance is always the second line of defense. If we can't explain by means of a regularity, but can explain by means of chance, then design is automatically precluded. There is thus an order of priority to explanation.[1] Within this order regularity has top priority, chance second, and design last.

It needs to be stressed, however, that this order of priority has nothing to do with one mode of explanation being somehow "preferable" or "better" than another. We are not offering a better explanation of, say, Nicholas Caputo's "golden arm" by attributing it to regularity or chance as opposed to design. Presumably, one or the other of these explanations is correct, so that the one to be preferred is the one that is correct. Nevertheless, as a matter of explanatory priority, we look to regularity and chance before we invoke design.

Explanatory priority is a case of Occam's razor. Accordingly, when any one of the three modes of explanation fails adequately to explain an event, we move to the mode of explanation at the next level of

[1] It is this order of priority and the concomitant sifting of explanatory options that justifies the name "Explanatory Filter."

complication. Note that explanations that appeal to regularity are indeed simplest, for they admit no contingency, claiming things always happen that way. Explanations that appeal to chance add a level of complication, for they admit contingency, but one characterized by probability. Most complicated are those explanations that appeal to design, for they admit contingency, but not one characterized by probability.

Examples of explanatory priority abound. To see that regularity has explanatory priority over design, recall Newton's classic error in calculating the dynamics of the solar system. Newton had thought that the dynamics of the solar system were unstable and therefore that the motions of the planets required slight periodic adjustments by the intervention of God (hence for Newton the proper mode of explanation for the dynamics of the solar system, though partially appealing to his laws of mechanics, also included an appeal to design, with design here taking the form of supernatural intervention). But when a century later the French mathematician Laplace showed that the dynamics of the solar system were relatively stable, and therefore did not require periodic divine adjustments, regularity was reinstated as the proper mode of explanation for the dynamics of the solar system.

Similarly, to see that regularity has explanatory priority over chance, consider a pair of dice that land snake-eyes (i.e., each die displays a one). If the dice are fair, this outcome has probability 1 in 36, an instance of what I call an intermediate probability (IP for short; what makes an event an IP event will become clear momentarily). Now under ordinary circumstances there is no problem attributing IP events, like the occurrence of snake-eyes, to chance. But suppose that in fact the dice are weighted so that each die will almost surely come up displaying a one (i.e., the dice are loaded). In this case the occurrence of snake-eyes is an HP event and would not be attributed to chance. Instead, the occurrence of snake-eyes would be attributed to a regularity, namely, the regular way dice land when weighted in certain prescribed ways.

Finally, to see that chance has explanatory priority over design, imagine that you and your aunt Jane both live in a small town, and that occasionally when you buy groceries at the one supermarket in town, you run into your aunt. Has your aunt carefully planned these meetings? Is she monitoring the food you purchase? Is she secretly trying to poison your pickles? Attributing to intelligent forces what ought properly be attributed to chance indicates paranoia and superstition. Maybe your aunt is secretly a mass murderer, and maybe her meetings with you at the supermarket are planned. But to demonstrate this you will need more than an innocent encounter at the supermarket.

To return now to the event E as it proceeds through the Explanatory Filter, suppose E is not an HP event, and has therefore passed to the next decision node, the node labeled "IP?" What needs to be determined here is whether E is an event of intermediate probability. Events of intermediate probability, or what I'm calling IP events, are the events we reasonably expect to occur by chance in the ordinary circumstances of life. Rolling snake-eyes with a pair of fair dice constitutes an IP event. Even someone winning a lottery where the probability of winning is as little as one in ten million will constitute an IP event once we factor in all the other people playing the lottery. IP events are sufficiently probable to give us no reason to suspect they resulted from anything other than chance. Thus, if an event E reaches the second decision node and is judged an IP event, we stop and attribute E to chance.

But suppose E is neither an HP nor an IP event. E therefore proceeds to the third and final decision node of the flowchart. In this case E is an event of small probability, or what I'm calling an SP event. As we saw in Chapter 1, SP events happen by chance all the time – flip a coin long enough and you'll participate in a highly improbable event. To eliminate chance it is therefore not enough simply to know that E is an SP event. Rather, an extraprobabilistic factor must also be introduced, what in Chapter 1 we referred to as specification. A probabilistic setup like tossing a coin 1000 times entails that some SP event will occur. If, however, this event is also specified, then we are justified in eliminating chance. It's the specified SP events (abbreviated sp/SP events) that cannot properly be attributed to chance. Richard Dawkins (1987, p. 8) makes this point as follows:

> Hitting upon the lucky number that opens the bank's safe is the equivalent, in our analogy, of hurling scrap metal around at random and happening to assemble a Boeing 747. Of all the millions of unique and, with hindsight equally improbable, positions of the combination lock, only one opens the lock. Similarly, of all the millions of unique and, with hindsight equally improbable, arrangements of a heap of junk, only one (or very few) will fly. The uniqueness of the arrangement that flies, or that opens the safe, [has] nothing to do with hindsight. It is *specified in advance*.

Specifications are common in statistics, where they are known as rejection regions. The basic problem of statistics is the testing of probabilistic hypotheses. Statisticians are continually confronted with situations in which some probability distribution is assumed to be operating, and then given the task of determining whether this probability distribution actually is operating. To do this statisticians set up a rejection region and then take a sample. If the observed sample falls within the rejection region, the presumption is that the probability distribution in question was not operating to produce the sample. On the other hand, if the observed sample falls outside the rejection region, the probability distribution in question is taken, at least provisionally, as adequate to account for the sample.[2]

Now there is an important difference between the logic of the Explanatory Filter and the logic of statistical hypothesis testing: to end up at the filter's terminal node labeled "design" is to sweep the field clear of all relevant chance explanations. This contrasts with statistical hypothesis testing, where eliminating one chance hypothesis typically opens the door to others. The sp/SP events of the Explanatory Filter exclude chance decisively, whereas the events that fall within the rejection regions of statistics may allow that some probability distribution other than the one originally suspected is operating.

For an event E to pass to the third decision node of the Explanatory Filter, it is therefore not enough to know that E has small probability with respect to some probability distribution or other. Rather, we must know that whatever probability distribution may have been responsible for E, it wasn't one for which the probability of E was either an HP or an IP event. Thus, unlike the statistician, who typically operates from a position of ignorance in determining what probability distribution is responsible for the samples he or she observes, before we even begin to send E through the Explanatory Filter, we need to know what probability distribution(s), if any, were operating to produce the event. Alternatively, the "design theorist" is in the business of categorically eliminating chance, whereas the statistician is in the business of ruling out individual probability distributions.

There is thus a crucial difference between the way statistics eliminates chance and the way the design inference eliminates chance.

[2] For a general description of statistical reasoning see Hacking (1965) as well as Howson and Urbach (1993). See Mood, Graybill, and Boes (1974, pp. 401–81) for an account of hypothesis testing.

When statistics eliminates chance, it is always a particular probability distribution (or set of probability distributions) that gets eliminated, with the question remaining what alternative probability distributions might be operating in its place. On the other hand, when the design inference eliminates chance, it leaves no room for alternative probability distributions. Statistics, therefore, eliminates chance only in the limited sense of rejecting one chance explanation while leaving open another. The design inference, on the other hand, eliminates chance in the global sense of closing the door to every relevant chance explanation.

To clarify the distinction between statistical inferences and design inferences, consider an example of each. When a statistician wants to determine whether a certain fertilizer will help improve the yield of a certain crop, she assumes initially that the fertilizer will have no effect. Her research presumably is going to be used by farmers to decide whether to use this fertilizer. Because fertilizer costs money, she doesn't want to recommend the fertilizer if in fact it has no positive effect on the crop. Initially she will therefore assume the fertilizer has no effect one way or the other.

Having made this assumption, she will now compare how the crop fares with both fertilized and unfertilized soil. She will therefore conduct an experiment, raising, let us say, one acre of the crop in fertilized soil, and another in unfertilized soil. In raising the crop, she will try to keep all sources of variation other than the fertilizer constant (e.g., water and sunshine). Our statistician does not know in advance how the crop will fare in the fertilized soil. Nevertheless, it's virtually certain she won't observe exactly the same yields from both acres.

The statistician's task, therefore, is to determine whether any difference in yield is due to the intrinsic effectiveness of the fertilizer or to chance fluctuations. Let us say that the fertilized crop yielded more than the unfertilized crop. Her task then is to determine whether this difference is significant enough to overturn her initial assumption that the fertilizer has no effect on the crop. If there is only a one percent increase in the fertilized crop over the unfertilized crop, this may not be enough to overthrow her initial assumption. On the other hand, a 100 percent increase in the fertilized crop would surely overthrow the statistician's initial assumption, providing evidence that the fertilizer is highly effective.

The point then is this. Our statistician started by assuming that the probability distribution characterizing crop yield for fertilized soil is the same as the probability distribution characterizing crop yield for unfertilized soil. She made this assumption because without prior knowledge about the effectiveness of the fertilizer, it was the safest assumption to make. Indeed, she doesn't want to recommend that farmers buy this fertilizer unless it significantly improves their crop yield. At least initially she will therefore assume the fertilizer has no effect on crops, and that differences in crop yield are due to random fluctuations. The truth of this initial assumption, however, remains uncertain until she performs her experiment. If the experiment indicates little or no difference in crop yield, she will stick with her initial assumption. But if there is a big difference, she will discard her initial assumption and revise her probability distribution. Note, however, that in discarding her initial assumption and revising her probability distribution, our statistician has not eliminated chance, but merely exchanged one chance hypothesis for another. Probabilities continue to characterize crop yield for the fertilized soil.[3]

Let us now contrast the probabilistic reasoning of this statistician with the probabilistic reasoning of a design theorist who, to use the preceding example by Richard Dawkins (1987, p. 8), must explain why a certain bank's safe that was closed earlier happens now to be open. Let us suppose the safe has a combination lock that is marked with a hundred numbers ranging from 00 to 99, and for which five turns in alternating directions are required to open the lock. We assume that precisely one sequence of alternating turns is capable of opening the lock (e.g., 34-98-25-09-71). There are thus ten billion possible combinations of which precisely one opens the lock (as Dawkins puts it, "of all the millions of unique and, with hindsight equally improbable, positions of the combination lock, only one opens the lock").

We assume our design theorist is adept at using the Explanatory Filter. He will therefore take the opening of the bank's safe, an event he will denote by E, and feed it into the filter. How does E fare at the first decision node? Since no regularities account for the opening of safes with combination locks, E is not an HP event. E therefore moves past

[3] Observe that throughout this statistical analysis we never obtain a causal story detailing the fertilizer's effect on growing plants. For both the fertilized and the unfertilized soils all we obtain are probability distributions characterizing expected yield.

the first decision node. How does E fare at the second decision node? With ten billion possibilities, only one of which opens the safe, random twirling of the combination lock's dial is exceedingly unlikely to open the lock. E therefore isn't an IP event either. E therefore lands at the third decision node. In particular, E is an SP event.

Now the crucial thing to observe here is that the status of E as an SP event is not open to question in the way the effectiveness of fertilizers was open to question in the last example. The design theorist's initial assessment of probability for the combination lock is stable; the statistician's initial assessment about the effectiveness of a given fertilizer is not. We have a great deal of prior knowledge about locks in general, and combination locks in particular, before the specific combination lock we are considering crosses our path. On the other hand, we know virtually nothing about a new fertilizer and its effect on crop yield until after we perform an experiment with it. Whereas the statistician's initial assessment of probability is likely to change, the design theorist's initial assessment of probability is stable. The statistician wants to exclude one chance explanation only to replace it with another. The design theorist, on the other hand, wants to exclude the only available chance explanation, replacing it with a completely different mode of explanation, namely, a design explanation.[4]

Let us now return to the passage of E through the Explanatory Filter. As an SP event, E proceeds past the first two decision nodes to the third and final decision node. There the crucial question is whether E is also specified. Is E an sp/SP event or merely an SP event? If the latter, then the opening of the bank's safe can legitimately be attributed to chance (just as apart from a specification a highly improbable sequence of coin flips can be attributed to chance). But of course E is specified. Indeed, the very construction of the lock's tumblers specifies which

[4] The safecracking example presents the simplest case of a design inference. Here there is only one possible chance hypothesis to consider, and when it is eliminated, any appeal to chance is effectively dead. Design inferences in which multiple chance hypotheses have to be considered and then eliminated arise as well. We might, for instance, imagine explaining the occurrence of a hundred heads in a row from a coin that is either fair or weighted in favor of heads with probability of heads 0.75. To eliminate chance and infer design in this example we would have to eliminate two chance hypotheses, one where the probability of heads is 0.5 (i.e., the coin is fair) and the other where the probability of heads is 0.75. To do this we would have to make sure that for both probability distributions a hundred heads in a row is an SP event, and then show that this event is also specified. In case still more chance hypotheses are operating, design follows only if each of these additional chance hypotheses gets eliminated as well, which means that the event has to be an SP event with respect to all the relevant chance hypotheses and in each case be specified as well.

one of the ten billion combinations opens the lock. E is therefore an sp/SP event and passes to the terminal node labeled "design." Our design theorist will therefore infer design, and in all likelihood conclude that an intelligent agent opened the safe (e.g., the safe's owner or a safe-cracking thief).

How to assign probabilities to events passing through the Explanatory Filter requires some clarification, especially how to decide whether an event E is an HP, IP, or SP event. I shall elaborate on the conception of probability relevant to the design inference and the Explanatory Filter in Chapter 3. But briefly, probability is a relation between background information and events. If this background information includes (1) that certain antecedent circumstances were satisfied, (2) that the occurrence of E has with at most rare exceptions attended these antecedent circumstances, and (3) that E actually has occurred, then it follows that E is an HP event and that a regularity explains E. Given background information satisfying (1)–(3), we answer "Yes" at the decision node labeled "HP?" and conclude that E is due to a regularity.

But suppose E doesn't satisfy (1)–(3). With regularity no longer in the running, the problem now is to determine whether E is due to chance or design. Moreover, the way to do this is to assume provisionally that chance was responsible for E, and then determine whether this assumption can be maintained. This accords with my earlier remarks about explanatory priority and Occam's razor. Having eliminated regularity, we are down to two explanations, chance or design. Since chance is the simpler explanation, barring strong countervailing reasons we prefer to explain by reference to chance as opposed to design. Such strong countervailing reasons arise when E is an sp/SP event – but we are getting ahead of ourselves.

Once we have eliminated regularity as an explanation of E, we therefore assume provisionally that E occurred by chance, and then determine the probability of E under this assumption. With our explanatory options reduced to chance and design, we identify the background information that accounts for how E could have arisen by chance. This background information enables us to assign a probability to E (or a range of probabilities). If this probability is an IP, we reject design and attribute E to chance. If, on the other hand, this probability is an SP, we then have the additional task of determining whether E is specified. Given our provisional assumption that E occurred by chance, the finding that E is a specified event of small probability acts as a probabilistic *reductio ad absurdum*, leading us to reject our provisional assumption of chance, and instead infer design (cf. Section 6.3).

Note that this way of opposing chance to design requires that we be clear what chance processes could be operating to produce the event in question. Suppose, for instance, I have before me a hundred pennies all of which have landed heads. What is the probability of getting all 100 pennies to exhibit heads? This probability depends on the chance process controlling the pennies. If, for instance, the chance process flips each penny individually and waits until it lands heads, after 200 flips there will be an even chance that all the pennies exhibit heads (and after a thousand flips it will be almost certain that all the pennies exhibit heads). If, on the other hand, the chance process operates by flipping all the pennies simultaneously, and does not stop until all the pennies simultaneously exhibit heads, it will require about 10^{30} such simultaneous flips for there to be an even chance that all the pennies exhibit heads.

But suppose we know there were no more than 200 occasions for flipping the pennies, and that all the pennies exhibit heads. If the pennies were flipped individually, the probability that all of them exhibit heads is 1/2. On the other hand, if the pennies were flipped simultaneously, the probability that all of them exhibit heads is less than one in 10^{27}. The first of these probabilities is an intermediate probability, whereas the second is for most purposes a small probability. Now 100 heads in a row is a specified event (cf. the discussion of randomness in Section 1.7). If the coins could have been flipped individually, our provisional assumption that chance produced 100 heads in a row would stand since in this case even though 100 heads in a row is specified, it is merely an IP event. On the other hand, if the coins could only have been flipped simultaneously, we would reject chance and attribute the 100 heads in a row to design since in this case 100 heads in a row is an sp/SP event. The chance process that flips coins individually and leaves them be after they land heads thus leads to significantly different conclusions from the chance process that flips coins simultaneously.

What the Explanatory Filter means by "chance" and "design" varies from situation to situation. Consider plagiarism. Person A writes X at time t_1. Person B writes X′ at time t_2, t_2 coming after t_1. Suppose

X′ looks suspiciously similar to X. We want to know whether B copied from A. For simplicity let's assume neither common cause explanations (i.e., that A and B both copied from some third source) nor intermediary cause explanations (i.e., that B copied from C_1 who copied from C_2 . . . who copied from C_n who in turn copied from A) hold. Thus, we assume that if B did not compose X′ independently, then B copied from A. If we now let E denote the writing of X′ at time t_2 by B, an investigation into whether B copied from A can be recast as an application of the filter to E. In this scenario, to say that E occurred by chance is to say that B produced X′ without recourse to X, whereas to say that E occurred by design is to say that B copied from A. But note that in either case E represents an intentional act.

To sum up this section, the Explanatory Filter faithfully represents our ordinary human practice of sorting through events we alternately attribute to regularity, chance, or design. In particular, passage through the flowchart to the terminal node labeled "design" encapsulates the design inference. This inference characterizes

(1) how copyright and patent offices identify theft of intellectual property
(2) how insurance companies keep themselves from getting ripped off
(3) how detectives employ circumstantial evidence to nail criminals
(4) how forensic scientists place individuals at the scene of a crime
(5) how skeptics debunk the claims of parapsychologists
(6) how scientists uncover data falsification
(7) how the SETI program detects extraterrestrial intelligences
(8) how statisticians and computer scientists distinguish random from nonrandom strings of digits.

Entire human industries depend critically on the design inference. Indeed, courts have sent people to the electric chair on account of this inference. A rigorous elucidation of the design inference is therefore of more than academic interest.

2.2 THE LOGIC OF THE INFERENCE

Just because the Explanatory Filter describes how we ordinarily sift our explanatory options does not mean it correctly prescribes how we should sift our explanatory options. To see that the filter is not merely descriptive, but also normative for sifting our explanatory options, we need to understand the logic that underlies it. The Explanatory Filter has a logical counterpart that is fully suited for the rigors of scientific inquiry. Specifically, there is a valid deductive argument that traces the passage of an event E through the Explanatory Filter from the initial node labeled "start" to the terminal node labeled "design." Here is this argument:

Premise 1: E has occurred.
Premise 2: E is specified.
Premise 3: If E is due to chance, then E has small probability.
Premise 4: Specified events of small probability do not occur by chance.
Premise 5: E is not due to a regularity.
Premise 6: E is due to either a regularity, chance, or design.
Conclusion: E is due to design.[5]

Let us walk through this argument in the concrete case where E is the opening of a safe's combination lock. Premise 1 simply says that E occurred. And indeed, E better have occurred for E to need explaining. Hence we take Premise 1 to be satisfied. Because E is specified (the very construction of the safe specifies which one of the many possible combinations opens it), it follows that Premise 2 is satisfied as well. Premise 3 says that if the safe was opened by chance, then the opening of the safe was an SP event. This is no problem either. Someone twirling the dial of the combination lock has to hit just the right combination. And on any reasonable lock, the probability of hitting the right combination will be exceedingly small. Premise 4 is the Law of Small Probability, or LSP for short. LSP is the key regulative principle governing small probabilities (cf. Section 1.1). Although LSP will by now seem quite plausible, its full justification will have to await Chapter 6. Thus, provisionally we take LSP to be satisfied as well. Premise 5 says that the safe didn't open as the result of a regularity.

[5] To say that an event E is "due to" either a regularity or chance or design is not to advance a causal story for E. Regularity, chance, and design, as developed in Section 2.1, constitute distinct modes of explanation whose appropriateness depends on whether an event exhibits certain features. The causal story one eventually tells, though constrained by the mode of explanation, ultimately depends on context. The statement "E is due to design" is thus a shorthand for "the proper mode of explanation for E is design." Similarly, "E is due to a regularity" and "E is due to chance" are shorthands for "the proper mode of explanation for E is a regularity" and "the proper mode of explanation for E is chance," respectively.

This premise too is satisfied since no known regularities account for the opening of safes with reasonably sophisticated combination locks. Premise 6 is the trichotomy rule. Trichotomy holds because regularity, chance, and design are mutually exclusive and exhaustive (cf. Section 2.1). Together these six premises entail the conclusion that the opening of the safe's combination lock must properly be attributed to design. And of course, except for the most trivial combination locks, we do attribute their opening not to regularities or chance, but to the intentional activity of a human agent.

The validity of the preceding argument becomes clear once we recast it in symbolic form (note that E is a fixed event and that in Premise 4, X is a bound variable ranging over events):

Premise 1: $oc(\mathrm{E})$
Premise 2: $sp(\mathrm{E})$
Premise 3: $ch(\mathrm{E}) \to SP(\mathrm{E})$
Premise 4: $\forall \mathrm{X}[oc(\mathrm{X})\ \&\ sp(\mathrm{X})\ \&\ SP(\mathrm{X}) \to \sim ch(\mathrm{X})]$
Premise 5: $\sim reg(\mathrm{E})$
Premise 6: $reg(\mathrm{E}) \vee ch(\mathrm{E}) \vee des(\mathrm{E})$
Conclusion: $des(\mathrm{E})$.

So formulated, the design inference constitutes a valid argument within the first-order predicate logic. There are six one-place predicates: $oc(\mathrm{E})$ = E has occurred; $sp(\mathrm{E})$ = E is specified; $SP(\mathrm{E})$ = E is an event of small probability; $reg(\mathrm{E})$ = E is due to a regularity; $ch(\mathrm{E})$ = E is due to chance; and $des(\mathrm{E})$ = E is due to design. The sentential connectives $\sim$, &, $\vee$, and $\to$ denote respectively *not, and, or*, and *if–then*. $\forall$X is the universal quantifier (read "for all X" or "for every X"). Premise 4, the Law of Small Probability, is the only premise employing a quantifier. The proof that this argument is valid is straightforward.[6]

Although the preceding formulation of the design inference is substantially correct, it needs to be refined. Three of the predicates, namely

[6] Instantiate Premise 4 by substituting E for X. Provisionally assume E is due to chance, i.e., $ch(\mathrm{E})$. Then by applying modus ponens to Premise 3, infer E is a small probability event, i.e., $SP(\mathrm{E})$. Conjoin $SP(\mathrm{E})$ with Premises 1 and 2. This conjunction is identical with the antecedent of Premise 4 after instantiation. Thus, by modus ponens it follows that E is not due to chance, i.e., $\sim ch(\mathrm{E})$. This contradicts our provisional assumption, and so, by *reductio ad absurdum*, it follows that E is indeed not due to chance, i.e., $\sim ch(\mathrm{E})$. But by disjunctive syllogism Premises 5 and 6 yield $ch(\mathrm{E}) \vee des(\mathrm{E})$. And since we just derived $\sim ch(\mathrm{E})$, another application of disjunctive syllogism yields $des(\mathrm{E})$. This is the desired conclusion.

ch, *SP*, and *sp*, need to be indexed by a chance hypothesis, which we'll denote by H. Thus, to say that an event E occurred by chance is to say that E occurred according to the chance hypothesis H; to say that E has small probability is to say that E has small probability with respect to H; and to say that E is specified is to say that E is specified with respect to H. Rewriting the predicates $ch(\cdot)$, $SP(\cdot)$, and $sp(\cdot)$ as respectively $ch(\cdot; \mathrm{H})$, $SP(\cdot; \mathrm{H})$, and $sp(\cdot; \mathrm{H})$ leads to the following reformulation of the design inference:

Premise 1: $oc(\mathrm{E})$
Premise 2: $sp(\mathrm{E}; \mathrm{H})$
Premise 3: $SP(\mathrm{E}; \mathrm{H})$
Premise 4: $\forall \mathrm{X}[oc(\mathrm{X})\ \&\ sp(\mathrm{X}; \mathrm{H})\ \&\ SP(\mathrm{X}; \mathrm{H}) \rightarrow \sim ch(\mathrm{X}; \mathrm{H})]$
Premise 5: $\sim reg(\mathrm{E})$
Premise 6: $reg(\mathrm{E}) \vee ch(\mathrm{E}; \mathrm{H}) \vee des(\mathrm{E})$
Conclusion: $des(\mathrm{E})$.

Except for Premise 3, the only change in this formulation is a substitution of predicates, substituting predicates indexed by H for nonindexed ones. As for Premise 3, dropping the conditional now makes sense because indexing *SP* by H predicates of E that as an event due to H it would have small probability (whether E actually occurred by chance thus becomes immaterial). The antecedent in $ch(\mathrm{E}) \rightarrow SP(\mathrm{E})$ therefore becomes redundant as soon as we index *SP* by H. The proof that this revised argument is valid is also straightforward.[7]

This last formulation of the design inference is adequate so long as only one chance hypothesis could be responsible for E. Thus, for the opening of a safe's combination lock, this last formulation is just fine. If, however, multiple chance hypotheses could be responsible for E, this formulation needs to be modified. Letting $\mathcal{H}$ denote all the relevant chance hypotheses that could be responsible for E, for the design inference to sweep the field clear of these chance hypotheses, it needs to be reformulated as follows:

Premise 1: $oc(\mathrm{E})$
Premise 2: $(\forall \mathrm{H} \in \mathcal{H})\ sp(\mathrm{E}; \mathrm{H})$

[7] Instantiate Premise 4 by substituting E for X. Conjoin Premises 1, 2, and 3. By modus ponens, infer $\sim ch(\mathrm{E}; \mathrm{H})$. By applying disjunctive syllogism to Premise 6 infer $reg(\mathrm{E}) \vee des(\mathrm{E})$. Now take Premise 5 and apply disjunctive syllogism to $reg(\mathrm{E}) \vee des(\mathrm{E})$. This yields $des(\mathrm{E})$, the desired conclusion.

Premise 3: $(\forall H \in \mathcal{H})\ SP(E; H)$
Premise 4: $(\forall X)(\forall H \in \mathcal{H})\ \{[oc(X)\ \&\ sp(X; H)\ \&\ SP(X; H)] \rightarrow \sim ch(X; H)\}$
Premise 5: $\sim reg(E)$
Premise 6: $reg(E) \vee (\exists H \in \mathcal{H})\ ch(E; H) \vee des(E)$
Conclusion: $des(E)$.

Except for one small refinement to be described in Chapter 6, this is the definitive formulation of the design inference. Note that $\exists H \in \mathcal{H}$ is a restricted existential quantifier (read "for some H in $\mathcal{H}$"). Note also that when $\mathcal{H} = \{H\}$ (i.e., when only one chance hypothesis could be responsible for E), this formulation reduces to the previous one.

So formulated, the design inference constitutes a valid deductive argument in the first-order predicate logic. To see this, fix a chance hypothesis H and consider the following argument:

Premise 1′: $oc(E)$
Premise 2′: $sp(E; H)$
Premise 3′: $SP(E; H)$
Premise 4′: $\forall X[oc(X)\ \&\ sp(X; H)\ \&\ SP(X; H) \rightarrow \sim ch(X; H)]$
Conclusion′: $\sim ch(E; H)$.

The validity of this argument is immediate: conjoin Premises 1′ through 3′, universally instantiate Premise 4′ substituting E for X, and then apply modus ponens. Since Conclusion′ eliminates only a single chance hypothesis H, to eliminate all the chance hypotheses relevant to E's occurrence, the preceding argument needs to hold for all H in $\mathcal{H}$. To represent this logically we place each line of the preceding argument under the scope of the quantifier $\forall H \in \mathcal{H}$. Premises 1′ through 4′ thereby become Premises 1 through 4, and Conclusion′ becomes $(\forall H \in \mathcal{H}) \sim ch(E; H)$.

A double application of disjunctive syllogism to Premise 6 now produces the desired conclusion. These applications successively strip the disjuncts $reg(E)$ and $(\exists H \in \mathcal{H})\ ch(E; H)$ from Premise 6. Specifically, $\sim reg(E)$ in Premise 5 strips $reg(E)$ from Premise 6 to yield $(\exists H \in \mathcal{H}) ch(E; H) \vee des(E)$. Next $(\forall H \in \mathcal{H}) \sim ch(E; H)$, which was derived from Premises 1 through 4 and is the negation of $(\exists H \in \mathcal{H})\ ch(E; H)$, strips the remaining obstacle to $des(E)$. This establishes the validity of the design inference.

Next, let us consider Premises 1 through 6 individually.

Premise 1: This premise asserts that E has occurred. Since the occurrence of an event can be asserted irrespective of any chance hypothesis that might explain its occurrence, there is no need in this premise to invoke a chance hypothesis H, nor for that matter the collection of chance hypotheses $\mathcal{H}$ relevant to E's occurrence. Since *oc* is independent of any chance hypothesis H, quantifying over $\mathcal{H}$ leaves Premise 1 unaffected.

Premise 2: This premise asserts that E is specified for all chance hypotheses relevant to E's occurrence (i.e., for all chance hypotheses in $\mathcal{H}$). Since an event has to be specified to eliminate chance, and since the design inference infers design by eliminating all relevant chance hypotheses, *sp*(E; H) has to be satisfied for all H in $\mathcal{H}$.

Premise 3: This premise asserts that E has small probability for all chance hypotheses relevant to E's occurrence (i.e., for all chance hypotheses in $\mathcal{H}$). Since an event has to have small probability to eliminate chance, and since the design inference infers design by eliminating all relevant chance hypotheses, *SP*(E; H) has to be satisfied for all H in $\mathcal{H}$.

Premise 4: The Law of Small Probability asserts that for an arbitrary event X and an arbitrary chance hypothesis H, if X occurred, is specified with respect to H, and has small probability with respect to H, then the occurrence of X was not governed by the chance hypothesis H. Formally this is expressed by writing

$$oc(\mathrm{X})\ \&\ sp(\mathrm{X}; \mathrm{H})\ \&\ SP(\mathrm{X}; \mathrm{H}) \rightarrow \sim ch(\mathrm{X}; \mathrm{H}).$$

Since in this formula X and H are unconstrained, we can eliminate the free variables and express the Law of Small Probability in closed form as follows:

$$\forall \mathrm{X}\ \forall \mathrm{H}[oc(\mathrm{X})\ \&\ sp(\mathrm{X}; \mathrm{H})\ \&\ SP(\mathrm{X}; \mathrm{H}) \rightarrow \sim ch(\mathrm{X}; \mathrm{H})].$$

But since this last formula holds for all chance hypotheses whatsoever, it holds for all chance hypotheses relevant to E's occurrence, that is, for all chance hypotheses H in $\mathcal{H}$. By restricting the universal quantifier $\forall$H to $\mathcal{H}$, we therefore obtain

$$\forall \mathrm{X}(\forall \mathrm{H} \in \mathcal{H})[oc(\mathrm{X})\ \&\ sp(\mathrm{X}; \mathrm{H})\ \&\ SP(\mathrm{X}; \mathrm{H}) \rightarrow \sim ch(\mathrm{X}; \mathrm{H})].$$

This last version of the Law of Small Probability coincides with Premise 4. We need it to derive $(\forall \mathrm{H} \in \mathcal{H}) \sim ch(\mathrm{E}; \mathrm{H})$ from Premises 1 through 4, the claim that E is not due to any of the chance hypotheses in $\mathcal{H}$.

Premise 5: This premise asserts that no regularity was responsible for E. We can represent a regularity **R** as an ordered pair ({A}, f) where {A} is a set of antecedent circumstances and f is a mapping from {A} to events. For a regularity **R** = ({A}, f) to account for E then means that for some A_0 in {A}, A_0 was satisfied and $f(A_0) = E$. To predicate *reg* of E may then be defined as follows:

$$reg(E) =_{\text{def}} \text{There is some regularity } \mathbf{R} \text{ that accounts for E.}$$

Justifying *reg*(E) is straightforward, for it amounts to no more than checking whether some regularity **R** accounts for E. On the other hand, justifying ~*reg*(E) tends to be trickier, for what must be argued is that no regularity accounts for E. In practice, the way ~*reg*(E) gets justified is by arguing that E is compatible with all relevant natural laws (natural laws are the regularities that govern natural phenomena, e.g., the laws of chemistry and physics), but that these natural laws permit any number of alternatives to E. In this way E becomes irreducible to natural laws, and thus unexplainable in terms of regularities.[8]

The safe-cracker example considered earlier in this section illustrates this method of eliminating regularities. The laws of physics prescribe two possible motions of the combination lock, viz., clockwise and counterclockwise turns. Dialing any particular combination is compatible with these possible motions of the lock, but in no way dictated by these motions. To open the lock by hitting the right combination is therefore irreducible to these possible motions of the lock.

This method of eliminating regularity applies not just to the opening of combination locks, but quite generally: The position of scrabble pieces on a scrabble board is irreducible to the natural laws governing the motion of scrabble pieces; the configuration of ink on a sheet of paper is irreducible to the physics and chemistry of paper and ink; the sequencing of DNA bases is irreducible to the bonding affinities between the bases; and so forth. In each case what defeats regularity is contingency – a contingency compatible with the regularities known to be operating, but in no way determined by them.

[8] This method for eliminating regularity is described in Lenoir (1982, pp. 7–8), Polanyi (1967; 1968), and Yockey (1992, p. 335).

Premise 6: This is the trichotomy rule. It asserts that E is due to either regularity or chance or design. The three disjuncts are mutually exclusive and exhaustive. Given that $\mathcal{H}$ includes all the relevant chance hypotheses that might explain E, to assert that E is due to chance means that there is some H in $\mathcal{H}$ that satisfies *ch*(E; H), that is, $(\exists H \in \mathcal{H})ch(E; H)$. This is the second disjunct in Premise 6. Since this disjunct requires that E be contingent, and since *reg*(E) precludes E from being contingent, it follows that the first two disjuncts of Premise 6 are mutually exclusive. To assure that the three disjuncts in Premise 6 are both mutually exclusive and exhaustive, it is therefore necessary to define design as the negation of the disjunction regularity-or-chance (cf. Section 2.1). We therefore define *des*(E) as the negation of $reg(E) \vee (\exists H \in \mathcal{H})ch(E; H)$. By the De Morgan rule and a standard logical manipulation with the existential quantifier, *des*(E) may therefore be defined as

$$des(E) =_{def} \sim reg(E) \;\&\; (\forall H \in \mathcal{H}) \sim ch(E; H).$$

E is therefore due to design just in case E is not due to a regularity and E is not due to chance. As the examples in Chapter 1 have made clear, this mode of explanation is nonvacuous and significant in its own right.

One last point about the logical structure of the design inference is worth noting: Just because the premises of the design inference will in most applications be only probable or assertible rather than certain or true in no way limits the significance of the design inference. The design inference constitutes a valid deductive argument. In a valid deductive argument the premises not only entail the conclusion, but also render the conclusion probable if the premises themselves are probable. This probabilistic relation between premises and conclusion is called "partial entailment." Within the logic of entailment, entailment automatically yields partial entailment (see Adams, 1975, ch. 1). Thus, whereas entailment guarantees the truth of the conclusion given the truth of the premises, entailment also guarantees that the conclusion has high probability if the premises have high probability. Entailment therefore confers not only truth, but also probability. It follows that the design inference is robust. It is not simply a piece of logic chopping. In particular, it is entirely suited to scientific inquiry and the uncertainties that so often attend scientific claims.

Although the hard work of this monograph still remains, conceptually our task is now much simpler. To complete our analysis of the design inference we need simply to explicate and justify the Law of Small Probability. This in turn reduces to accomplishing three objectives: (1) explicate the predicate *sp* (i.e., say what it is for an event to be specified); (2) explicate the predicate *SP* (i.e., say what it is for an event to have small probability); and (3) justify the Law of Small Probability (i.e., show why specified events of small probability cannot legitimately be referred to chance). All the remaining chapters of this monograph are devoted to accomplishing these objectives. Chapters 3 and 4 lay the logical and mathematical foundations. Chapter 5 explicates specification. Chapter 6 explicates small probabilities and justifies the Law of Small Probability. The remainder of this chapter presents a case study of the design inference (Section 2.3) and shows the connection between design and intelligent agency (Section 2.4).

2.3 CASE STUDY: THE CREATION–EVOLUTION CONTROVERSY

Design inferences occur widely in the creation–evolution controversy. Arguments by evolutionists that support biological evolution and arguments by creationists that oppose biological evolution frequently and self-consciously employ the logic of the design inference. To see this, let us consider two such arguments, one an antievolutionary argument by creationists Clifford Wilson and John Weldon, the other a proevolutionary argument by the Darwinist Richard Dawkins. First consider the antievolutionary argument by Wilson and Weldon (1978, pp. 320–3):

> In the October, 1969, issue of *Nature* magazine, Dr. Frank Salisbury... examined the chance of one of the most basic chemical reactions for the continuation of life taking place. This reaction involved the formation of a specific DNA molecule.... He calculated the chance of this molecule evolving on 10^{20} hospitable planets.... He concluded that the chances of just this one tiny DNA molecule coming into existence over four billion years... as *one chance* in 10^{415}.... This shows that life simply could not originate in outer space, period.
>
> [Yet] Dr. George Wald, Nobel Prize-winning biologist of Harvard University, stated several years ago: "One only has to contemplate the magnitude of [the] task to concede that spontaneous generation of a living organism is impossible. Yet here we are – as a result I believe, of spontaneous generation."

[This type of reasoning] shows how far even brilliant men will go to escape the idea of God being their Creator....

Next consider the following proevolutionary argument by Richard Dawkins (1987, pp. 45, 49):

[One in 10^{190}] is the chance against happening to hit upon haemoglobin by luck.... It is amazing that you can still read calculations like my haemoglobin calculation, used as though they constituted arguments *against* Darwin's theory. The people who do this, often expert in their own field, astronomy or whatever it may be, seem sincerely to believe that Darwinism explains living organization in terms of chance – "single-step selection" – alone. This belief, that Darwinian evolution is "random," is not merely false. It is the exact opposite of the truth. Chance is a minor ingredient in the Darwinian recipe, but the most important ingredient is cumulative selection which is quintessentially *non*random.

There is a big difference, then, between cumulative selection (in which each improvement, however slight, is used as a basis for future building), and single-step selection (in which each new "try" is a fresh one). If evolutionary progress had to rely on single-step selection, it would never have got anywhere. If, however, there was any way in which the necessary conditions for *cumulative* selection could have been set up by the blind forces of nature, strange and wonderful might have been the consequences. As a matter of fact that is exactly what happened on this planet.

Both the argument by Wilson and Weldon, and the argument by Dawkins can be unpacked as design inferences – for Wilson and Weldon as a successful design inference, for Dawkins as a failed design inference. A design inference attempts to establish whether an event is due to design. If we now take the event in question to be the occurrence of life on planet Earth, and denote this event by LIFE, then the design inference assumes the following form (cf. Section 2.2):

Premise 1: LIFE has occurred.
Premise 2: LIFE is specified.
Premise 3: If LIFE is due to chance, then LIFE has small probability.
Premise 4: Specified events of small probability do not occur by chance.
Premise 5: LIFE is not due to a regularity.
Premise 6: LIFE is due to regularity, chance, or design.
Conclusion: LIFE is due to design.

Since Wilson and Weldon, as well as Dawkins are attempting to explain LIFE, let us consider how their respective arguments conform to this pattern of inference. First Dawkins. Dawkins resolutely refuses to countenance design as a proper mode of explanation for LIFE. Dawkins thus rejects the conclusion of the design inference. But since the design inference constitutes a valid logical argument, for Dawkins to reject the conclusion he must reject at least one of the premises. But which one?

Let us run through the premises individually. Is Premise 1 a problem for Dawkins? Obviously not. LIFE has clearly occurred. What about Premise 2? Is LIFE specified? Dawkins (1987, p. 9) is quite definite about affirming this premise: "Complicated things have some quality, specifiable in advance, that is highly unlikely to have been acquired by random chance alone. In the case of living things, the quality that is specified in advance is . . . the ability to propagate genes in reproduction." So Premise 2 isn't a problem for Dawkins either. Indeed, no evolutionist or creationist I know denies that LIFE is specified.

Consider next Premise 4, the Law of Small Probability. Here too Dawkins finds nothing objectionable. Consider, for instance, the following remark (Dawkins, 1987, pp. 139, 145–6):

> We can accept a certain amount of luck in our explanations, but not too much. . . . In our theory of how we came to exist, we are allowed to postulate a certain ration of luck. This ration has, as its upper limit, the number of eligible planets in the universe. . . . We [therefore] have at our disposal, if we want to use it, odds of 1 in 100 billion billion as an upper limit (or 1 in however many available planets we think there are) to spend in our theory of the origin of life. This is the maximum amount of luck we are allowed to postulate in our theory. Suppose we want to suggest, for instance, that life began when both DNA and its protein-based replication machinery spontaneously chanced to come into existence. We can allow ourselves the luxury of such an extravagant theory, provided that the odds against this coincidence occurring on a planet do not exceed 100 billion billion to one.

Dawkins is restating the Law of Small Probability. LIFE is a specified event whose probability better not get too small. Thus, Premise 4 is not a problem for Dawkins either.

What about trichotomy, that LIFE is properly explained either by regularity, chance, or design? Here the very title of Dawkins's book – *The Blind Watchmaker* – makes clear that Premise 6 is not a problem either. Dawkins's title alludes to William Paley's (1802) famous watchmaker argument. For Dawkins, however, the watchmaker is blind, implying that the watchmaker is conditioned solely by regularity and chance. Along with most evolutionists, Dawkins holds that regularity and chance together are adequate to explain LIFE. Dawkins therefore holds to trichotomy, albeit a truncated trichotomy in which one of the disjuncts (i.e., design) is vacuous.

That leaves Premises 3 and 5. Dawkins appears to accept Premise 5 – that LIFE is not due to a regularity. All the same, because Dawkins never assigns an exact probability to LIFE, he never settles whether LIFE is a high probability event and thus could legitimately be attributed to a regularity. Dawkins seems mainly interested in showing that the occurrence of life on earth is probable enough, not in determining whether this probability is so close to unity to justify calling it a high probability event. Moreover, given the importance Dawkins attaches to probabilities, it appears that chance and contingency are essential to his understanding of LIFE. Thus, we have reason to think Dawkins accepts Premise 5.

This leaves Premise 3. Dawkins rejects Premise 3. Dawkins's appeal to cumulative selection makes this clear. According to Dawkins (1987, p. 49), "Chance is a minor ingredient in the Darwinian recipe, but the most important ingredient is cumulative selection which is quintessentially *non*random." The difference between cumulative selection and what Dawkins calls single-step selection can be illustrated with a coin tossing example. Suppose you want to know whether by tossing a hundred pennies you can ever expect to observe a hundred heads simultaneously. In the single-step selection scenario you put the pennies in a box, shake the box, and see if all the pennies simultaneously exhibit heads. If not, you keep repeating the process until all the pennies simultaneously exhibit heads. In the cumulative selection scenario, on the other hand, you shake the box as before, but every time you look inside the box, you remove the pennies that exhibit heads. You stop when all the pennies have been removed and are exhibiting heads (an example like this was treated more fully in Section 2.1).

Now it's clear that with the single-step selection scenario you will never observe all the pennies exhibiting heads – the odds are too much against it. On the other hand, it's equally clear that with the cumulative selection scenario you'll see all the pennies exhibiting heads very quickly. Dawkins's point, then, is that the natural processes responsible for LIFE act by cumulative selection, and therefore render LIFE reasonably probable. This he regards as the genius of Darwin, finding a naturalistic means for rendering probable what naively we take to be highly improbable. Dawkins therefore rejects Premise 3. Moreover, having rejected one of the premises in the design inference, Dawkins is under no obligation to draw the conclusion of the design inference, to wit, that the proper mode of explanation for LIFE is design. Indeed, Dawkins explicitly rejects this conclusion.

Where do Wilson and Weldon come down on the six premises of the design inference? Like Dawkins, Wilson and Weldon hold to Premises 1, 2, 4, and 6. Premises 1 and 6, though not explicitly stated, are clearly presupposed in their argument. Premises 2 and 4, on the other hand, are also presupposed, but imperfectly expressed since Wilson and Weldon do not have a well-developed notion of specification. Their version of the Law of Small Probability is Borel's Single Law of Chance (Wilson and Weldon, 1978, p. 321): "Emile Borel ... formulated a basic law of probability. It states that the occurrence of any event where the chances are beyond one in 10^{50} ... is an event which we can state with certainty will *never* happen – no matter how much time is allotted, no matter how many conceivable opportunities could exist for the event to take place."[9]

Wilson and Weldon are here employing a pretheoretic version of the Law of Small Probability, and one that omits specification. But since small probabilities have to combine with specifications to eliminate chance (exceedingly improbable unspecified events, after all, happen by chance all the time – see Chapter 1 and Section 2.1), a cleaned-up version of their argument would substitute Premises 2 and 4 for their pretheoretic version of the Law of Small Probability. Hence, as with Dawkins, we may regard Wilson and Weldon as affirming Premises 1, 2, 4, and 6.

But while Dawkins is not entirely clear about where he stands on Premise 5 and is perfectly clear about rejecting Premise 3, Wilson and Weldon accept both these premises. From the vast improbability of a certain DNA molecule, Wilson and Weldon (1978, p. 321) infer that LIFE is vastly more improbable still, and thus conclude that it is impossible for LIFE to originate anywhere in the universe by chance.

[9] Wilson and Weldon are citing Borel (1962, p. 28), the relevant portion of which was quoted in Section 1.1. In appealing to Borel, Wilson and Weldon are not misrepresenting him. Nevertheless, because Borel never developed an adequate conception of specification, his version of the Law of Small Probability is inadequate, and by implication so is Wilson and Weldon's. Cf. Borel (1962; 1963) and Knobloch (1990, p. 228).

For them LIFE is therefore neither the product of a regularity of nature nor an event with anything other than a very small probability. Besides Premises 1, 2, 4, and 6, Wilson and Weldon therefore accept Premises 3 and 5 as well. And having accepted the six premises of the design inference, by force of logic they conclude that the proper mode of explanation for LIFE is design. Note, however, that their identification of design with the activity of an intelligent agent – much less the God of Scripture – does not follow by the force of this logic.

These two arguments, the one by Wilson and Weldon, and the other by Dawkins, provide an object lesson for how design inferences arise in the creation–evolution controversy. The design inference constitutes a valid logical argument. Moreover, creationists and evolutionists alike tend not to controvert Premises 1, 2, 4, and 6. Thus, when creationists and evolutionists dispute the conclusion of the design inference, the dispute is over Premises 3 and 5: *If LIFE is due to chance, how improbable was it?* and *Is LIFE due to a regularity?* If Premises 3 and 5 both hold, then the conclusion of the design inference follows. If there is reason to doubt either of these premises, then the conclusion is blocked.

One thing, however, is clear. Creationists and evolutionists alike feel the force of the design inference. At some level they are all responding to it. This is true even of those who, unlike Dawkins, think LIFE is extremely unlikely to occur by chance in the known physical universe, but who nevertheless agree with Dawkins that LIFE is properly explained without reference to design. For instance, advocates of the Anthropic Principle like Barrow and Tipler (1986) posit an ensemble of universes so that LIFE, though highly improbable in our own little universe, is nevertheless virtually certain to have arisen at least once in the many, many universes that constitute the ensemble of which our universe is a member.

On this view LIFE is the winning of a grand lottery in which our universe happens to be the lucky ticket holder: The fact that our universe was lucky enough to beget LIFE is perhaps surprising, but no reason to look for explanations other than chance – much as a lottery winner, though no doubt surprised at winning, need not look for explanations other than chance since somebody (some universe) had to win. Our sense of surprise is due to a selection effect – that we should be so lucky. The relevant probability, however, is not the vast improbability that anyone in particular should win, but the extremely high probability that some (unspecified) lottery player would be sure to win. It's the high probability that someone will be selected that transforms what started as a seeming impossibility into a virtual certainty. Thus, whereas Dawkins rejects Premise 3 and offers the standard Darwinian mechanism as grounds for his rejection, Barrow and Tipler reject Premise 5, positing an ensemble of universes so that LIFE is sure to arise somewhere in this ensemble.

Positing an ensemble of universes isn't the only way to undercut Premise 5. Some theorists think our own little universe is quite enough to render LIFE not only probable, but virtually certain. Stuart Kauffman (1993, p. 340), for instance, identifies LIFE with "the emergence of self-reproducing systems of catalytic polymers, either peptides, RNA, or others." Adopting this approach, Kauffman (1993, p. 288) develops a mathematical model in which "autocatalytic polymer sets ... are expected to form spontaneously." Kauffman (1993, p. 287) attempts to lay the foundation for a theory of life's origin in which LIFE is not a lucky accident, but an event that is fully to be expected: "I believe [life] to be an expected, emergent, collective property of complex systems of polymer catalysts. Life, I suggest, 'crystallizes' in a phase transition leading to connected sequences of biochemical transformations by which polymers and simpler building blocks mutually catalyze their collective reproduction." Kauffman is not alone in explaining LIFE as a regularity of nature. Prigogine and Stengers (1984, pp. 84, 176), Wicken (1987), Brooks and Wiley (1988), and de Duve (1995) all share this same commitment.

To sum up, whereas creationists accept all six premises of the design inference, evolutionary biologists, to block the conclusion of the design inference, block Premises 3 and 5. Thus Darwin, to block Premises 3 and 5, had to give himself more time for variation and selection to take effect than many of his contemporaries were willing to grant (even though Lord Kelvin, the leading physicist in Darwin's day, estimated the age of the earth at 100 million years, Darwin regarded this age as too low for his theory). Thus Dawkins, to block Premises 3 and 5, and sustain his case for the blind watchmaker, not only gives himself all the time Darwin ever wanted, but also helps himself to all the conceivable planets that might exist in the known physical universe. Thus Barrow and Tipler, to block Premises 3 and 5, and give credence to their various anthropic principles, not only give themselves all the time and planets that Dawkins ever wanted, but also help themselves to a generous serving of universes (universes that are by definition causally inaccessible to us).[10] Thus Kauffman, to block Premises 3 and 5, and explain LIFE entirely in terms of natural processes operating on the earth (and hence without recourse to an ensemble of universes), invokes laws of self-organization whereby LIFE might arise spontaneously. From the perspective of the design inference, all these moves are moves against Premises 3 and 5, and therefore moves to block the conclusion of the design inference.

2.4 FROM DESIGN TO AGENCY

The logic of the Explanatory Filter is eliminative – to infer design is to eliminate regularity and chance. Yet in practice, to infer design is not simply to eliminate regularity and chance, but to detect the activity of an intelligent agent. Though defined as a negation, design delivers much more than a negation. Apparently lacking in causal pretensions, design sharply constrains our causal stories. There is an intimate connection between design and intelligent agency, a connection made clear by the Explanatory Filter. The aim of this section is to elucidate this connection.

To see why the filter is so well suited for recognizing intelligent agency, we need to understand what it is about intelligent agents that reveals their activity. The principal characteristic of intelligent agency is *directed contingency*, or what we call *choice*. Whenever an intelligent agent acts, it chooses from a range of competing possibilities. This is true not just of humans, but of animals as well as of extraterrestrial intelligences. A rat navigating a maze must choose whether to go right or left at various points in the maze. In trying to detect an extraterrestrial intelligence, SETI researchers assume such an intelligence could choose from a range of possible radio transmissions, and then attempt to match the observed transmissions with patterns regarded as sure indicators of intelligence. Whenever a human being utters meaningful speech, a choice is made from a range of possible sound combinations that might have been uttered. Intelligent agency always entails discrimination, choosing certain things and ruling out others.

[10] See Barrow and Tipler (1986), as well as critiques of the Anthropic Principle by van Inwagen (1993, ch. 8) and Leslie (1989).

Given this characterization of intelligent agency, the crucial question is how to recognize it. Intelligent agents act by making a choice. How then do we recognize that an intelligent agent has made a choice? A bottle of ink spills accidentally onto a sheet of paper; someone takes a fountain pen and writes a message on a sheet of paper. In both instances ink is applied to paper. In both instances one among an almost infinite set of possibilities is realized. In both instances a contingency is actualized and others are ruled out. Yet in one instance we ascribe agency, in the other chance. What is the relevant difference? Not only do we need to observe that a contingency was actualized, but we ourselves need also to be able to specify that contingency. The contingency must conform to an independently given pattern, and we must be able independently to formulate that pattern (cf. Sections 1.1 and 1.2 as well as Chapter 5). A random ink blot is unspecifiable; a message written with ink on paper is specifiable. Wittgenstein (1980, p. 1e) made the same point as follows: "We tend to take the speech of a Chinese for inarticulate gurgling. Someone who understands Chinese will recognize *language* in what he hears. Similarly I often cannot discern the *humanity* in man."

In hearing a Chinese utterance, someone who understands Chinese not only recognizes that one from a range of all possible utterances was actualized, but is also able to specify the utterance as coherent Chinese speech. Contrast this with someone who does not understand Chinese. In hearing a Chinese utterance, someone who does not understand Chinese also recognizes that one from a range of possible utterances was actualized, but this time, because lacking the ability to understand Chinese, is unable to specify the utterance as coherent speech. To someone who does not understand Chinese, the utterance will appear gibberish. Gibberish – the utterance of nonsense syllables uninterpretable within any natural language – always actualizes one utterance from the range of possible utterances. Nevertheless, gibberish, by corresponding to nothing we can understand in any language, also cannot be specified. As a result, gibberish is never taken for intelligent communication, but always for what Wittgenstein calls "inarticulate gurgling."

The actualization of one among several competing possibilities, the exclusion of the rest, and the specification of the possibility that was actualized encapsulate how we recognize intelligent agents. Actualization–Exclusion–Specification – this triad – provides a general scheme for recognizing intelligence, be it animal, human, or extraterrestrial. Actualization establishes that the possibility in question is the one that actually occurred. Exclusion establishes that there was genuine contingency (i.e., that there were other live possibilities, and that these were ruled out). Specification establishes that the actualized possibility conforms to a pattern given independently of its actualization.

Now where does choice, that defining characteristic of intelligent agency, figure into this criterion? The problem is that we never witness choice directly. Instead, we witness actualizations of contingency that might be the result of choice (i.e., directed contingency), but that also might be the result of chance (i.e., blind contingency). Now there is only one way to tell the difference – specification. Specification is the only means available to us for distinguishing choice from chance, directed contingency from blind contingency. Actualization and exclusion together guarantee we are dealing with contingency. Specification guarantees we are dealing with a directed contingency. The Actualization–Exclusion–Specification triad is therefore precisely what we need to identify choice and therewith intelligent agency.

Psychologists who study animal learning and behavior have known of the Actualization–Exclusion–Specification triad all along, albeit implicitly. For these psychologists – known as learning theorists – learning is discrimination (cf. Mazur, 1990; Schwartz, 1984). To learn a task an animal must acquire the ability to actualize behaviors suitable for the task as well as the ability to exclude behaviors unsuitable for the task. Moreover, for a psychologist to recognize that an animal has learned a task, it is necessary not only to observe the animal making the appropriate behavior, but also to specify this behavior. Thus, to recognize whether a rat has successfully learned how to traverse a maze, a psychologist must first specify the sequence of right and left turns that conducts the rat out of the maze. No doubt, a rat randomly wandering a maze also discriminates a sequence of right and left turns. But by randomly wandering the maze, the rat gives no indication that it can discriminate the appropriate sequence of right and left turns for exiting the maze. Consequently, the psychologist studying the rat will have no reason to think the rat has learned how to traverse the maze. Only if the rat executes the sequence of right and left turns specified by the psychologist will the psychologist recognize that the rat has learned how to traverse the maze. Now it is precisely the learned behaviors we regard as intelligent in animals. Hence it is no surprise that the same scheme for recognizing animal learning recurs for recognizing intelligent agents generally, to wit, actualization, exclusion, and specification.

This general scheme for recognizing intelligent agents is but a thinly disguised form of the Explanatory Filter: For the filter to eliminate regularity, one must establish that a multiplicity of possibilities is compatible with the given antecedent circumstance (recall that regularity admits only one possible consequence for the given antecedent circumstance; hence to eliminate regularity is to establish a multiplicity of possible consequences). Next, for the filter to eliminate chance, one must establish that the possibility actualized after the others were ruled out was also specified. So far the match between this general scheme for recognizing intelligent agency and how the Explanatory Filter infers design is exact. Only one loose end remains – the role of small probabilities. Although small probabilities figure prominently in the Explanatory Filter, their role in this general scheme for recognizing intelligent agency is not immediately apparent. In this scheme one among several competing possibilities is actualized, the rest are excluded, and the possibility that was actualized is specified. Where in this scheme are the small probabilities?

The answer is that they are there implicitly. To see this, consider again a rat traversing a maze, but this time take a very simple maze in which two right turns conduct the rat out of the maze. How will a psychologist studying the rat determine whether it has learned to exit the maze? Just putting the rat in the maze will not be enough. Because the maze is so simple, the rat could by chance just happen to take two right turns, and thereby exit the maze. The psychologist will therefore be uncertain whether the rat actually learned to exit this maze, or whether the rat just got lucky. But contrast this now with a complicated maze in which a rat must take just the right sequence of left and right turns to exit the maze. Suppose the rat must take 100 appropriate right and left turns, and that any mistake will prevent the rat from exiting the maze. A psychologist who sees the rat take no erroneous turns and in short order exit the maze will be convinced that the rat has indeed learned how to exit the maze, and that this was not dumb luck. With the simple maze there is a substantial probability that the rat will exit the maze by chance; with the complicated maze this is exceedingly improbable.

It's now clear why the Explanatory Filter is so well suited for recognizing intelligent agency: for the Explanatory Filter to infer design coincides with how we recognize intelligent agency generally. In general, to recognize intelligent agency we must establish that one from a range of competing possibilities was actualized, determine which possibilities were ruled out, and then specify the possibility that was actualized. What's more, the competing possibilities that were ruled out must be live possibilities, sufficiently numerous so that specifying the possibility that was actualized cannot be attributed to chance. In terms of probability, this just means that the possibility that was specified has small probability. All the elements in the general scheme for recognizing intelligent agency (i.e., Actualization–Exclusion–Specification) find their counterpart in the Explanatory Filter. It follows that the filter formalizes what we have been doing right along when we recognize intelligent agents. The Explanatory Filter pinpoints how we recognize intelligent agency.

3

Probability Theory

3.1 THE PROBABILITY OF AN EVENT

Our aim throughout the next four chapters is to explicate and justify the Law of Small Probability (LSP). To accomplish this aim, let us start by identifying the conception of probability we shall be using. Conceptions of probability abound. Typically they begin with a full theoretical apparatus determining the range of applicability as well as the interpretation of probabilities. In developing our conception of probability, I want to reverse this usual order, and instead of starting with a full theoretical apparatus, begin by asking what minimally we need in a conception of probability to make the design inference work.

One thing that becomes clear immediately is that we do not need a full-blown Bayesian conception of probability. Within the Bayesian conception propositions are assigned probabilities according to the degree of belief attached to them. Given propositions E and H, it makes sense within the Bayesian conception to assign probabilities to E and H individually (i.e., $\mathbf{P}(E)$ and $\mathbf{P}(H)$) as well as to assign conditional probabilities to E given H and to H given E (i.e., $\mathbf{P}(E \mid H)$ and $\mathbf{P}(H \mid E)$). If E denotes evidence and H denotes a hypothesis, then of particular interest for the Bayesian probabilist is how believing E affects belief in the hypothesis H. Bayes's theorem is said to answer this question, relating the probability of H given E (i.e., $\mathbf{P}(H \mid E)$, known as the posterior probability) to the probability E given H (i.e., $\mathbf{P}(E \mid H)$, known as the likelihood) and the probability of H by itself (i.e., $\mathbf{P}(H)$, known as the prior probability):

$$\mathbf{P}(H \mid E) = \frac{\mathbf{P}(E \mid H)\mathbf{P}(H)}{\mathbf{P}(E \mid H)\mathbf{P}(H) + \sum_j \mathbf{P}(E \mid H_j)\mathbf{P}(H_j)}.$$

Here the H_j are alternate hypotheses that together with H are mutually exclusive and exhaustive. Bayes's theorem is therefore of particular interest if one wants to understand the degree to which evidence confirms one hypothesis over another.

It follows that Bayes's theorem has little relevance to the design inference. Indeed, confirming hypotheses is precisely what the design inference does not do. The design inference is in the business of eliminating hypotheses, not confirming them. Given an event E and a chance hypothesis H, for the design inference to eliminate H, what needs to be established is that the probability of E given H (i.e., $\mathbf{P}(E \mid H)$) is small enough. On the other hand, what does not need to be established is the probability of H given E (i.e., $\mathbf{P}(H \mid E)$). Because the design inference is eliminative, there is no "design hypothesis" against which the relevant chance hypotheses compete, and which must then be compared within a Bayesian confirmation scheme. Thus, we shall never see a design hypothesis D pitted against a chance hypothesis H so that E confirms D better than H just in case $\mathbf{P}(D \mid E)$ is greater than $\mathbf{P}(H \mid E)$. This may constitute a "Bayesian design inference," but it is not the design inference stemming from the Explanatory Filter.

Of the three types of probabilities that appear in Bayes's theorem – posterior probabilities, likelihoods, and prior probabilities – only one is relevant to the design inference, namely, the likelihoods. The probability of an event given a chance hypothesis is the only type of probability we need to consider. Posterior probabilities and prior probabilities play no role. This I take to be a huge advantage of the design inference. Posterior probabilities can typically be established only via prior probabilities, and prior probabilities are often impossible to justify.

Only in special cases can prior probabilities be assigned with any degree of confidence (e.g., medical tests that test whether someone has contracted a given disease). Usually, however, prior probabilities cannot be assigned with confidence, and so one is left appealing to some variant of the indifference principle. Broadly speaking, the indifference principle asserts that given a set of mutually exclusive and exhaustive possibilities, the possibilities are to be treated as equiprobable unless there is reason to think otherwise. Unless properly nuanced, the indifference principle leads to paradoxes like the one of Bertrand (cf. Howson and Urbach, 1993, p. 60). And even when it avoids paradox, the indifference principle can lead to such ridiculous claims as the one by Laplace (see Zabell, 1988, p. 173), who, in assuming that the earth was about 5000 years old, claimed "it is a bet of 1826214 to one that [the sun] will rise again tomorrow."

These remarks about the Bayesian conception of probability apply equally to the logical and epistemic conceptions of probability (cf. respectively Keynes, 1921; Plantinga, 1993, chs. 8–9), and indeed to any conception of probability that in assigning probabilities to propositions (or statements), finds it important to move freely from a conditional probability like **P**(E | H) to its inverse probability **P**(H | E) (usually by means of Bayes's theorem) as well as to the unconditioned probabilities **P**(E) and **P**(H). This is not to say that the Bayesian or logical or epistemic conceptions of probability are incompatible with the design inference – the probabilities they assign work perfectly well with it. Nonetheless, these conceptions commit one to a much bigger (and more controversial) probabilistic apparatus than required by the design inference. All that the design inference requires of a probabilistic apparatus is that it assign probabilities to events.

What, then, is the probability of an event? To answer this question observe first of all that the probability of an event is never the probability of an event *simpliciter*, but always the probability of an event in relation to certain background information. Consider for instance the case of someone we'll call John. John, let us say, belongs to a local athletic club. What is the probability that John will show up at the athletic club Friday night to work out? The event in question, and the one whose probability we wish to determine, is the appearance of John at the athletic club Friday night. What probability shall we assign to this event? The answer clearly depends on the background information with which we approach this event. Change the background information and the probability changes as well. Here are a few possibilities:

(1) We never heard of John. We don't know who John is, nor the athletic club John attends. With this paucity of background information we can't even begin to assign a probability to John showing up at the athletic club Friday night.
(2) We know John well. We know that it is John's birthday on Friday, and that he never works out on his birthday because he prefers instead to carouse with his friends. We also know that the athletic club where John works out does not tolerate drunken revelry. It

therefore seems highly improbable that John will turn up at the athletic club Friday night.

(3) We know John well, but this time we know that John is a physical fitness addict. In fact, we know that for the past five years John has not missed a Friday night workout at the athletic club. We also know that John is in perfect health and has not been sick in years. It therefore seems quite probable that John will turn up at the athletic club Friday night.

(4) Finally, we know John well, and we know that John only works out three to four nights a week, favoring no day of the week over another. It therefore seems as probable as not that John will turn up at the athletic club Friday night.

The probability of an event can therefore be understood as the probability of an event E in relation to certain background information H. What sort of relation? The following definition provides the answer:

Definition. *The* **probability** *of an event E with respect to background information H, denoted by* ***P****(E | H) and called "the probability of E given H," is the best available estimate of how likely E is to occur under the assumption that H obtains.*

The remaining sections of this chapter will be devoted to unpacking this definition. Yet before we do this, it will be instructive to compare this definition with the propensity interpretation of probability. The propensity interpretation holds that probabilities are physically real properties of experimental arrangements (cf. von Mises, 1957, p. 14; Popper, 1959; and Giere, 1973, p. 473). Accordingly, probability is a relation between events and experimental arrangements. A given experimental arrangement has an inherent tendency (i.e., propensity) to produce a given event. If the experimental arrangement could be repeated indefinitely, this inherent tendency would express itself as a limiting relative frequency. Within a propensity interpretation this limiting relative frequency, capturing as it does the propensity of an experimental arrangement to produce an event, defines the probability of an event.

Although our definition of probability can accommodate the propensity interpretation, the converse fails. To accommodate the propensity interpretation it is enough to associate an experimental arrangement A with background information H that adequately describes A, and then set $\mathbf{P}(E \mid H)$ equal to the propensity-theoretic probability of E given A. On the other hand, accommodating our definition of probability within the propensity interpretation is not possible. The problem is that while an experimental arrangement is always fixed and comprehensive, background information is typically partial and fluid, subject to change in the light of further information.

Consider again the case of John showing up at the athletic club Friday night given that we know he's a physical fitness addict. Our background information leads us, rightly, to assign a high probability to this event. The actual experimental arrangement in which John finds himself, however, may be one where John has had a brain hemorrhage and is in no position to convey himself to the athletic club. Factoring in this new information would lead us, rightly, to lower the probability of John showing up at the athletic club. But the actual experimental arrangement might also include John's best friend, in a fervor of nostalgia, forming the intention to wheel John's unconscious body to the athletic club. Again, factoring in this new information would lead us, rightly, to increase the probability of John showing up at the athletic club.

Clearly, none of these probabilities assigned in the light of changing background information can legitimately be regarded as a propensity. But perhaps the problem is that we were being too restrictive with our background information. Is it possible to factor in enough information so that our background information comprehensively describes the experimental arrangement in which John finds himself? And if we could, would the probability we determined on the basis of this background information then constitute a propensity? Quantum mechanical systems that are simple, stylized, and irreducibly random may be amenable to this approach. But John showing up at the athletic club depends on a plethora of factors, including human intentionality. It's not at all clear that background information can capture all these factors, nor that the resulting probability would constitute a propensity. Even so, the decisive obstacle facing the propensity interpretation, and preventing it from accommodating our definition of probability, is that it has no way of assigning probabilities in the light of partial background information.

3.2 EVENTS

In unpacking our definition of probability, let us begin with events. An event is any actual or possible occurrence in space and time. Events can be relatively simple, as when a coin lands heads, or more complicated, as when a sperm and egg cell unite to form a zygote. Events can be historical, as in the dropping of the atomic bomb on Nagasaki in World War II. Events can be counterfactual, as in the dropping of the atomic bomb on Tokyo in World War II. Events are one-shot occurrences, tied to specific times, places, and things. This is not to say that an event cannot take a long to time to happen or might not be diffused through considerable swatches of space – e.g., the event of the solar system assuming its present form. It is, however, to say that events always possess specificity. Thus, a coin landing heads is an event only if we are considering a specific coin tossed at a specific time and place.[1]

Because we often apply Boolean operators to events, it is convenient to broaden our notion of event so that events are closed under Boolean operations. Thus, we define events as including not only one-shot occurrences, but also negations, conjunctions, and disjunctions of events. Thus, we may think of generic events as built up recursively from elementary events (i.e., the one-shot occurrences) via Boolean operations.

Often it will be convenient to treat events abstractly. Thus, we may speak of tossing a fair coin, without reference to any actual coin, place, or time. But such "abstract events" are simply classes of events grouped together because they share a common feature of interest. Thus, to speak abstractly of a coin landing heads, without reference to a specific coin, serves as a convenient way of talking about a general class of coin tosses. The probability of such an "abstract event" is the range of probabilities associated with all the individual events constituting it.

Events are not beliefs. My belief that I can successfully jump a twelve-foot chasm is not the same as the event that I successfully jump the chasm. What's more, my degree of belief that I can jump the chasm may be unrelated to the probability that I successfully jump the chasm if pursued, say, by a pack of bloodhounds. I may be fully convinced that the occult power I derive from yoga meditation

[1] Here I am following Borel (1963), for whom probability was the probability of a single event.

will enable me to jump the chasm successfully. But if I am grossly overweight and out of shape, the probability that I shall successfully cross the chasm will be small.

Finally, even though events are distinct from objects, they are closely related. Events typically presuppose objects and objects are typically produced by events. Consider, for instance, a dictionary resting on my bookshelf. I have just now adjusted it, changing its position. As a result, the dictionary has just now participated in an event, namely the changing of its position. Though presupposed in this event, the dictionary is not an event. On the other hand, the dictionary itself was produced by an event, most immediately, its printing and binding at a printer. It follows that even though we cannot, strictly speaking, submit an object to the Explanatory Filter and thus perform a design inference on it, nevertheless we can submit the event that produced the object to the Explanatory Filter. In particular, we can assign probabilities to objects by assigning probabilities to the events that produced the objects.

3.3 BACKGROUND INFORMATION

To assess the probability of an event E is to assess its probability relative to certain background information H. Such background information consists of items of information, which are simply claims about the world.[2] The background information H for assessing the probability of an event E decomposes into items of information $H_1, H_2, \ldots, H_n$ such that H is the conjunction of these items:

$$H = H_1 \,\&\, H_2 \,\&\, \cdots \,\&\, H_n$$

(the ampersand here represents the standard logical symbol for the conjunction "and"). As with events, negations, conjunctions, and disjunctions of items of information are also items of information. Thus, the background information H, as a conjunction of items of information, can itself be viewed as an item of information.

To illustrate how probability relates events and background information, consider the following example. Suppose we are given an

[2] Here I am following Stalnaker (1984, p. 65) in identifying "items of information" with propositions or claims. Ordinary usage sometimes restricts information to true claims about the world. In contrast, the background information relative to which probability assignments are made can be either factual or counterfactual.

event E that consists of Frank attending a certain party tomorrow. What are the chances that E will happen (i.e., that Frank will attend the party)? To assess the probability of this event, let us assume we are given the following items of information. H_1 – Frank is ambivalent about parties (he can take them or leave them). H_2 – Susan is going to be at the party, and Frank is crazy about Susan. H_3 – George, Frank's boss, is going to be at the party, and Frank would just as soon avoid him. H_4 – it's going to be a huge bash, so it's likely for Frank to get lost in the shuffle and avoid his boss. Given these four items of information, the probability of E needs to be assessed relative to the background information

$$H = H_1 \ \& \ H_2 \ \& \ H_3 \ \& \ H_4.$$

Moreover, given H, it seems reasonably probable that Frank will show up at the party: Frank doesn't actively dislike parties, Susan is going to be there, he's crazy about Susan, and his boss, whom he would rather avoid, can, it seems, be avoided.

An important fact to remember about background information is that probabilities can change drastically when either old items of information are deleted or new items of information are added. Suppose, for instance, we delete H_2 and H_4 from H. In this case the relevant background information becomes $H' = H_1 \ \& \ H_3$. All we know now is that Frank is ambivalent about parties and that his boss, whom he would rather avoid, is going to be there. Relative to H′, the probability of E happening seems quite small. On the other hand, we may augment our original H to include the additional item of information H_5, viz., that Frank has been in a serious car accident and is hospitalized lying in a coma. In this case the relevant background information becomes

$$H'' = H \ \& \ H_5 = H_1 \ \& \ H_2 \ \& \ H_3 \ \& \ H_4 \ \& \ H_5.$$

The addition of H_5 overwhelms our previous items of information. The probability of Frank showing up at the party tomorrow now becomes virtually nil.

The probability of an event E relative to background information H does not depend on whether the information in H actually obtains, or for that matter whether the event E has actually occurred. It is simply irrelevant to **P**(E | H) whether the information in H is accurate, verified, true, or assertible, or for that matter whether E has occurred.

Rather, what is important is how likely it would be for E to occur under the supposition that all the information in H obtains. The probability of an event E is therefore a conditional probability, conditional on the information H, with no presumption that the event or the information accurately reflects the actual world.

Finally, items of information are capable of describing events, but are not limited to events. Suppose, for instance, I wish to assign a probability to Ronnie washing my car. Suppose my background information includes (1) that I have asked Ronnie to wash my car, (2) that I have paid Ronnie $10 to wash my car, (3) that Ronnie is a conscientious soul who is eager to please people, and (4) that Ronnie views me as a cult figure and worships me. Given this background information it seems highly probable that Ronnie will wash my car. But notice, of these four items of information only the first two describe events – asking Ronnie to wash my car, and paying Ronnie to do it. In contrast, the third item describes a character trait of Ronnie's whereas the fourth describes a misguided belief. Every event can be described by a corresponding item of information, but some items of information do not describe events.

3.4 LIKELIHOOD

Having defined events and background information, we need next to define the relation between them. Specifically, given an event E and background information H we need to define how E and H together induce the probability $\mathbf{P}(E \mid H)$. To see what's at stake, consider the case of an English professor who for the past twenty years has been teaching the same course each year. Suppose each year the course enrollment has reached the maximum capacity of thirty students. Because the course is a college requirement, the professor is confident that in the upcoming year the course will once again attract thirty students. Now for a little diversion, each time the professor teaches this course she asks her students to give their birthdates. To the professor's surprise, in fifteen out of the twenty times she has taught the course, at least two students shared a birthday. The professor is not alone in being surprised. Indeed, most people find it unlikely that in a class of thirty students two will share a birthday.

Because our college professor teaches in the humanities, let us follow the popular stereotype of attributing to her little mathematical sophistication. Thus, we assume she lacks the statistical training for thoroughly analyzing this apparent anomaly. Nevertheless, let us suppose that even though this result seems to her counterintuitive, she decides it is more likely than not that in a class of thirty students at least one pair of students will share a birthday. In fact, given that in seventy-five percent of her classes the students shared a birthday, let us say she adopts a frequentist approach and assigns a probability of 0.75 to the event that the next time she teaches this course, her class will have a shared birthday.

Without any statistical or probabilistic training, the best our college professor can do is employ past relative frequencies to assign probabilities. We may therefore ask what probability she ought to assign if she possessed further mathematical training. William Feller, who possessed such training, proposed the following assignment of probability (Feller, 1968, p. 33 – the italics are his):

> The birthdays of r people form a sample of size r from the population of all days in the year. The years are not of equal length, and we know that the birth rates are not quite constant throughout the year. However, in a first approximation, we may take a random selection of people as equivalent to random selection of birthdays and consider the year as consisting of 365 days. [Given] these conventions ... [we conclude] *that the probability that all r birthdays are different equals*
>
> $$p = \frac{(365)_r}{365^r} = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{r-1}{365}\right).$$
>
> Again the numerical consequences are astounding. Thus for $r = 23$ people we have $p < 1/2$, that is, *for* 23 *people the probability that at least two people have a common birthday exceeds* $1/2$.

Because this result is counterintuitive, it is called the "birthday paradox." Since there are 365 days in the year, our first thought is that for two of r people to have a better than even chance of sharing a birthday, about half the number of days in the year will be required. Our first thought is to fix a given individual and then calculate the odds that the others share a birthday with this individual. The reason far fewer people will do is that we must consider all paired comparisons between individuals, and not simply comparisons to a fixed individual. Thus, in a class of 30 students there are 435 ways (>365 days in a year) of pairing the students, that is, ways the students can potentially share a birthday. It turns out, therefore, to be quite likely that in a class of thirty students at least two will share a birthday. In fact, according to Feller's formula, in a class of thirty students, the probability is 0.77 that at least two share a birthday.

Despite her lack of statistical training, our college professor was therefore quite close to the mark in assigning a probability of 0.75. Nevertheless, because her approach to probabilities relied exclusively on relative frequencies and because her sample size was small, she might easily have been much further from the mark. Given her training, our college professor did the best she could. On the other hand, Feller's probabilistic analysis represents a significant improvement. Even so, Feller's analysis is not the improvement to end all improvements. As Feller himself notes, his analysis can be further improved by factoring in leap years and differential birth rates according to time of year. And no doubt this analysis could be improved still further by taking into account yet other factors.

The lesson here is that the probability one assigns to an event depends on how effectively one is able to utilize background information and relate it back to the event. And this in turn depends on the training, experience, computational tools, and noetic faculties of the probabilist. The key word here is "effectively." Background information is like a mine whose resources need to be fully exploited. Given an event E and background information H, the probability **P**(E | H) never drops out for free. What H has to tell us about E needs to be as effectively utilized as possible. Given her lack of training in statistics, our English professor was limited in what she could extract from her background information. On the other hand, given his extensive training in statistics and probability, William Feller was able to make much better use of the same background information (though not perfect use).

The idea of effectively utilizing background information to assess probability is related to what statisticians call sufficiency. Sufficiency is all about making optimal use of one's data. Invariably the statistician tries to understand a population by collecting data from it (called "taking a sample"), and then drawing inferences from the data back to the population. Now the problem with data is that left to themselves they tell us nothing about how to infer back to the population. Suppose, for instance, we've taken a random sample, recorded 100 measurements, and now want to estimate the population mean. By themselves, these 100 measurements tell us nothing about the population mean.

To be of use, data have to be mined. If we suppose the 100 measurements are normally distributed, the optimal way to mine these data is by taking the sample mean: add all 100 measurements together and divide by 100. This is the sufficient statistic for the population mean (Mood et al., 1974, p. 313). As a sufficient statistic, it makes optimal use of these data. Of course, we can also make less than optimal use of these data. We can, for instance, ignore all but the smallest and largest measurement, add them together, and divide by 2. This statistic would also estimate the population mean, but would be grossly inefficient, ignoring most of the data. Moreover, it would introduce bias if the population were distributed asymmetrically about the population mean.[3]

The mining metaphor is worth taking seriously. A sufficient statistic is one that has mined the data so thoroughly that it can tell us nothing more about the parameter of interest. So too, background information may be mined so thoroughly that it can tell us nothing more about the event in question. Working a mine till it is exhausted is of course an ideal, and one we rarely attain. Nevertheless, it is an ideal to which we aspire, and one that regulates our assignment of probabilities. Just as in statistics sufficiency is the regulative ideal relating data to population parameters, so for probability *likelihood* is the regulative ideal relating background information to events:

Definition. *The* **likelihood** *of an event E with respect to background information H, denoted by* $\Lambda(E \mid H)$ *and called "the likelihood of E given H," is that number, or range of numbers,*[4] *in the unit interval* [0, 1] *denoting how likely E is to occur under the assumption that H obtains and upon the condition that H is as effectively utilized as possible.*

Implicit in this definition is that likelihoods are appropriately calibrated. There are two ways to calibrate likelihoods, both of which agree. The first is the frequentist approach. According to it, to say that the likelihood of an event E given H has value λ is to say that if an experimental set-up in which H obtains could be repeated indefinitely,

[3] For an introduction to sufficiency see Mood et al. (1974, pp. 299–314). For a full theoretical account see Le Cam (1986, ch. 5).

[4] Cf. Peter Walley's "imprecise probabilities" and Henry Kyburg's "interval-valued probabilities." See respectively Walley (1991) and Bogdan (1982).

so that the outcome of one experiment does not affect the outcomes of the others, then the relative frequency with which E occurs in these experiments has limiting value λ. This was the approach originally adopted by von Mises (1957), and one that continues to enjoy popularity. The second way of calibrating likelihood is the classical approach. According to it, to say that the likelihood of an event E given H has value λ is to say E is as likely to occur as randomly selecting one of m designated possibilities from a total of N possibilities, where (1) $0 \leq m \leq N$, (2) all the possibilities are symmetric, and (3) λ equals m/N. This second approach, though apparently limited to likelihoods that are rational numbers, is perfectly adequate for all practical purposes since our precision in measurement is always limited. Moreover, if we take mathematical limits of rational likelihoods, we can attain any likelihood in the unit interval.

Both approaches employ idealizations. With the frequentist approach the idealization consists in an indefinite repetition of an experiment under independent and identical conditions, as well as the supposition that in such an indefinite repetition relative frequencies will converge to a limiting value. With the classical approach the idealization consists in an abstract mechanism from which one may randomly sample among N distinct symmetric possibilities. Examples of the latter include flipping a rigid, homogeneous disk with distinguishable sides; throwing a rigid, homogeneous cube with distinguishable faces; selecting a ball from an urn that has been thoroughly mixed and whose contents consist of balls all having the same size, shape, rigidity, and homogeneity, though differing in color; and shuffling a pack of cards each of which are identical save for the pattern printed on the card. "Random sampling" and "symmetric possibilities" are irreducible concepts within the classical approach.

In practice, both approaches for calibrating likelihood work in tandem. The abstract mechanisms of the classical approach are readily approximated by real mechanisms like coins, dice, cards, and urns containing balls. Repeated random sampling from these real mechanisms typically yields relative frequencies that closely approximate the likelihoods specified by the classical approach. In this way the two approaches coalesce. Since my aim is not to explicate the theoretical underpinnings of the frequentist and classical approaches, but merely to indicate the rationale for why likelihoods assume certain values as opposed to others (e.g., why 1/13 as opposed to 1/14 is the likelihood of drawing an ace from an ordinary deck of playing cards), I'll leave the discussion here. The theoretical underpinnings of these approaches are taken up in the works of Keynes (1921 [1988]), von Mises (1957), Reichenbach (1949), and Laplace (1814).

A likelihood can be a single number between 0 and 1, a range of numbers between 0 and 1, or be undefined. Given as background knowledge H the claim that Ronnie washed my car, and given the event E that Ronnie did indeed wash my car, $\Lambda(E \mid H)$ is the likelihood of a tautology and therefore equals 1, a single number. On the other hand, given as background knowledge that I shall be randomly sampling a ball from either of two urns where one urn contains a single white ball and three red balls and the other contains three white balls and a single red balls (all the balls are identical save for color), and given that I know nothing about which urn I shall be sampling (i.e., my background information includes nothing for assessing prior probabilities), the likelihood $\Lambda(E \mid H)$ of drawing a red ball (= E) given my background information (= H) is the interval [1/4, 3/4], i.e., all numbers between 1/4 and 3/4. Finally, let E = *Ronnie washed my car* and H = *All the theorems of arithmetic.* Because H has no relevance to E, $\Lambda(E \mid H)$ is undefined.[5]

Now that we know what the numbers mean, let's return to the definition of likelihood and consider what it is for background information to be as effectively utilized as possible. At issue here is not simply the effective, but the maximally effective use of background information. What's more, except in the simplest cases the maximally effective use of background information is beyond human capacities. The question therefore arises just what capacities would facilitate the maximally effective use of background information. Complete knowledge of all mathematical truths (including all the theorems of statistics and probability), complete knowledge of the causal structure of the universe, complete knowledge of every possible event as well as the conditions under which it can occur, unlimited computational power, and the ability to reason without error should do it. A being with these capacities – call it an ideal rational agent – regardless whether it exists, and regardless what form it takes (whether a transcendent god, a Peircean ideal

[5] In these examples the relation between events and background information was sufficiently simple that a definite determination of likelihoods was possible. In general, likelihoods can only be estimated (cf. Section 3.5).

community, or a Kantian regulative ideal), could make maximally effective use of background information in determining likelihoods.

The mere possibility of such ideal rational agents renders likelihood an objective relation between events and background information. To see this, consider all ideal rational agents across all possible worlds. If for an event E and background information H all these agents agree in their assignment of likelihood, then $\Lambda(E \mid H)$ is univocally defined across all possible worlds. If on the other hand they disagree, define $\Lambda(E \mid H)$ as the smallest subset of the unit interval [0, 1] that includes all the likelihoods proposed by the ideal rational agents. This too defines $\Lambda(E \mid H)$ univocally across all possible worlds. Finally, if at least one ideal rational agent decides the likelihood of E given H is undefined, then take $\Lambda(E \mid H)$ as undefined. This too assigns $\Lambda(E \mid H)$ a univocal sense across all possible worlds. Holding univocally across all possible worlds, the likelihood relation is therefore independent of what we know or believe, and therefore objective. Moreover, since there are many likelihoods we can evaluate precisely and which assume point values, it follows this relation is nontrivial (most statistical probabilities are likelihoods and assume point values – see Plantinga, 1993, pp. 139–41).

3.5 THE BEST AVAILABLE ESTIMATE

Just because likelihoods are objective does not mean we can determine what they are. Objectivity and intelligibility are always separate issues. We see this preeminently in mathematics, which consists entirely of objective truths, only some of which we can know to be true. So too, we can determine only some likelihoods exactly. Typically, though, the best we can do is determine likelihoods approximately. This calls for estimates.

To estimate a likelihood is to assign an event a number between 0 and 1 (or a range of numbers between 0 and 1) that, in the light of relevant background information, reflects how likely the event is to occur. All of us make such estimates of likelihood all the time (though frequently we don't make explicit numerical assignments). Given what I know about myself, I judge it highly unlikely (i.e., I estimate the likelihood close to 0) that I'll play the upcoming state lottery, much less win it; I judge it moderately likely (i.e., I estimate the likelihood better than a half) that I'll get a haircut within the next three weeks; and I judge it highly likely (i.e., I estimate the likelihood very close to 1) that I won't live to see the twenty-second century. Such estimates of likelihood we call probabilities. Probabilities are not only commonplace, but also indispensable for navigating through life's uncertainties.

Because probabilities attempt to bridle uncertainty, the claim is sometimes made that probability is a measure of ignorance. This characterization of probability is misleading. Estimating the likelihood of an event in the light of background information presupposes a grasp of (1) the relevant background information, (2) the range of events compatible with this background information, and (3) the art of handling probabilities. In other words, we need to know quite a bit to estimate likelihoods. It seems, therefore, inappropriate to describe what we are doing when we estimate likelihoods as "measuring ignorance." Ignorance typically connotes a lack of knowledge that can be remedied, and should have been remedied, but wasn't because of negligence. Estimating likelihoods in the light of relevant background information is nothing of the sort. When estimating likelihoods, we have typically remedied as much ignorance as we are able. What uncertainty remains is due not to negligence, but to the difficulties inherent in ruling out alternate possibilities.

Although we make estimates of likelihood all the time, our estimates can also be wrong. Wrong in what sense? Wrong not in the sense of diverging from the true likelihood (estimation always tolerates some degree of error), but wrong in the sense that given what we know about the art of handling uncertainty (statistical theory, causal relations, contingency, etc.), we could have come up with a better estimate of likelihood. We go wrong in our estimates of likelihood for all the usual reasons we commit errors, such as sloppiness, dogmatism, and self-deception. But this is never an excuse. Anyone who goes to Las Vegas thinking he or she has a better than even chance of beating the casinos (short of a card counter, or someone who has a fix with the casino) ought to know better, and has made some terribly wrong estimates of likelihood.

Estimating likelihood is not something we do well instinctively. A striking example comes from the field of cognitive psychology. At one time psychologists Peterson and Beach (1967) claimed statistical theory provided a reasonably accurate model for how people make decisions in the face of uncertainty. According to Peterson and

Beach, people left to their own devices come up with estimates of likelihood reasonably close to those of statisticians. To be sure, these estimates might need to be refined through more exact probabilistic analyses, but on the whole people's intuitive estimates of likelihood were thought not seriously to depart from the estimates of likelihood prescribed by statistical theory.

In a celebrated series of articles, Kahneman and Tversky (1972; 1973) (also Tversky and Kahneman, 1971; 1973) provided convincing contrary evidence that people left to their own pretheoretic devices are not "sloppy Bayesians" or "somewhat imperfect probability theorists." Rather, Kahneman and Tversky found that people demonstrate consistent biases in their departure from optimal statistical decision making. Their findings held even for trained statisticians when these were deprived of paper and pencil. Thus, even trained statisticians fail to estimate the likelihood of events according to statistical theory unless paper and pencil are in hand and statistical theory is consciously being applied to the problem at hand. The connection between statistical theory on the one hand, and the way people ordinarily reach decisions in the face of uncertainty on the other, was therefore found to be not even close.[6]

To appreciate how people can go wrong as "intuitive probabilists," consider the following case in which someone is asked to estimate a likelihood (these are the sorts of counterexamples Kahneman and Tversky used to refute Peterson and Beach):

Imagine you are at the doctor's office. You have recently had tests performed to check whether you have a certain disease – disease X. The tests have come back positive. You know that the false positive rate for the test is 4% and the false negative rate is 3%. That is to say, for those who don't have the disease, only 4% of the time does the test indicate the presence of the disease; whereas for the people who actually have the disease, only 3% of the time does the test fail to detect the disease. Hence for those who don't have the disease, the test is 96% accurate, whereas for those who have the disease the test is

[6] Likewise, Bayesian decision theorists (e.g., Earman, 1992; Howson and Urbach, 1993) avoid suggesting that we are intuitive Bayesians when left to our own devices. Instead, they produce Dutch Book arguments, demonstrating that we had better align our decision making with Bayesian principles since otherwise we'll lose money. Bayesianism not only prescribes how to be rational, but also ties rationality to that great arbiter of all things, the bottom line. Even Carnap (1955 [1970], p. 450) took this approach: "A man considers several possible ways for investing a certain amount of money. Then he can – in principle, at least – calculate the estimate of his gain for each possible way. To act rationally, he should then choose that way for which the estimated gain is highest."

97% accurate. Suppose now that you also happen to know that the disease is rare in the population at large – say that no more than 1% of the population has the disease. With this background information, and given that your test came back positive, what do you estimate is the likelihood that you actually have contracted disease X?

Presented with a case like this, most people estimate their likelihood of having contracted the disease as quite high, typically better than eighty or ninety percent. As it turns out, however, this estimate is unduly high. In fact, a simple (and in this case sound) application of Bayes's theorem shows that the chance of having contracted the disease is less than twenty percent. For if we let E denote the contracting of disease X, and let H denote the test coming back positive, and factor in what we've been given about false positives, false negatives, and the incidence of the disease within the general population, then the probability we are after is $\mathbf{P}(E \mid H)$, which according to Bayes's theorem is just

$$\begin{aligned}\mathbf{P}(E \mid H) &= \frac{\mathbf{P}(H \mid E)\mathbf{P}(E)}{\mathbf{P}(H \mid E)\mathbf{P}(E) + \mathbf{P}(H \mid \sim E)\mathbf{P}(\sim E)} \\ &= \frac{(0.97)(0.01)}{(0.97)(0.01) + (0.04)(0.99)} \\ &= 0.1968.\end{aligned}$$

What, then, makes an estimate a good estimate of likelihood? To answer this question recall what makes something a likelihood in the first place. Given an event E and background information H, for a number λ to be the likelihood of E given H means that λ denotes how likely E is to occur once H has been as effectively utilized as possible. An ideal rational agent omnicompetent in the art of handling uncertainties knows how to extract from H everything relevant for determining how likely E is to occur. How does this apply to our estimates of likelihood? Though less than ideal rational agents, we all know something about the art of handling uncertainty, some more than others. The key question, therefore, is whether we are making the best use of what we do know about handling uncertainty. Given perfect knowledge in the art of handling uncertainty, an ideal rational agent can make maximally effective use of H in assessing likelihood. Given our limited knowledge in the art of handling uncertainty, the question remains whether we are nonetheless utilizing H as effectively as we can in assessing likelihood.

Recall the case of the college professor in Section 3.4. Though without formal statistical training, by appealing to relative frequencies the college professor made the best estimate she could of the likelihood that her upcoming class would have a shared birthday. On the other hand, in virtue of his statistical training William Feller was able to make better use of the same background information, and thereby improve the college professor's estimate of likelihood. Yet given what the college professor and William Feller respectively knew about the art of handling uncertainties, both offered perfectly good estimates of likelihood. It is important to realize that neither estimate was subjective or relative in the sense of merely reflecting personal taste. As soon as their knowledge about the art of handling uncertainties was determined, what could count as a good estimate of likelihood was determined as well.

As estimates of likelihood, probabilities therefore fail to be objective in the unqualified sense of being completely independent of what we know or believe. On the other hand, probabilities are objective in the qualified sense that once we factor in what we know about the art of handling uncertainty, the probability of an event given background knowledge is determined, and not the further subject of human idiosyncrasy.[7] This qualified objectivity of probability implies also a normativity for probability. People with the same knowledge in the art of handling uncertainty confronted with the same event and relating it to the same background information should assign roughly the same probabilities.

Even so, there is no simple answer to the question what separates the good from the bad estimates of likelihood – those we want to dignify with the appellation "the probability of an event" from those we don't. There is no algorithm that for every event–information pair invariably outputs the right probability. Neither do any of the traditional conceptions of probability – whether subjective, logical,

[7] Though adopting a logical conception of probability, Keynes (1921, p. 35) fully appreciated this point: "The degree of probability, which it is rational of *us* to entertain, does not presume perfect logical insight, and is relative in part to the secondary propositions which we in fact know; and it is not dependent upon whether more perfect logical insight is or is not conceivable. It is the degree of probability to which those logical processes lead, of which our minds are capable.... If we do not take this view of probability, if we do not limit it in this way and make it, to this extent, relative to human powers, we are altogether adrift in the unknown; for we cannot ever know what degree of probability would be justified by the perceptions of logical relations which we are, and must always be, incapable of comprehending."

classical, or frequentist – adequately decide this question. The mathematical theory of probability is only about 350 years old. In that time it has been continually developing new methods and ideas for estimating likelihoods. The probability of an event is therefore not etched in stone, but subject to continual refinement over time. What's more, assigning a probability to estimate likelihood often requires a combination of methods and ideas.

Often but not always. A frequentist conception of probability is no doubt the right one for determining the failure rate of light bulbs, and more generally the failure rate of devices that wear out over time. Thus, to test the endurance of a brand of light bulbs, the technicians at Underwriter Laboratories will simply take a bunch of these light bulbs and run them continuously until they fail. The average rate of failure and the variability of failure rates will be recorded, and thereafter serve as baselines for the light bulbs. Such baselines are established within a frequentist conception of probability. The frequentist conception of probability, however, remains limited. Missionaries being fattened up by cannibals have little reason to think that the frequency with which they have been fed until now reliably gauges whether they will be fed tomorrow – indeed, every day that passes gives them more reason to think that tomorrow they themselves will furnish the main course.

Even assigning a probability to something as seemingly straightforward as tossing a coin requires a host of factors be taken into account, including both frequentist and classical conceptions of probability. For if evenly balanced coins on average showed experimentally a strong predisposition to land heads, we should not assign equal probabilities to heads and tails. Yet the fact that in an even number of tosses coins only infrequently land heads and tails equally often does not imply that our best estimate of likelihood should be other than 50–50. Recorded relative frequency confirms that the two faces of a coin should each be assigned a probability close to fifty percent. Stochastic independence among coin tosses and symmetry properties then settle the matter, fixing the probabilities at exactly fifty percent.

Moreover, the view that coin trajectories follow the laws of mechanics and can therefore, at least in principle, be predicted in advance (thereby for any given toss collapsing probabilities to either zero or one) has had to be modified in the light of our more recent understanding of nonlinear systems, of which the human body is a conspicuous example. Yes, once a coin is in the air, and we know its position and momentum, we can predict how the coin will land. But prior to its being tossed, physics is incapable of determining the precise impulse that a complex nonlinear system like a human being will impart to the coin.[8] Thus, we see that a variety of considerations come into play when assessing probability even for something as simple and stylized as flipping a balanced, rigid disk with distinguishable sides.

Nor are Bayesian conceptions of probability without their limitations. As we've seen, both in this section and in Section 3.1, Bayesian conceptions of probability invariably face the problem of how to assign prior probabilities. Only in special cases can prior probabilities be assigned with any degree of confidence (e.g., medical tests). So long as the priors remain suspect, so does any application of Bayes's theorem. On the other hand, when the priors are well-defined, Bayes's theorem works just fine, as does the Bayesian conception of probability. To sum up, then, there is no magic bullet for assigning probabilities.

To define probability as the best available estimate of likelihood is therefore to embrace fallibilism and progressivism as ineliminable aspects of probability. How we determine the best available estimate of the likelihood of an event is in the end a function of our accumulated knowledge for dealing with uncertainty. Such knowledge consists of norms and practices that are constantly being corrected and refined. Such knowledge arises within a community of discourse and reflects the practices of those judged by the community as expert in the estimation of likelihoods. As the best available estimate of the likelihood of an event in the light of relevant background information, the probability of an event is the estimate on which this group of experts agrees.[9] What gets assigned as the probability of an event is therefore historically contingent. I offer no algorithm for assigning probabilities nor theory for how probabilities change over time. For practical purposes it suffices that a community of discourse can settle on a fixed estimate of likelihood.

Each word in the phrase "the best available estimate" does important work in the definition of probability. "Estimate" signifies that an

[8] For nonlinear dynamics and chaos see Lasota and Mackey (1985) and Waldrop (1992).

[9] In describing who is qualified to assign probabilities, Alvin Plantinga (1993, p. 195) remarks, "While we may not be thinking of a veritable Mozart of probabilities, we are not thinking of your average probability duffer either. When it comes to the deliverances of reason, what counts is the best, or nearly the best, that human beings can do."

event is being assigned a number between 0 and 1 (or a range of numbers, or no number at all) which, in the light of relevant background information, reflects how likely it is for an event to occur. The expression "best available" ensures that the relevant community of discourse is taking full advantage of its accumulated knowledge for dealing with uncertainty (e.g., the latest advances in statistical theory). This means that assignments of probability are self-critical and that the knowledge employed to assign probabilities is constantly being revised, corrected, and improved. As estimates of likelihood by experts in a community of discourse, probabilities are always on the table, ready to be discussed, revised, or even trashed. Finally, the definite article "the" indicates that the relevant community of discourse agrees on the estimate, whether it be a single number or a range of numbers or no number at all. Within a community of discourse at a particular time probability is therefore uniquely determined.

3.6 AXIOMATIZATION OF PROBABILITY

Distinct conceptions of probability require distinct axiomatizations.[10] The conception of probability presented in this chapter requires an axiomatization whose probability measures are point-valued, finitely additive, partial functions that apply to sentences (an interval-valued generalization is straightforward). Here is the axiomatization: Let **S** be a collection of sentences that contains the sentence T (tautology), is closed under the unitary operator $\sim$ (negation) as well as under the binary operators & (conjunction) and $\vee$ (disjunction), and has a binary relation $\perp$ (disjointness). $\sim$, &, and $\vee$ satisfy the usual laws of sentential logic (e.g., de Morgan's laws, distributivity of & over $\vee$ and vice versa, and commutativity of & and $\vee$). We define two binary relations on **S** in terms of $\perp$: $A \Rightarrow B$ iff_{def} $A \perp (\sim B)$ and $A \Leftrightarrow B$ iff_{def} $A \Rightarrow B$ and $B \Rightarrow A$. Then $\perp$ satisfies the following conditions:

(1) $\perp$ is symmetric, that is, for every A and B in **S**, if $A \perp B$, then $B \perp A$.
(2) For every A in **S**, $A \perp (\sim A)$.
(3) For every A, B, and C in **S**, if $A \Rightarrow C$, then $(A \& B) \Rightarrow C$.
(4) For every A, B, and C in **S**, if $(A \vee B) \Rightarrow C$, then $A \Rightarrow C$.

[10] For examples of distinct axiomatizations of probability see Gudder (1988), Dubins and Savage (1976), and de Finetti (1974).

(5) For every A, B, and C in **S**, if $A \perp B$ and $C \Rightarrow B$, then $A \perp C$.
(6) It's not the case that $T \perp T$.
(7) For all A in **S**, $A \Rightarrow T$, $(A \vee {\sim}A) \Leftrightarrow T$, and $(A \;\&\; {\sim}A) \Leftrightarrow ({\sim}T)$.

In case $A \perp B$ we'll say A and B are either disjoint, perpendicular, or mutually exclusive; in case $A \Rightarrow B$ we'll say A entails B; and in case $A \Leftrightarrow B$ we'll say A and B are equivalent.

Given a set **S** satisfying these properties, a probability measure **P** on **S** is then by definition any function satisfying the following axioms:

(A1) **P** is a partial function[11] from the Cartesian product $\mathbf{S} \times \mathbf{S}$ into the unit interval [0, 1] (the value of **P** for any ordered pair (A, B) in its domain is denoted by $\mathbf{P}(A \mid B)$).
(A2) For any A and B in **S**, $\mathbf{P}(A \mid B)$ is defined only if B is not equivalent to ~T. Moreover, if B is not equivalent to ~T, and if B entails A, then $\mathbf{P}(A \mid B)$ is defined and equals 1.
(A3) For any A, A′, B, and B′ in **S**, if A is equivalent to A′, B is equivalent to B′, and $\mathbf{P}(A \mid B)$ is defined, then $\mathbf{P}(A' \mid B')$ is defined as well and equals $\mathbf{P}(A \mid B)$.
(A4) For any A, B, and C in **S**, if $A \perp B$ and $\mathbf{P}(A \mid C)$, $\mathbf{P}(B \mid C)$, and $\mathbf{P}(A \vee B \mid C)$ are all defined, then $\mathbf{P}(A \mid C) + \mathbf{P}(B \mid C) = \mathbf{P}(A \vee B \mid C)$.
(A5) For any A and B in **S**, if $\mathbf{P}(A \mid B)$, $\mathbf{P}(A \,\&\, B \mid T)$, and $\mathbf{P}(B \mid T)$ are all defined, then $\mathbf{P}(A \mid B) = \mathbf{P}(A \,\&\, B \mid T) / \mathbf{P}(B \mid T)$. If we define **P** also as a partial function on **S** that to each A in **S** assigns $\mathbf{P}(A)$ ($=_{def}$ $\mathbf{P}(A \mid T)$), then the last equation reads $\mathbf{P}(A \mid B) = \mathbf{P}(A \,\&\, B)/\mathbf{P}(B)$.

How does this formalism accommodate the probability of an event E given background information H (i.e., the definition of probability given in Section 3.1)? As we noted in Section 3.3, every event can be described by a corresponding item of information. If we now interpret the sentences in **S** as items of information, then some of these sentences describe events. Thus, the probability $\mathbf{P}(E \mid H)$ can be reinterpreted within the preceding axiomatization as the probability of the sentence in **S** describing E given the sentence in **S** describing H. Once the correspondence between sentences on the one hand, and

[11] Making probability measures partial rather than total functions allows for gaps in probability assignments to be built directly into our probabilistic formalism.

events and information on the other is established, our entire discussion about the probability of an event assimilates to the preceding axiomatization. Because it is natural to speak of the probability of an event conditional on information instead of the probability of one sentence conditional on another, and because there is little chance of confusion going from one way of speaking to the other, I shall use both ways of speaking freely and interchangeably.

Given a probability **P**(E | H), one needs to keep in mind that adding further background information to H can drastically affect this probability. Save for probabilities identically zero or one, the addition of novel information always threatens to overturn a previous assignment of probability. Indeed, if **P**(E | H) is strictly between 0 and 1 and if the underlying probability space is sufficiently fine-grained,[12] as is often the case, then it is possible to specify additional information I such that **P**(E | H & I) falls anywhere between 0 and 1. Moreover, additional information is always available to collapse probabilities strictly to 1 and 0 (whether the underlying probability space is fine-grained or not): for I = E & H, **P**(E | H & I) = 1 and for I = H & (~E), **P**(E | H & I) = 0 (we assume all the probabilities here are defined).

Because I shall want to connect the probability measures of this chapter with the complexity measures of the next chapter, it will be convenient to give a slight reformulation of the probability measures defined in axioms A1–A5. Thus, instead of defining the probability measure **P** simply as a partial function on the Cartesian product of sentences **S** × **S**, it will be convenient also to define **P** as a partial function on the Cartesian product **S** × *pow*(**S**), where *pow*(**S**) is the powerset of **S**, that is, the collection of all subsets of **S**. The correspondence between the two types of probability is now easily established: for **P** defined on **S** × **S**, and evaluated at (E, H) in **S** × **S**, the corresponding **P** defined on **S** × *pow*(**S**) is evaluated at (E, {H}). Thus, **P**(E | H) = **P**(E | {H}), where the probability measure on the left side of the equation is defined on **S** × **S**, and the probability measure on the right side is defined on **S** × *pow*(**S**).

[12] What I mean by "sufficiently fine-grained" is that for any A such that **P**(A) > 0 and any r such that **P**(A) > r > 0, there is a sentence A′ that entails A satisfying **P**(A′) = *r* (if we're dealing with a Boolean algebra of sets, A′ will be a subset of A). Many probability measures satisfy this condition (e.g., the uniform probability on the unit interval [0, 1]). The general class of these probability measures consists of the nonatomic probability measures on uncountable complete separable metric spaces (see Parthasarathy, 1967, pp. 53–5).

Alternatively, for **P** defined on **S** × *pow*(**S**), and evaluated at (E, **H**) in **S** × *pow*(**S**) where **H** = {H_1, H_2, . . . , H_n}, the corresponding **P** defined on **S** × **S** is evaluated at (E, H_1 & H_2 & ⋯ & H_n). Thus, **P**(E | **H**) = **P**(E | H_1 & H_2 & ⋯ & H_n), where the probability measure on the left side of the equation is defined on **S** × *pow*(**S**), and the probability measure on the right side is defined on **S** × **S**. To establish the correspondence when **H** contains infinitely many members of **S** we need to extend "&" to an infinitary conjunction. Moreover, when **H** is empty, **P**(E | **H**) = **P**(E | T) = **P**(E) (an empty conjunction is by convention a tautology). Whether defined on **S** × **S** or **S** × *pow*(**S**), these formulations of **P** are entirely equivalent. In particular, it is perfectly straightforward to reformulate axioms A1–A5 for probability measures defined on **S** × *pow*(**S**).

4

Complexity theory

4.1 THE COMPLEXITY OF A PROBLEM

Our aim remains to explicate and justify the Law of Small Probability. Two pillars undergird this law, one probabilistic, the other complexity–theoretic. In the last chapter we elucidated the probabilistic pillar. Here we elucidate the complexity–theoretic pillar. Complexity theory, like probability theory, is a theory of measurement.[1] Whereas probability theory measures the likelihood of an event, complexity theory measures the difficulty of a problem. Specifically, complexity theory measures how difficult it is to solve a problem Q given certain resources R. To see how complexity theory works in practice, let us examine the most active area of research currently within complexity theory, namely, computational complexity theory.

Computational complexity pervades every aspect of computer science. Whatever the computational problem, a programmer has to consider how the available computational resources (= R) contribute to solving the problem (= Q). If a problem is intractable, the programmer won't want to waste time trying to solve it. Intractability occurs either if the problem has no algorithm that allows it even in principle to be solved on a computer, or if all the algorithms that solve the problem consume so many computational resources as to make solving the problem impracticable (either by requiring too much time or memory). Programmers therefore have a vital stake in computational complexity theory. By definition computational complexity theory handles one task: inputting algorithms and outputting the computational resources needed to run those algorithms.

Stated in this way computational complexity theory appears to do little more than reckon the cost of doing computational business. Nevertheless, the very idea of reckoning such costs has had profound implications for computer science and mathematics. The

[1] For general theoretical accounts of what it means to measure something, see Torgerson (1958), Suppes and Zinnes (1963), and Coombs, Dawes, and Tversky (1970).

celebrated open problem whether P (the polynomial time algorithms) equals NP (the nondeterministic polynomial time algorithms) is just one problem formulated entirely within computational complexity theory. To show that P is contained in NP is straightforward. The difficult question is whether the conceivably broader class of algorithms NP in fact coincides with the known subclass P (P is significant because it comprises the only class of algorithms known to be computationally tractable). A solution to this problem has implications for everything from logic (is the satisfiability problem of the propositional calculus practicable?) to real-world optimization problems (can we in real-time solve the traveling salesperson problem?). Questions at the heart of probability (e.g., randomness) and about the nature of mathematical proof (e.g., interactive proof systems) have also been recast in terms of computational complexity theory.[2]

Thus, philosophers too, and not just mathematicians and computer scientists, have a stake in computational complexity theory (see Earman, 1986, ch. 8; Wimsatt, 1980). Nevertheless, because computational complexity theory operates with the restricted notion of a computational resource, the problems it can pose and the solutions it can give are limited to this type of resource. Now the resources relevant to the design inference certainly include computational resources, but also exceed them. For this reason I shall propose a general approach to complexity, and one that incorporates computational complexity as a special case. Together with probability theory this generalized approach to complexity undergirds the Law of Small Probability.

The main object of study within complexity theory is the complexity measure. Complexity measures are measures of difficulty, measuring how difficult it is to solve a problem using the resources given. Thus, complexity measures measure such things as cost, time, distance, work, or effort. In measuring how difficult it is to solve a problem using resources, complexity measures assess not the actual contribution the resources have already made toward solving the

[2] For a precise formulation of the problem whether P equals NP see Garey and Johnson (1979, ch. 2). Various complexity–theoretic approaches to randomness have been proposed. See for instance Kolmogorov (1965) and the more recent work of van Lambalgen (1989) for a space complexity approach; Goldreich, Goldwasser, and Micali (1986) for a time complexity approach; and Dembski (1991; in press) for a "pattern" complexity approach. Finally, for interactive proof systems see Balcazar, Diaz, and Gabarro (1990, ch. 11).

problem, but what remains to be done in solving the problem once the resources are in hand (e.g., with resources of \$100 in hand and a problem of acquiring a total of \$1000, the complexity of the problem is not the \$100 already available, but the \$900 yet needed).

The computational complexity measures from computer science are a special case of complexity measures. Computational complexity is measured in terms of either time (i.e., number of elementary computational steps per second) or space (i.e., size of memory, usually measured in bits or bytes) or some combination of the two. The more difficult a problem, the more time and space are required to run the algorithm that leads to a solution. Time and space are the programmer's resources for transacting computational business. The business of computation is solving computational problems. To stay in business the efficient programmer strives to minimize the expenditure of computational resources.

Complexity theory generalizes computational complexity theory. Complexity theory takes problem–resource pairs (Q, R), conceived now quite generally, and assesses to what degree the resources R contribute to solving the problem Q. Moreover, it does this by measuring what remains undone to solve Q once R is in hand. The easier/harder it is to solve Q using R, the smaller/larger the complexity of (Q, R). Thus, when the complexity of (Q, R) is minimal, one faces but minimal difficulty solving Q using R (e.g., when the resources already contain a complete solution to the problem). Conversely, when the complexity of (Q, R) is high, one faces considerable difficulty solving Q using R (e.g., when Q is insoluble on the basis of R).

Since complexity assesses the difficulty of solving a problem using the resources given, complexity is a binary relation between problems and resources. The following definition characterizes this relation:

Definition. *The* **complexity** *of a problem Q with respect to resources R, denoted by* $\varphi(Q \mid R)$[3] *and called "the complexity of Q* given *R," is the best available estimate of how difficult it is to solve Q under the assumption that R obtains.*

This definition closely parallels the definition of probability given in Section 3.1. Recall,

[3] On occasion I shall also use other Greek letters, especially ψ, ξ, and χ, to denote complexity measures.

Definition. *The* **probability** *of an event E with respect to background information H, denoted by* $P(E \mid H)$ *and called "the probability of E given H," is the best available estimate of how likely E is to occur under the assumption that H obtains.*

The similarity between complexity and probability is not purely formal. In navigating through life's uncertainties, we invariably need to do two things: (1) estimate how likely it is for certain events to happen and (2) estimate how difficult it is for us to solve problems arising from those events. Probability theory and complexity theory thus work in tandem.

Sometimes the connection between probability and complexity is so close that the two notions end up being mathematically equivalent. Consider, for instance, the case of a safecracker whose problem is to open a safe (cf. Section 2.1). Let us suppose the safe has a combination lock marked with a hundred numbers ranging from 00 to 99, and for which five turns in alternating directions are required to open the lock. As usual, we assume only one sequence of alternating turns can open the lock (e.g., 34-98-25-09-71). There are thus ten billion possible combinations, of which precisely one opens the lock.

The opening of the safe may now be treated from both a probabilistic and a complexity–theoretic point of view. From a probabilistic point of view we may ask how likely is it for the safecracker, given only one opportunity to try a possible combination, to open the combination lock. Let us assume the safecracker has no background information to narrow down the range of possible combinations.[4] Here the event in question is the opening of the lock by the safecracker, and the relevant background information, whatever else it may be, is such that it fails to narrow down the range of possible combinations. The relevant probability is therefore one in ten billion, or 10^{-10}.

[4] It's worth noting that safecrackers by and large do not have special background information that enables them to reduce the number of possibilities on a combination lock. In general, safecrackers open safes by cracking them – i.e., by actually breaking the mechanism – not by finessing their way past the safe's mechanism, as by listening to the fall of tumblers. As Gleick (1992, p. 189) notes, "the lore notwithstanding, the chief tools of successful safecrackers were crowbars and drills. Safes were cracked; holes were torn in their sides; handles and dials were torn off. When all else failed, safes were burned. The safeman used 'soup' – nitroglycerin." Without such tools (= additional resources), safecrackers have no advantage over the rest of us.

Alternatively, from a complexity–theoretic point of view we may ask how difficult is it for the safecracker to open the lock. Without tools to force the lock and without information to narrow down the possible combinations, the safecracker cannot judge whether a combination opens the safe except by checking it. As the maximum number of possible combinations the safecracker may have to check, 10^{10} therefore measures the difficulty of opening the safe. Since it is more convenient to take the logarithm of this number, and since the logarithm of 10^{10} is by definition the information inherent in an event of probability 10^{-10} (see Hamming, 1986; as well as Section 4.6), the likelihood of opening the safe and the complexity of opening the safe are mathematically equivalent.

Although probability and complexity can turn out to be equivalent notions, generally this is not the case. For instance, the probability of a massive solar flair erupting between ten and eleven o'clock tomorrow morning depends not at all on any problems we are capable of solving. Alternatively, sorting an array of numbers according to magnitude depends on the correctness of our sorting algorithms and our computational power for executing these algorithms, and not at all on probabilistic considerations.

Yet despite such differences, probability and complexity parallel each other point for point (cf. Chapter 3): *Event* corresponds to *problem*, *background information* to *resources*, and *likelihood* to *difficulty*. Just as *probability* assesses likelihood of events in the light of background information, so too *complexity* assesses difficulty of problems in the light of resources. Moreover, both definitions depend on a community of discourse producing *best available estimates*, of likelihood in the one case, of difficulty in the other. For the definitions of probability and complexity to make sense it is necessary that estimates, whether of likelihood or of difficulty, be properly calibrated. In the case of probability, classical and frequentist approaches to probability enabled us to calibrate numerical assignments of probability (cf. Section 3.4). In the case of complexity, the notion of a complexity bound enables us to calibrate numerical assignments of complexity (cf. Section 4.5). Just as likelihood constitutes an objective relation between events and background information, so too difficulty constitutes an objective relation between problems and resources. Just as good estimates of likelihood require knowledge in the art of handling uncertainty, so too good estimates of difficulty require knowledge in the art of problem-solving.

4.2 PROBLEMS AND RESOURCES

The next four sections unpack the definition of complexity. We begin with problems. Though philosophers devote considerable energy to investigating specific problems, philosophers seem to have devoted little attention to explicating the nature of problems per se. In Bertrand Russell's *Problems of Philosophy*, problems are not themselves taken to constitute a problem. Philosophers of science Larry Laudan (1977) and Imre Lakatos (1970) make problem-solving the centerpiece of their philosophy of science, but say little, except by way of example, about what problems are. Cognitive scientists seek to understand human problem-solving, yet when coming to terms with the nature of problems per se typically do not progress beyond claims like the following (Bourne, Dominowski, and Loftus, 1979, p. 232): "Problems come in all shapes and sizes but generally share the characteristic that the individual must discover what to do in order to achieve a goal." In dictionaries and encyclopedias of philosophy, "problem" taken by itself (as opposed to specific problems like "the problem of other minds" or "the mind–body problem") neither appears as an entry nor is cited in any index. Since so many philosophers resist the urge to explicate problems per se, I shall do the same, treating problems as irreducible and foundational.

Given a problem, we next define resources as whatever is given to solve the problem. In identifying the resources for solving a problem, we need to be careful to include the skills and capacities of the agent who will attempt to solve the problem. For instance, to prove a mathematical theorem requires not only paper and pencil, but also the skills of a mathematician. Moreover, the degree of mathematical skill possessed by the mathematician will greatly affect whether and with what facility a theorem gets proven. A world class number theorist can solve problems in number theory that I cannot hope to solve, solve others easily that I could solve only with great effort, and solve still others that I too can solve easily, but which someone without mathematical training cannot solve. In short, paper plus pencil plus the skills of a world class number theorist is a far more effective resource for solving problems in number theory than paper plus pencil plus the skills of a number-theoretic duffer like myself.

Agents engaged in problem-solving vary greatly. Often they are individual human beings (as in the safecracker example of Section 4.1). Many problems, however, require whole teams of human problem solvers. For instance, the 10,000-page proof of the classification theorem for finite simple groups is the solution to a problem requiring the joint effort of hundreds of mathematicians spanning several decades (see Gorenstein 1983; 1986). The relevant agent here was therefore not a lone super-genius mathematician, but rather the community of mathematicians known as finite group theorists. Nor are agents necessarily human. A honey bee performing a dance is an agent who by dancing solves the problem of informing other bees where nectar is located. So too, computers can be treated as agents that solve problems by employing algorithms.[5]

The resources of complexity theory and the background information of probability theory are interconvertible notions. Certainly it's always possible to treat background information as a resource for solving a problem. For instance, when a safecracker attempts to open the combination lock of a safe, if the safecracker also possesses information about what the first two numbers of the safe's combination are, this background information constitutes a resource that substantially decreases the difficulty of opening the lock. Conversely, it's always possible to use background information to describe the resources for solving a problem. For instance, even though a key resource helping me to write this monograph is a Macintosh SE/30 computer, the description of this resource also serves as background information. Thus, we can view resources and background information as interconvertible. This interconvertibility will be exploited throughout Chapters 5 and 6, where the resources for a complexity measure φ serve double-duty as background information for a probability measure $\mathbf{P}$.

[5] Treating computers as agents raises some interesting issues in the philosophy of mathematics. Tymoczko (1979), for instance, questioned whether Appel, Haken, and Koch's (1977) computer-assisted proof of the four color theorem was legitimate by charging that the computer, rather than assuming its ordinary role as a passive resource, was usurping the role of human agents in surveying the proof's validity (Tymoczko was here appealing to Wittgenstein's requirement that mathematical proofs be humanly surveyable – see Wittgenstein, 1983, pp. 95, 143 ff.). Detlefsen and Luker (1980, p. 804) disagreed, insisting that mathematicians employ all sorts of resources to assist them in proving theorems, among which the computer is unexceptional.

A problem's difficulty will appear greater than it actually is if the resources for solving the problem are used ineffectively. Resources can be ineffectively used in either of two ways – by underutilizing them or by wasting them. For example, if my resources comprise a 300-pound refrigerator and a pickup truck, and my problem is to transport the refrigerator from my house to my friend's house five miles away, then I had better use the pickup. To be sure, I could avoid using the pickup and try moving the refrigerator manually the five miles to my friend's house. But by refusing to use the pickup, I make my problem appear more difficult than it actually is. The defect here is underutilizing available resources. Alternatively, I can make full use of the pickup, but instead of driving directly with the refrigerator to my friend's house, decide first to drive to Niagara Falls and back. The defect this time is not underutilizing available resources, but wasting them. As with underutilizing available resources, wasting them makes my problem appear more difficult than it actually is.

Accurately assessing a problem's difficulty therefore requires avoiding the twin defects of underutilizing and wasting resources. Indeed, we can go so far as to say that resources are effectively utilized to the degree that these twin defects are avoided. Thus, to say that resources are as effectively utilized as possible is to say these twin defects have been entirely eliminated. This is of course an ideal, and one we rarely attain. Nevertheless, it is an ideal to which we aspire, and one that regulates our assessment of difficulty. Just as within probability theory likelihood is the regulative ideal relating background information to events, so within complexity theory difficulty is the regulative ideal relating resources to problems:

Definition. *The* **difficulty** *of a problem Q with respect to resources R, denoted by* $\Delta(Q \mid R)$ *and called "the difficulty of Q given R," is that number, or range of numbers, in the interval* $[0, \infty]$ *denoting how difficult Q is to solve under the assumption that R obtains and upon the condition that R is as effectively utilized as possible.*

In this definition $\Delta(Q \mid R)$ measures how difficult Q remains once R has been as effectively utilized as possible. Moreover, degree of difficulty increases with $\Delta(Q \mid R)$ so that $\Delta(Q \mid R) = 0$ indicates minimal difficulty and $\Delta(Q \mid R) = \infty$ maximal difficulty. Implicit in this definition is that degree of difficulty has been appropriately calibrated. This topic is taken up in Section 4.5.

As with likelihood, difficulty can assume a single value, a range of values, or remain undefined. How difficult is it to open a lock of which we know nothing about its construction? This question admits no answer, leaving difficulty undefined. How difficult is it to open a lock if all we know is that it is one of several models, some of which are more secure than others. In this case difficulty will span a range of values. How difficult is it to open a lock that is already open? Here there is no difficulty at all. The difficulty is zero, a single value.

Although evaluating difficulty requires utilizing resources as effectively as possible, typically we fall short of this ideal. Given human limitations, some underutilization and waste typically creep into our problem-solving. With less than perfect facility in the art of problem-solving, we tend to make less than maximally effective use of the resources we are given. Indeed, maximally effective utilization of resources presupposes perfect facility in the art of problem-solving. A being with this capacity – call it an ideal rational agent – regardless whether it exists, and regardless what form it takes, could evaluate difficulty precisely. Now the mere possibility of such ideal rational agents renders difficulty an objective relation between problems and resources. Justifying this claim is virtually identical to justifying the objectivity of likelihood in Section 3.4, only this time our ideal rational agents specialize in problem-solving as opposed to uncertainty (the one additional complication here is that difficulty needs to be appropriately calibrated – see Section 4.5).

Just because difficulty is objective does not mean we can evaluate it precisely. Typically, the best we can do is estimate it approximately. To estimate the difficulty of a problem Q given resources R is to assign a number between 0 and ∞ (or a range of numbers) that, given the resources R, reflects how difficult it is to solve Q, 0 indicating minimal difficulty and ∞ indicating maximal difficulty (often ∞ signifies impossibility). All of us make such estimates of difficulty all the time (though typically we don't make explicit numerical assignments). As with estimates of likelihood, estimates of difficulty are indispensable to navigating through life. The following examples illustrate just how widespread the practice of estimating difficulty is.

(1) When building, plumbing, or electrical contractors place bids with their clients, they estimate how difficult it will be to solve a problem (i.e., complete a job) given the resources at the contractor's disposal. The estimate of difficulty here corresponds to the cost of labor and materials.
(2) Computer scientists routinely estimate the difficulty of solving a computational problem by taking the currently best algorithm that solves the problem and determining the algorithm's running time. In this case the estimate of difficulty is the running time of an algorithm on a computer.
(3) The odds that odds-makers in Las Vegas assign to sporting events reflect the difficulty that one team is expected to encounter in defeating another. In this case probabilities are doing the work of estimating difficulty.
(4) The whole field of optimization within the industrial sciences can be construed as minimizing the difficulty of solving certain problems, thereby maximizing output and minimizing incurred costs. Within optimization theory, estimates of difficulty range over everything from straight monetary costs to distances a traveling salesperson has to cover.

Estimates of difficulty sometimes keep us from attempting to solve problems not because they are intractable, but because we deem them too time-consuming or tedious. Consider for instance the case of William Thurston, the premier mathematician in low-dimensional topology, who a few years back wouldn't check a supposed proof of the Poincaré Conjecture in dimension 3. The Poincaré Conjecture in dimension 3 has been a long-standing open problem in mathematics whose resolution the mathematical community eagerly awaits. Nevertheless, Thurston refused to examine the proposed proof in detail, his reason being not only that the methods used in the proof were in his opinion inadequate, but more importantly because he judged it would take him several months to work through the details of the proof. Thurston therefore left the problem of checking the proof to his students. As it turned out, Thurston's intuitions were correct – the supposed proof contained an unfixable error.

Although we make estimates of difficulty all the time, our estimates can also be wrong. Wrong in what sense? Wrong not in the sense of diverging from the true difficulty (estimation always tolerates some degree of error), but wrong in the sense that given what we know about the art of problem-solving, we could have come up with a better estimate of difficulty. Sometimes we go wrong by thinking a problem more difficult than it actually is, at other times by thinking a problem easier than it actually is. Cases of underestimating difficulty include the following:

(1) In the 1970s medical doctors thought that by the 1980s they would be able to produce an artificial heart that could serve patients almost as well as a fully functional natural heart. The problem of producing such an artificial heart has in fact proved much more difficult, with artificial hearts unable to keep patients alive for more than a matter of months.
(2) In the 1950s it was thought that with the advent of the computer, computational natural language translators were just around the corner. The hope in particular was that such translators could help us keep up with the Russians during the Sputnik craze. By the 1960s these hopes were effectively dashed. Since then work on natural language processing has continued, but with limited success, and nowhere near the full natural language translators that had been envisioned.
(3) Around the turn of the century the mathematician Hermann Minkowski regarded the four-color problem as a fairly easy problem whose solution had been delayed only because no one of his mathematical caliber had yet to attack the problem (Minkowski was never modest about his abilities). When Minkowski actually did set his sights on the problem, he found it much more difficult than he expected. Indeed, its solution had to wait another eighty years, and then was possible only with the aid of a computer. (See Appel et al., 1977 for the solution to the four-color problem and Reid, 1986 for a biographical sketch of Minkowski.)
(4) How difficult is it for inorganic matter to organize itself spontaneously and produce life? Until the last century the spontaneous generation of organisms (even multicellular organisms) was taken as a matter of course. Since then, however, beginning notably with

the work of Louis Pasteur, chemical evolution and the origin of life have come to appear much more difficult problems. To be sure, many scientists hold that some form of prebiotic evolution has occurred. But the simplistic view that flies simply pop into existence out of household dust has long been abandoned. (See Thaxton, Bradley, and Olsen, 1984 for a critical review of origin of life studies.)

On the other hand, cases of overestimating difficulty are quite common as well:

(1) A bank president regards a bank vault as impregnable to burglars. A software designer thinks the security measures she has implemented render an operating system safe from hackers. The proud owner of a new Porsche Turbo thinks the car's security system will make it difficult for thieves to steal. Nevertheless, in each instance a breach of security frequently proves far easier than was previously suspected.
(2) How difficult is it for a human being to fly? How difficult is it to put a person on the moon? How difficult is it to travel to Mars? In the decades preceding these accomplishments few thought it possible. Technology frequently outstrips our expectations. Technology can make the solution of problems far easier than we previously imagined.
(3) After World War I the allies placed such constraints on Germany (in terms of war reparations, limits on the size of the German army, and sanctions) as to make it difficult for Germany to recover and again threaten Europe. World War I was to be the war that ended all wars. As it turned out, Germany was soon enough back on its feet posing a threat to Europe.
(4) At the turn of the century the mathematician Gottlob Frege claimed to have derived arithmetic entirely from the laws of logic (Frege's system employed no specifically arithmetical axioms). Based as it was supposed on purely logical laws, Frege's system should have been logically consistent. In particular, no amount of deductive "inferencing" should have been able to derive a contradiction from Frege's system. Deriving a contradiction from Frege's system was supposed to be logically impossible, and hence infinitely difficult (complexity $= \infty$). Nevertheless,

> Bertrand Russell's paradox made short work of Frege's system, showing that a contradiction issued from it rather easily.[6]

What, then, makes an estimate a good estimate of difficulty? To answer this question recall how we defined difficulty. Given a problem Q and resources R, for a number δ to be the difficulty of Q given R means that δ denotes how difficult Q is to solve once R has been as effectively utilized as possible. An ideal rational agent omnicompetent in the art of problem-solving knows how to extract from R everything relevant for assessing Q's difficulty. How is this relevant to our own estimates of difficulty? Though less than ideal rational agents, we all know something about the art of problem-solving, some more than others. The key question, therefore, is whether we are making the best use of what we do know about problem-solving. Given perfect facility in the art of problem-solving, an ideal rational agent can make maximally effective use of R, and thereby determine difficulty exactly. Given our limited facility in the art of problem-solving, the question remains whether we are nonetheless utilizing R as effectively as we can in estimating difficulty.

Such estimates of difficulty are what we mean by complexity. It is important to realize that such estimates are neither subjective nor relative in the sense of merely reflecting personal taste. As soon as we factor in what we know about the art of problem-solving, what can count as a good estimate of difficulty is fully determined. As estimates of difficulty, complexities therefore fail to be objective in the unqualified sense of being completely independent of what we know or believe. On the other hand, complexities are objective in the qualified sense that once we factor in what we know about the art of problem-solving, complexities are determined, and not the further subject of human idiosyncrasy. This qualified objectivity of complexity implies also a normativity for complexity. People with the same knowledge in the art of problem-solving confronted with the same problem and relating it to the same resources should assign roughly the same complexities.

[6] As Frege (1985, p. 214) put it, "Hardly anything more unfortunate can befall a scientific writer than to have one of the foundations of his edifice shaken after the work is finished. This was the position I was placed in by a letter of Mr. Bertrand Russell, just when the printing of this volume was nearing its completion. It is a matter of my Axiom (V)." This remark appears in the appendix to volume II of Frege's *Grundgesetze der Arithmetik* (*The Basic Laws of Arithmetic*). Russell's paradox had demonstrated that inherent in Frege's system was a contradiction.

Even so, there is no simple answer to the question what separates the good from the bad estimates of difficulty – those we want to dignify with the appellation "the complexity of a problem" from those we don't. There is, for instance, no algorithm that for every problem–resource pair invariably outputs the right complexity. Neither is there a systematic approach to complexity that regulates how this question is to be answered across the board. What with complexity covering everything from computation to cooking, complexity is simply too broad a notion to yield complete systematization. Computational complexity theory, much less complexity theory generally, is less than fifty years old. In that time complexity theory has been continually developing new methods and ideas for estimating difficulty. The complexity of a problem is therefore not etched in stone, but subject to continual refinement over time.

Still, one guideline for preventing bad estimates of difficulty is worth keeping in mind. Namely, we must avoid artificially inflating our estimates of difficulty by permitting useless activity that wastes precious resources. A computer programmer, for instance, can purposely write inefficient code, introducing, say, a loop that iterates a million times, but has no effect on the output of the algorithm. Lawyers and accountants can pad their time sheets by spending far more time on the account of a client than they actually need to. Taxicab drivers can take a long, circuitous route between two points when a simple, direct route exists. In estimating difficulty we must conscientiously strive to eliminate all superfluous activity. Granted, this is not always clear or possible. If we are new in a city, we can be fooled by a dishonest taxicab driver, and thus think getting from point A to point B is more difficult than it actually is. Nevertheless, as soon as we discover a superfluous element in the solution of a problem, we eliminate it, and no longer factor it into our estimates of difficulty. Our estimates of difficulty need to avoid as much useless padding as possible.

To define complexity as the best available estimate of difficulty is therefore to embrace fallibilism and progressivism as ineliminable aspects of complexity. How we go about finding the best available estimate of the difficulty of a problem is in the end a function of our accumulated knowledge for dealing with a wide cross-section of related problems. Such knowledge consists of norms and practices that are constantly being corrected and refined. Such knowledge arises within a community of discourse and reflects the practices of those judged by the community as expert in the estimation of difficulty. As the best available estimate of the difficulty of a problem given certain resources, the complexity of a problem is the estimate on which this group of experts agrees. What gets assigned as the complexity of a problem is therefore historically contingent. I offer no algorithm for assigning complexities nor theory for how complexities change over time. For practical purposes it suffices that a community of discourse can settle on a fixed estimate of difficulty.

Each word in the phrase "the best available estimate" does important work in the definition of complexity. "Estimate" signifies that a problem is being assigned a number between 0 and ∞ (or a range of numbers, or no number at all) that, given certain resources, reflects how difficult it is to solve the problem. The expression "best available" ensures that the relevant community of discourse is taking full advantage of its accumulated knowledge for solving problems (e.g., the latest advances in computational complexity theory). This means that assignments of complexity are self-critical and that the knowledge employed to assign complexities is constantly being revised, corrected, and improved. As estimates of difficulty by experts in a community of discourse, complexities are always on the table, ready to be discussed, revised, or even trashed. Finally, the definite article "the" indicates that the relevant community of discourse agrees on the estimate, whether it be a single number or a range of numbers or no number at all. Within a community of discourse at a particular time complexity is therefore uniquely determined.

4.4 AXIOMATIZATION OF COMPLEXITY

In the definition of complexity, two questions still need to be answered. The first is how to axiomatize complexity theory, the second is what do the numerical values of complexity signify. We take up the first question here, the second in the next section. In any axiomatization of complexity theory the primary object of study is the complexity measure. To define a complexity measure φ we start with a nonempty set **S** whose elements will be referred to generically as *sentences*. **S** will typically comprise the sentences of a formal language. Nevertheless, as far as the formal apparatus of complexity is concerned, exactly what **S** is or how its elements are interpreted doesn't matter. All that is required is that **S** be a nonempty set. Thus, it can happen that the "sentences" on which a complexity measure is defined are propositions, possible worlds, elements from a Boolean or σ-algebra, geographical locations, or even the very problems to which estimates of difficulty are being assigned.

Whereas in Section 3.6 it was assumed that the sentences associated with probability measures always come with additional structures (i.e., negation, conjunction, and disjunction), no such additional structures are presupposed for the sentences associated with complexity measures. Nevertheless, the sentences associated with a complexity measure always have the option of supporting additional structures. The most common of these optional additional structures will be called negation and denoted by the symbol $\sim$, which will signify a function from **S** into itself such that for any A in **S**, $\sim A \neq A$. In case **S** is not a collection of sentences from a formal language, but, say, a Boolean algebra, $\sim$ may have the additional property of being an involution (i.e., for any A in **S**, $\sim\sim A = A$), as when $\sim$ denotes complementation of sets. In addition to $\sim$, **S** may support binary operators & and $\vee$, called respectively conjunction and disjunction, which map $\mathbf{S} \times \mathbf{S}$ into **S**. **S** may of course support still other structures.

The problem–resource pairs of **S** are next defined as any ordered pairs of the form (C, **A**) from the Cartesian product $\mathbf{S} \times pow(\mathbf{S})$ ($pow(\mathbf{S})$ = the powerset of **S**, i.e., the collection of all subsets of **S**). The sentence C therefore represents a problem Q whereas the collection of sentences **A** represents the resources R relevant for solving Q. A complexity measure on **S** is then a partial function whose domain comprises these problem–resource pairs of **S**. More specifically,

Definition. *A* **complexity measure** *on a collection of sentences S is any partial function φ from the Cartesian product $S \times pow(S)$ into the nonnegative extended reals* $[0, \infty]$ *that, so long as (C, A) is in the domain of φ, satisfies the following redundancy condition*:

$$\textit{For } C \in S \textit{ and } A \in pow(S), \textit{ if } C \in A, \textit{ then } \varphi(C \mid A) = 0.$$

Note that as a partial function, φ need not be defined everywhere on the Cartesian product $\mathbf{S} \times pow(\mathbf{S})$ (cf. conditional probabilities of the form $\mathbf{P}(C \mid A)$ which are undefined whenever $\mathbf{P}(A)$ equals 0). This is not to say that complexity measures can't be total functions, defined on the whole of **S** × *pow*(**S**) – in some instances they are, as with computational complexity measures. But often, as with information measures (which are really just probability measures in disguise – see Section 4.6), they will only be partial functions.

For C in **S** and **A** a subset of **S**, we write $\varphi(C \mid \mathbf{A})$ even though the standard way of expressing a function of two variables is $\varphi(C, \mathbf{A})$. We choose the notation $\varphi(C \mid \mathbf{A})$ by analogy with conditional probabilities, thinking of **A** as supplying the resources available for solving the problem C. Thus, we think of $\varphi(C \mid \mathbf{A})$ as calculating how difficult it is to solve C once **A** is given, with $\varphi(C \mid \mathbf{A}) = 0$ indicating minimal difficulty and $\varphi(C \mid \mathbf{A}) = \infty$ indicating maximal difficulty. When writing $\varphi(C \mid \mathbf{A})$, we shall speak of conditioning C on **A**. It is convenient to collect all complexity measures defined on a collection of sentences **S** into a set. Thus, we let *Comp*(**S**) denote the set of all complexity measures on **S**.[7]

For A and C in **S**, the complexity of the problem–resource pair (C, A) is next defined as $\varphi(C \mid \{A\})$. To simplify the notation we denote $\varphi(C \mid \{A\})$ also by $\varphi(C \mid A)$. $\varphi(C \mid A)$ then estimates how difficult it is to solve C from A alone. Notice that by the redundancy condition $\varphi(A \mid A) = 0$. The difficulty of establishing some sentence D apart from any additional resources can next be defined as $\varphi(D \mid \emptyset)$ ($\emptyset$ = null set). To simplify the notation we denote $\varphi(D \mid \emptyset)$ also by $\varphi(D)$.[8]

If φ is a complexity measure on a collection of sentences **S**, we refer to the pair (**S**, φ) as a *complexity system*. Associated with any complexity system (**S**, φ) is a subcollection $Ax_\varphi(\mathbf{S})$ of **S** called the *axioms* or *implicit resources* of (**S**, φ). $Ax_\varphi(\mathbf{S})$ is the collection of sentences in **S** that are immediate under φ – zero complexity is required to solve the problems represented by the axioms. Formally

$$Ax_\varphi(\mathbf{S}) =_{\text{def}} \{A \in \mathbf{S} : \varphi(A) = 0\}.$$

If **S** and φ are clear from the context, we'll denote $Ax_\varphi(\mathbf{S})$ simply by *Ax*.

[7] The reader who wants to compare this notion of complexity measure with the standard computational notion should consult Davis and Weyuker (1983, pt. 4). For a machine-independent approach to computational complexity, see also Blum (1967).

[8] Note that conditioning on the null-set for a complexity measure corresponds to conditioning on a tautology for a probability measure. This correspondence is not accidental. Probability measures, when reformulated as information measures, are a type of complexity measure. See Section 4.6.

How does the preceding formalism accommodate the complexity of a problem Q given resources R (i.e., the definition of complexity given in Section 4.1)? To apply this formalism to an actual problem–resource pair (Q, R), we need to interpret the sentences in **S** as descriptions of both the problem Q and the resources R. Thus, the complexity of the problem–resource pair (Q, R) is within the preceding formalism unpacked as the complexity of a sentence C in **S** describing Q given a collection of sentences **A** in **S** describing the resources in R. Once the correspondence between sentences on the one hand and problems and resources on the other is established, we can assimilate our treatment of complexity in Sections 4.1–4.3 to this formalism.[9]

Because it is more natural to speak of the complexity of a *problem* in relation to *resources* than the complexity of an *individual sentence* in relation to a *collection of sentences*, I shall prefer the former mode of expression. Moreover, for a problem–resource pair (Q, R) and a complexity measure φ I shall freely write $\varphi(Q \mid R)$, even though (Q, R) is not technically in the domain of definition of φ ($\varphi(Q \mid R)$ needing properly to be interpreted as the complexity of a sentence describing Q given a collection of sentences describing R). Correspondingly, for a sentence C and a collection of sentences **A** I shall freely refer to the pair (C, **A**) as a problem–resource pair, even though technically (C, **A**) is just an ordered pair consisting of sentences from a formal language.

If we now review the treatment of probability at the end of Section 3.6, we see that both probability measures and complexity measures are partial functions on a Cartesian product of sentences $\mathbf{S} \times pow(\mathbf{S})$. This common domain of definition for both probability and complexity measures will prove extremely useful in the next chapter. Because of this common domain of definition, sentences in **S** can alternately describe events, background information, problems, and resources. Thus in particular, it will be possible to let background information for a probability measure serve double-duty as resources for a complexity measure (a collection of sentences describing background information for a probability measure may equally well

[9] Implicit in this identification is that sentences that describe the same problems and resources yield the same assignments of complexity under substitution. For instance, if C and D are sentences both describing the problem Q, and if **A** and **B** are collections of sentences both describing the resources R, then for any complexity measure φ that assigns a complexity to the problem–resource pair (Q, R), $\varphi(C \mid \mathbf{A}) = \varphi(D \mid \mathbf{B})$.

describe resources for a complexity measure, and vice versa). Having sentences serve double-duty in this way will be essential to the account of specification given in Chapter 5.

4.5 CALIBRATION THROUGH COMPLEXITY BOUNDS

In the definition of complexity, one question still needs to be answered: What do the numerical values of complexity signify? So far all I have said about complexities is that they range between 0 and ∞, with 0 signifying minimal difficulty (typically no difficulty at all), ∞ signifying maximal difficulty (typically impossibility), and the bigger complexities signifying the more difficult problems. Thus, all I have done is attribute a certain ordinal property to complexities. What I have not done is explain the significance of the numerical values assigned by complexity measures.

Because complexity is so much more general a notion than probability (in Section 4.6 we'll see that probability is a special case of complexity), calibrating complexities is nowhere near as neat as calibrating probabilities. There is a universality to probability that simply does not hold for complexity. To say that an event has probability $1/2$ is to say that it is as likely to occur as it is for a fair coin to land heads. Probabilities are always comparable to a standard (e.g., coin tossing) and therefore always commensurable (cf. Section 3.4).

Nothing like this holds for complexities. To say that a problem has complexity 100 may mean that it requires 100 processing steps on a computer, or that it requires 100 bits of information to specify, or that it requires 100 dollars to procure, or what have you. None of these complexities is commensurable in any straightforward way. Within certain contexts it may, for instance, be possible to equate number of processing steps executed on a mainframe computer with a dollar amount measuring the cost to a computer facility for executing so-and-so-many computational steps. But such an equation will depend on the current state of computer technology, what it costs to maintain and run the mainframe computer, not to mention what a given computer facility thinks it can get away with charging its users. All of this is highly contextual and ad hoc. There is consequently no standard against which all complexity measures can be compared as there is for probability. The best we can hope for is certain limited pairwise comparisons between distinct complexity measures. The dual notions of tractability and intractability bounds provide the key to these comparisons.

Although a complexity measure φ estimates the difficulty of solving a problem Q given resources R, it does not tell us the degree of difficulty *up to* which Q can be solved by means of R, nor the degree of difficulty *beyond* which Q cannot be solved by means of R. The degree of difficulty up to which a problem can still be solved will be called a *tractability bound*, whereas the degree of difficulty beyond which a problem can no longer be solved will be called an *intractability bound*. Tractability and intractability bounds are always nonnegative extended real numbers (i.e., numbers in the extended real interval $[0, \infty]$). We denote tractability bounds with the Greek letter λ and intractability bounds with the Greek letter μ. Also, we refer to tractability and intractability bounds generically as *complexity bounds*. Given a complexity measure φ, a problem Q, resources R, a tractability bound λ, and an intractability bound μ, we say that the problem Q is *tractable* if $\varphi(Q \mid R) < \lambda$ and *intractable* if $\varphi(Q \mid R) > \mu$. Note that a tractability bound λ can never exceed its corresponding intractability bound μ, and that the two need not be equal.

To see what's at stake with tractability and intractability bounds, imagine that what currently is the top of the line supercomputer can perform its computations at the rate of 1 teraflop (i.e., a trillion floating point operations per second). Suppose further that we, as operators of this supercomputer, want to use it to solve various computational problems. If we now include this supercomputer among our resources R, we can then define a complexity measure φ that to each computational problem Q assigns the minimum number of floating point operations required on the supercomputer to solve Q, a number we denote by $\varphi(Q \mid R)$.

What are plausible tractability and intractability bounds for φ? A tractability bound characterizes the degree of difficulty up to which we are still able to solve a problem Q given resources R. Even if we owned the supercomputer outright and could use it any way we saw fit, it seems implausible that we should be able to spend more than a year of computation time on any given problem (presumably we purchased the supercomputer to meet a multiplicity of needs, and not just to solve a single problem). Since there are on the order of 10^8 seconds in a year and since our supercomputer can perform 10^{12} floating point operations per second, we therefore do not expect to be able to handle any computational problem requiring more than $10^8 \times 10^{12} = 10^{20}$ floating point operations. In this way $\lambda = 10^{20}$ becomes a tractability bound for φ.

On the other hand, we may choose to ignore practical limitations on the use of our supercomputer. Indeed, if we can devote one year on the supercomputer to a given problem, why not two? And if two, why not three? And if three, why not four? And if . . .? At some point, however, these questions must terminate. The duration of the universe, for instance, obviously presents an upper bound on the number of years our supercomputer, or any computer for that matter, can be devoted to a particular problem. Since an intractability bound characterizes the degree of difficulty beyond which we are no longer able to solve a problem Q given resources R, to play it safe we might imagine generation upon generation of computer scientist, each intent on obtaining a solution to the same problem, pursuing the solution to this problem on our supercomputer for, let us say, a billion billion years (which by all accounts exceeds the heat-death/collapse of the known physical universe). In this case, with 10^{18} (= a billion billion) years to perform computations, with less than 10^8 seconds per year, and with the ability to perform 10^{12} floating point operations per second, even in this science-fiction scenario our supercomputer cannot handle computational problems requiring more than $10^{18} \times 10^8 \times 10^{12} = 10^{38}$ floating point operations. In this way $\mu = 10^{38}$ becomes an intractability bound for φ.

Observe that the tractability bound $\lambda = 10^{20}$ is strictly less than the intractability bound $\mu = 10^{38}$. A gap like this between tractability and intractability bounds is common. If there is a practical possibility of running a supercomputer for one year devoted to one particular computational problem, then there is also a practical possibility of running it for two, three, or even ten years. Similarly, if a billion billion years is incredibly generous as a bound above which a supercomputer cannot even in principle be run, then so is a hundred-million billion years. Typically between a tractability bound and its corresponding intractability bound there lies a demilitarized zone in which levels of difficulty slightly bigger than the tractability bound will still characterize tractable problems, in which levels of difficulty slightly smaller than the intractability bound will still characterize intractable problems, and in which the levels of difficulty that remain between the two bounds do not clearly characterize tractable or intractable problems.[10]

It follows that neither tractability nor intractability bounds will in general be unique. To be sure, they may be unique, and they may even be equal. This would be the case, for instance, if our supercomputer had a built-in explosive device so that after precisely one year of continuous operation the computer was sure to blow up. In this case, any computation requiring less than a year would be tractable and any computation requiring more than a year would be intractable. Thus, in this instance λ and μ would be equal.

In general, however, tractability and intractability bounds for a given complexity measure will neither be equal nor uniquely specified. Nevertheless, since a tractability bound signifies a degree of difficulty up to which problems are still doable, and an intractability bound signifies a degree of difficulty beyond which problems are no longer doable, there is an obvious rule of thumb for selecting tractability and intractability bounds: take tractability bounds as large as possible and intractability bounds as small as possible. Although this rule of thumb doesn't uniquely specify tractability and intractability bounds, it does narrow the field considerably. Moreover, it has the effect of making our tractability and intractability bounds as informative as possible. Suppose that λ and λ' are both tractability bounds with $\lambda' < \lambda$. Then for any problem that λ' tells us is tractable, λ will tell us it is tractable as well, since $\varphi(Q \mid R) < \lambda'$ entails $\varphi(Q \mid R) < \lambda$. Similarly, suppose that μ and μ' are both tractability bounds with $\mu' > \mu$. Then for any problem that μ' tells us is intractable, μ will tell us it is intractable as well, since $\varphi(Q \mid R) > \mu'$ entails $\varphi(Q \mid R) > \mu$.

Tractability and intractability bounds are as much estimates of difficulty as are complexities, only this time the estimates are made not for individual problem–resource pairs, but for whole classes of them (i.e., the doable and nondoable problems). Everything said in

[10] Although complexity bounds cannot avoid the problem of vagueness, what vagueness they encounter is innocuous and does not undermine the utility of the predicates "(Q, R) is tractable" and "(Q, R) is intractable." As Bas van Fraassen (1980, p. 16) rightly observes, "a vague predicate is usable provided it has clear cases and clear counter-cases." "Tractable" and "intractable" are usable in van Fraassen's sense because complexity bounds can be so defined that complexities less than a tractability bound *clearly* signal problems that are solvable, and complexities greater than an intractability bound *clearly* signal problems that are not solvable. There is a vast literature on vagueness (cf. Goguen, 1968–69; Fine, 1975; Kaufmann, 1975; Klir and Folger, 1988; and McGee, 1991).

Section 4.3 about assigning numerical values of complexity therefore holds for assigning tractability and intractability bounds. In particular, they are assigned in a community of discourse and reflect the norms and practices of those in the community judged as expert in estimating the difficulty of problems.

Assigning appropriate tractability and intractability bounds to a complexity system (S, φ) depends especially on the types of resources available for solving problems. If φ, for instance, measures the number of elementary computational steps needed for an algorithm to solve a computational problem, and if one's resources are limited to computers available before 1960, then the relevant complexity bounds, call them λ_{1960} and μ_{1960}, will be relatively small (relatively few computational steps are required to exhaust our computational resources). If, on the other hand, φ has available as resources machines present in 1990, then the relevant complexity bounds, call them λ_{1990} and μ_{1990}, will be substantially larger (this time our computational resources enable us to execute far more computational steps). Thus, even for the same complexity measure φ, different types of resources issue in different complexity bounds (in this example technological advance marks a change in resource type).

The question remains, What does a complexity measure signify when it assigns a numerical estimate of difficulty? This question now admits a limited answer in terms of tractability and intractability bounds. Given a complexity measure φ with tractability and intractability bounds λ and μ respectively, and a problem Q that is to be solved with resources R; Q is respectively tractable, indeterminate, or intractable depending on whether $\varphi(Q \mid R)$ is strictly less than λ, between λ and μ, or strictly greater than μ. Complexity bounds therefore make it possible to compare distinct complexity measures according to which problems they render tractable, indeterminate, or intractable. This wraps up the definition of complexity given in Section 4.1. The next two sections examine two commonly occurring types of complexity measures.

4.6 INFORMATION MEASURES

Probability measures are disguised complexity measures. As they stand, however, probability measures are not complexity measures. Not only is the scaling off (complexity measures assume values in the extended real interval [0, ∞] whereas probability measures assume values in the unit interval [0, 1]), but also the directionality of these measures is reversed. To see this, recall the safecracker in Section 4.1 who tries to open a safe whose combination lock allows ten billion possible combinations. As before, we suppose the safe is sufficiently well constructed that the safecracker cannot do better than randomly twirl the dial on the safe's combination lock and hope by chance to hit the right combination. Under these circumstances it's clear that the difficulty that confronts the safecracker in opening the safe increases as the safecracker's likelihood of opening the safe by chance decreases. Thus, we see that complexity and probability covary in opposite directions: Probabilities closer to zero correspond to complexities closer to infinity, while probabilities closer to one correspond to complexities closer to zero (imagine a safe with only one or two possible combinations – the probability/likelihood of opening the safe now goes up to one whereas the complexity/difficulty of opening it goes down to zero).

Once we adjust for scaling and directionality, it becomes entirely straightforward to view probability measures as disguised complexity measures, with the disguise involving nothing more than a change in direction and scale. Indeed, any order-reversing one-to-one correspondence between [0, 1] and [0, ∞] mapping 0 to ∞ and 1 to 0 transforms a probability measure into a complexity measure (by an order-reversing function from reals to reals, I mean a function f such that for $a < b$, $f(a) > f(b)$). Thus, if f is such a correspondence and **P** is a probability measure, $f \circ \mathbf{P}$ will be a complexity measure (the little circle "o" between f and **P** signifies composition of functions).[11]

It's clear that such fs are not uniquely determined. Indeed, given the way I've characterized f, f need not even be continuous. Nevertheless, of all the possible order-reversing correspondences between the unit interval [0, 1] and the extended real interval [0, ∞] taking 0 to

[11] At the end of Section 3.6 we extended the definition of probability measures from $S \times S$ to $S \times pow(S)$. Hence, for the probability measure **P** and the one-to-one order-reversing map f that takes 0 to ∞ and 1 to 0, $f \circ \mathbf{P}$ has the right domain of definition for a complexity measure (complexity measures being defined on Cartesian products of the form $S \times pow(S)$). To see that $f \circ \mathbf{P}$ actually is a complexity measure, it remains to check that $f \circ \mathbf{P}$ satisfies the redundancy condition. Suppose therefore that $E \in S$, that $\mathbf{H} \subset S$ where $\mathbf{H} = \{H_1, H_2, \ldots, H_N\}$, and that E is one of the H_is, i.e., $E \in \mathbf{H}$. Then H_1 & H_2 & ⋯ & H_N entails E, and so long as $\mathbf{P}(E \mid \mathbf{H})$ is defined, $\mathbf{P}(E \mid \mathbf{H}) = \mathbf{P}(E \mid H_1 \,\&\, H_2 \,\&\, \cdots \,\&\, H_N) = 1$. But since f takes 1 to 0, it follows that $f \circ \mathbf{P}(E \mid \mathbf{H}) = 0$. $f \circ \mathbf{P}$ therefore satisfies the redundancy condition and is a complexity measure.

∞ and 1 to 0, one correspondence stands out, and this is the negative logarithm to the base 2, that is, $f = -\log_2$. This canonical method for transforming probability measures into complexity measures stems from Claude Shannon's theory of communication (see Shannon and Weaver, 1949), with the complexity measure $-\log_2 \circ \mathbf{P}$ used to measure the "Shannon information" crossing a communication channel.

Complexity measures that are probability measures transformed by $-\log_2$ will henceforth be referred to as *information measures*. Thus, given a probability measure $\mathbf{P}$ defined on a collection of sentences $\mathbf{S}$, $\varphi = -\log_2 \circ \mathbf{P}$ defines an information measure on $\mathbf{S}$. For simplicity we also write the composition of the two functions $-\log_2$ and $\mathbf{P}$ as $-\log_2 \mathbf{P}$ (thus $\varphi = -\log_2 \mathbf{P}$). If φ is an information measure on the collection of sentences $\mathbf{S}$, we refer to the complexity system $(\mathbf{S}, \varphi)$ as an *information system*. As we saw in Section 4.4, associated with any complexity system $(\mathbf{S}, \varphi)$ is a subcollection $Ax_\varphi(\mathbf{S})$ of $\mathbf{S}$, which we called the *axioms* or *implicit resources* of $(\mathbf{S}, \varphi)$. $Ax_\varphi(\mathbf{S})$ was there defined formally as the collection $\{A \in \mathbf{S} : \varphi(A) = 0\}$. Since $\varphi = -\log_2 \mathbf{P}$, in terms of the probability measure $\mathbf{P}$, $Ax_\varphi(\mathbf{S})$ equals $\{A \in \mathbf{S} : \mathbf{P}(A) = 1\}$ (since $-\log_2(1) = 0$). Any axiom A of $(\mathbf{S}, \varphi)$ therefore describes an event of probability 1.

From the vantage of Shannon information, the axioms of an information measure are otiose, corresponding as they do to events of probability 1. Events of probability 1 rule out no possibilities. Because events of probability 1 comprise the complete range of possibility, they are utterly uninformative about what did and did not happen. In the case of coin tossing, for instance, that either heads or tails will land is an event of probability 1. To be informed of this event – that after the coin was tossed, either heads or tails landed – is therefore to acquire no new information, since we presumably knew what the range of possibility was in the first place. This contrasts with RMS systems, whose axioms are not trivial (see the next two sections, but especially Section 4.8).

What advantage does an information measure φ have over its corresponding probability measure $\mathbf{P}$? And why apply $-\log_2$ to $\mathbf{P}$ instead of some other order reversing map that takes 0 to ∞ and 1 to 0? Since φ and $\mathbf{P}$ are interdefinable ($\varphi = -\log_2 \mathbf{P}$ and $\mathbf{P} = 2^{-\varphi}$), any complexity-theoretic analysis in terms of φ can be translated into a corresponding probabilistic analysis in terms of $\mathbf{P}$, and vice versa. Moreover, if our aim is simply to transform $\mathbf{P}$ into some complexity measure or other, any number of other transformations could serve us equally well (e.g., $f(x) = (1 - x)/x$, what is known as the inverse odds ratio).

The rationale for transforming probability measures by $-\log_2$ derives from communication theory. The most convenient way for communication theorists to measure information is in bits (this accounts for the base 2 in the logarithm). Any message sent across a communication channel can be viewed as a sequence of 0s and 1s. Characters from more complicated alphabets can be represented as strings of 0s and 1s (cf. the ASCII code, which uses strings comprising eight 0s and 1s to represent the characters on a typewriter), with whole words and sentences in turn represented as strings of such character strings. In this way all communication can be reduced to the transmission of sequences of 0s and 1s. Given this reduction, the obvious way for communication theorists to measure complexity is in number of bits (i.e., number of 0s and 1s) transmitted across a communication channel. Communication theorists refer to this measure of complexity as information.

The connection between information and probability now follows as soon as we let sequences of 0s and 1s represent events. Imagine, for instance, that a spy in the field wants to inform headquarters of the exact position of the enemy. There are any number of positions the enemy might have taken, each of which constitutes a possible event. The enemy, however, has assumed one concrete position which the spy must now represent in a sequence of 0s and 1s and transmit to headquarters. How long (i.e., complex) does this string of 0s and 1s have to be? The answer obviously depends on the number of possible positions the enemy might have assumed.

Suppose, for instance, the enemy has assumed 1 of 2^n possible positions. There are 2^n bit strings of length n (i.e., strings of the form 100110· · ·0110 of length n). Hence there is a one-to-one correspondence between bit strings of length n and the 2^n possible enemy positions, one of which has actually been assumed. Once such a correspondence is in place, to represent the actual enemy position, it is enough to point to the bit string of length n that corresponds to that position.

If we now suppose that each of the possible positions the enemy might assume is equiprobable (this assumption is surely unrealistic, but it helps to draw the connection between information and probability), then the probability of the enemy taking a given position is the reciprocal of the total number of possible positions the enemy could have assumed, in this case 2^{-n}. Since information as a measure of complexity is just the length of a bit string, and since n bits are required to represent any one of 2^n equiprobable events (in this case, enemy positions), it makes sense to extend the notion of information from bit strings to equiprobable events so that the information inherent in any of 2^n equiprobable events is by definition $n = -\log_2(2^{-n}) = -\log_2$(probability of any such event). The obvious further extension for arbitrary events A with respect to arbitrary probability measures **P** is to define the information inherent in A as $-\log_2 \mathbf{P}(A)$. Here, then, is the rationale for calling $\varphi = -\log_2 \mathbf{P}$ an information measure.

It follows that information measures are complexity measures in more than just name. Of course, as objects satisfying the definition of complexity measures given in Section 4.4, information measures are complexity measures on purely formal grounds. Information measures, however, also connect with our intuitive understanding of complexity. Complexity is always a measure of difficulty, assessing how difficult it is to do this given that you can do that. For information measures the relevant type of difficulty is length of bit strings, assessing how difficult is it to represent events by means of bit strings, difficulty being identified with length of bit strings.[12]

4.7 RMS MEASURES

Given a nonempty set **S**, an *RMS measure* on **S** is defined as any partial function φ from the Cartesian product $\mathbf{S} \times pow(\mathbf{S})$ into the

[12] In this discussion of information measures I seem inadvertently to have slipped into the probabilistic language of "events" and "background information," as opposed to sticking with the language appropriate to information measures, that is, the complexity–theoretic language of "problems" and "resources." Strictly speaking, probability measures are defined for event–information pairs whereas complexity measures are defined for problem–resource pairs. Since information measures are complexity measures, we may well ask what are the problem–resource pairs to which they are being applied. As for resources, these readily subsume background information. As for problems, they have been implicit throughout this discussion, since it was never events themselves that were considered, but rather events that needed to be represented by sequences of 0s and 1s – and representing an event in this way does constitute a problem. The probabilistic language of "events" and "background information" is innocuous in the case of information measures because the identification between the probability of an event and the information required to represent an event is perfectly straightforward.

nonnegative extended reals $[0, \infty]$ satisfying the following conditions:

(4.7.1) *Redundancy*: For $C \in \mathbf{S}$ and $\mathbf{A} \in pow(\mathbf{S})$, if $C \in \mathbf{A}$, then $\varphi(C \mid \mathbf{A}) = 0$.

(4.7.2) *Monotonicity*: For $C \in \mathbf{S}$ and $\mathbf{A}, \mathbf{B} \in pow(\mathbf{S})$, if $\mathbf{A} \subset \mathbf{B}$ then $\varphi(C \mid \mathbf{A}) \geq \varphi(C \mid \mathbf{B})$.

(4.7.3) *Subadditivity*: For $C, D \in \mathbf{S}$ and $\mathbf{A} \in pow(\mathbf{S})$, $\varphi(D \mid \mathbf{A}) \leq \varphi(C \mid \mathbf{A}) + \varphi(D \mid \mathbf{A} \cup \{C\})$.

These conditions need only hold where φ is defined. RMS is an acronym for Redundancy-Monotonicity-Subadditivity. As in the axiomatization of complexity given in Section 4.4, the elements of the set **S** on which an RMS measure is defined will be referred to generically as *sentences*.

An RMS measure is therefore a complexity measure satisfying two additional requirements, monotonicity and subadditivity (redundancy holds by definition for all complexity measures – see Section 4.4). For a collection of sentences **S** and an RMS measure φ, the ordered pair $(\mathbf{S}, \varphi)$ will be referred to as an *RMS system* (cf. complexity systems in Section 4.4). The collection of all RMS measures on a set of sentences **S** will be denoted by *RMS*(**S**) (cf. *Comp*(**S**) in Section 4.4). Note that a handy triangle inequality follows immediately from the monotonicity and subadditivity conditions:

(4.7.4) For all A, B, and C in **S**, $\varphi(C \mid A) \leq \varphi(B \mid A) + \varphi(C \mid B)$.

RMS measures can be thought to measure the *work* or *effort* that must be expended to solve a problem given a set of resources. Thus, for an RMS measure φ, $\varphi(C \mid \mathbf{A})$ can be interpreted as the amount of effort needed to solve C given the resources **A**.[13] For RMS measures the redundancy condition then indicates that if a problem has already been solved and its solution is among the given resources, no further effort need be expended to solve the problem. Hence (4.7.1). The monotonicity condition in turn indicates that the effort that must be expended to solve a problem decreases as the resources needed for solving the problem are increased. Thus, $\varphi(C \mid \mathbf{A})$ decreases as **A**

[13] In general, the elements of S will represent problems and resources rather than actually coinciding with them. Thus, technically it would be more accurate to interpret $\varphi(C \mid \mathbf{A})$ as the amount of effort needed to solve the problem represented by C given the resources represented by **A**. This is a niggling point.

increases. Hence (4.7.2). Complexity measures that satisfy condition (4.7.2) only will be called *monotonic complexity measures.*

Finally, the subadditivity condition indicates that the effort needed to solve a fixed problem from fixed resources cannot exceed the following the sum: (the effort needed to solve an intermediate problem from the initial resources) + (the effort needed to solve the original problem from initial resources together with the solution to the intermediate problem, which now serves as an additional resource). In the subadditivity condition, the intermediate problem serves double duty, once as a problem to be solved via the initial resources, and once as a solution (to the intermediate problem) that together with the initial resources helps solve the original problem.[14] Thus, comparing the most direct route for solving D with the indirect route of first solving C and then solving D using the solution to C – the initial resources **A** staying fixed – we expect the direct route to D to require less effort than the indirect route through C. Hence (4.7.3). Complexity measures that satisfy condition (4.7.3) only will be called *subadditive complexity measures*.

Although conditions (4.7.1)–(4.7.3) seem reasonable for modeling work and effort, conditions (4.7.1) and (4.7.2) are nonetheless open to criticism in this regard. The difficulty with these two conditions is that they seem to ignore the cost of utilizing resources that are so large as to be unwieldy. Indeed, the more resources the merrier is the upshot of conditions (4.7.1) and (4.7.2). Resources, however, need not always be an asset. One of the artificial intelligence community's biggest challenges is devising search strategies that distinguish relevant from irrelevant information (see VanLehn, 1989).[15]

Consider for instance a large database. If we treat the items of information it contains as resources, then according to conditions (4.7.1) and (4.7.2) this large database is resource rich, and is therefore better at reducing effort than smaller databases. On the other hand, if we treat size of the database as an obstacle in the sense that search

[14] For RMS measures, what is the solution to a problem in one instance can be a resource in another. This is especially true of real-world examples in which effort is expended: Thus, in one instance the problem may be constructing hammer and chisel from nondescript pieces of metal, whereas in another the problem may be sculpting a statue using that same hammer and chisel.

[15] Cf. Aristotle who in the opening of his *Metaphysics* asserted that knowledge is always a good thing. In an information age, where the sheer volume of information is spinning out of control, too much knowledge/information can actually be a bad thing.

and access times increase with size, then a large database requires greater expenditures of effort than smaller databases. This example becomes particularly poignant if we think of augmenting the database with a mound of trivial and useless information whereby most of the database becomes irrelevant to the actual information we are trying to derive.[16]

Against this criticism I would simply point out that search times can themselves be represented in terms of effort. Thus, in the database example we may have to consider two types of effort, one that treats the individual items of information in the database as assets, no matter how trivial or irrelevant, and another that treats the entire database as an obstacle, with the database becoming less and less tractable as its size increases. Depending on the purposes at hand we may focus on one form of effort in place of another. Alternatively, we may want to consider some combined measure of effort that balances the advantage of increased information with the disadvantage of having to search through it.

In contrast, condition (4.7.3), the subadditivity requirement, seems largely uncontroversial in modeling work or effort. Because this condition is stated as an inequality (i.e., the effort needed to solve a problem directly is less than or equal to the effort needed to solve a problem indirectly via the solution of some intermediate problem), it represents the common occurrence that costs diminish as production size increases. Imagine a printer who must print 1000 copies of a book. The subadditivity requirement implies that the cost of printing 1000 copies in one batch cannot exceed – and might actually be strictly less than – the cost of printing 500 copies twice in two batches. In fact, we know from experience that printing 500 copies twice will be more expensive than printing 1000 copies once. Of course, the subadditivity requirement also permits strict equality, as when solving an intermediate problem is the quickest way to solving a given problem.

Although the preceding account of RMS measures is new, specific examples of RMS measures are well-known and easily recognized. I want to review some of these examples now. These examples will demonstrate the scope and diversity of RMS measures. Here, then, is a brief catalogue:

[16] I'm indebted to Frank Döring for this line of criticism.

Example 4.7.1 (Distance). *Imagine sending a light signal from point A to point B (suppose A and B are at rest in some inertial frame). Then with a light source at A as the resource and getting to B as the problem, one example of an RMS measure relating the two is the distance separating A and B. In this way the metrics of point-set topology become RMS measures. Note that the triangle inequality* (4.7.4) (*i.e., that the direct distance between two points never exceeds the indirect distance through some intermediate point) follows immediately from the subadditivity condition. Note also that if the problem is getting a signal to B as quickly as possible from one of many sources (e.g., A, A′, A″, ...), then the relevant "measure of effort" is the distance from B of the source closest to B. Hence, consistent with the monotonicity condition, the more (re)sources, the less effort.*

More formally, we may consider a metric space (M, d). M is a nonempty set and d is a function on the twofold Cartesian product of M into the nonnegative extended reals $[0, \infty]$. *M and d satisfy the following conditions:*[17]

(4.7.5) *Nonnegativity:* $d(x, y) \geq 0$ *for all* x *and* y *in* M, *with* $d(x, y) = 0$ *if* $y = x$.

(4.7.6) *Symmetry:* $d(x, y) = d(y, x)$ *for all* x *and* y *in* M.

(4.7.7) *Triangle inequality:* $d(x, z) \leq d(x, y) + d(y, z)$ *for all* x, y, *and* z *in* M.

Now define the following function φ *on* $M \times pow(M)$ ($x \in M$ *and* $A \subset M$):

(4.7.8) $\varphi(x \mid A) = \inf\{d(x, y) \mid y \in A\}$ (*i.e., the* $d(x, A)$ *of point-set topologists*).

φ *measures the minimum distance from x to the set A (the nonmathematician might want to think of x as a ship at sea and A as land with* $\varphi(x \mid A)$ *measuring the nearest distance to land). It is immediate that* φ *satisfies conditions* (4.7.1)–(4.7.3) *for RMS measures.*

[17] Observe that I am presenting here a slightly more general notion of metric space than is customary. Normally the distance between any pair of points in a metric space is finite, i.e., normally d takes values in the nonnegative reals $[0, \infty)$ rather than in the nonnegative *extended* reals $[0, \infty]$. Also, the definition I give permits distinct points to have zero distance. What I'm calling a "metric" thus corresponds to what mathematicians call a "semimetric." See Wilansky (1983, p. 12).

Example 4.7.2 (Blocks Worlds). *Imagine an initial configuration of blocks that is to be rearranged into a certain final configuration. Arranging the blocks to fit the final configuration is the problem to be solved; the initial configuration constitutes the resources; and the complexity of solving the problem with the given resources is the effort qua work (in the sense of force × distance) needed to rearrange the blocks. If initial and final configurations are identical, no effort need be expended (cf. redundancy). If the final configuration contains more blocks than the initial configuration, no amount of rearranging is going to attain the final configuration (in this case effort becomes infinite). This example generalizes easily to the full notion of work as defined in classical physics.*

Example 4.7.3 (Computational Complexity Theory[18]). *Given a program P, the effort needed to run it is the number of elementary computational steps that must be executed before the program halts –* $\varphi(P) = n$ *for some natural number n. Given a program P that incorporates another program Q as a subroutine, we can ask how quickly P halts given that no costs are incurred every time P calls Q – the answer being given by* $\varphi(P \mid Q)$. *If P incorporates several subroutines, such as* $\{Q_1, Q_2, \ldots, Q_k\}$, *we can ask how quickly P halts given that no costs are incurred every time P calls any of the subroutines in* $\{Q_1, Q_2, \ldots, Q_k\}$ *– the answer being given by* $\varphi(P \mid \{Q_1, Q_2, \ldots, Q_k\})$. φ *is easily shown to be an RMS measure.*

Example 4.7.4 (Mathematical Models and the Material Conditional). *Mathematical models can be viewed as* $\{0, \infty\}$*-valued RMS measures. Given a formal language* **L** *that generates a collection of sentences* **S** *by means of logical connectives, a model* **M** *for* **L** *assigns to each sentence* $A \in \mathbf{S}$ *a truth value which can be denoted by* $\mathbf{M}(A)$ (= *true or false*). **M** *interacts with the connectives* $\sim$, &, *and* $\vee$ *as follows: for* $A, B \in \mathbf{S}$, $\mathbf{M}(\sim A) =$ *true* iff $\mathbf{M}(A) =$ *false*; $\mathbf{M}(A \,\&\, B) =$ *true* iff $\mathbf{M}(A) =$ *true and* $\mathbf{M}(B) =$ *true*; $\mathbf{M}(A \vee B) =$ *true* iff *at least one of* $\mathbf{M}(A)$ *and* $\mathbf{M}(B)$ *is true. The collection of all true sentences with respect to* **M**, *known as the theory of* **M**, *is usually denoted by* th(**M**) (= $\{A \in \mathbf{S} \mid \mathbf{M}(A)$ *is true*$\}$). *Associated with* **M** *is therefore an*

[18] Garey and Johnson (1979) is the place to start for computational complexity theory. Balcazar et al.'s (1988; 1990) two volumes present a more thorough exposition of the subject.

RMS measure φ *on* **S** *such that* $\varphi(C \mid A) = \infty$ *precisely when* **A** $\subset$ th(**M**) *and* $C \notin$ th(**M**), 0 *otherwise.*

Thus, for $A \in$ **S**, **M**(A) = *true* (*resp. false*) iff $\varphi(A) = 0$ (*resp.* ∞). *Moreover, if* **L** *employs the horseshoe connective "*$\supset$*" and* **M** *interprets "*$\supset$*" as a material conditional (i.e.,* $A \supset C$ *is false* iff *A is true and C is false), then for A, C* $\in$ **S**, **M**$(A \supset C)$ = *false* iff **M**(A) = *true and* **M**(C) = *false* iff $A \in$ th(**M**) *and* $C \notin$ th(**M**) iff $\varphi(C \mid A) = \infty$. *Analyzing the truth of the material conditional* $A \supset C$ *by means of a mathematical model* **M** *can therefore be accomplished equally well by means of the corresponding* $\{0, \infty\}$*-valued RMS measure* φ. *In the next example we shall see that RMS measures enable us also to characterize logical entailment.*

Example 4.7.5 (Minimum Proof Length Measures). *As a final example I want to consider a type of RMS measure that arises very naturally for formal systems. By a* formal system *I mean a collection of sentences* **S** *together with a consequence relation* **R**. **S** *will typically be a collection of well-formed formulas from a language* **L** *built up recursively through the application of various logical connectives.* **R** *will then be a collection of inference rules, which may be conceived as mappings that take nonempty collections of sentences in* **S** *to individual sentences in* **S** *(for instance, for any p and q in* **S**, *disjunctive syllogism takes* $\{p \vee q, \sim p\}$ *into q*).[19] *We shall always assume that* **R** *includes* reductio ad absurdum: *for any p and q in* **S**, *there is a mapping in* **R** *that takes* $\{p, \sim p\}$ *into q. Note that* **S** *is always closed under the inference rules of* **R**.

Every subset **B** *of* **S** *issues in a set of consequences, called the* deductive closure *of* **B** *with respect to* **R** *(i.e., all sentences in* **S** *provable from* **B** *via* **R**). *We'll denote the deductive closure of* **B** *with respect to* **R** *by* **R**(**B**). *For convenience we'll define a* deductive system *as any ordered pair* (**B**, **R**) *where* **B** *is a subset of* **S**, *usually referred to as the axioms of the deductive system, and* **R** *is a collection of inference rules under which* **S** *is closed. A proof with respect to the deductive system* (**B**, **R**) *is now an* $(N+1)$*-tuple of sentences* $(B_0, B_1, \ldots, B_N)$ *where each* B_i *is either an element of* **B** *or the output of some inference*

[19] We could permit **R** to include inference rules that map the empty collection of sentences in S to individual sentences in S. Such inference rules, however, serve the same function as the background assumptions qua axioms we consider momentarily. By restricting our attention to nondegenerate inference rules that operate on nonempty collections of sentences, a useful separation between inference rules and axioms is maintained.

rule $r \in \boldsymbol{R}$ applied to some subset of B_js for $j < i$. In this way $(B_0, B_1, \ldots, B_N)$ is said to be a proof of B_N. $\boldsymbol{R}(\boldsymbol{B})$ is then the totality of these B_Ns provable from $\boldsymbol{B}$ via $\boldsymbol{R}$. Note that we always take N to be finite and therefore restrict proofs to N-tuples of finite length. Note also that $(\boldsymbol{B}, \boldsymbol{R})$ is inconsistent just in case there is some $A \in \boldsymbol{S}$ such that $A \in \boldsymbol{R}(\boldsymbol{B})$ and $\sim A \in \boldsymbol{R}(\boldsymbol{B})$.

Suppose next that $\pi = (B_0, B_1, \ldots, B_N)$ is a proof of the sentence B_N within the deductive system $(\boldsymbol{B}, \boldsymbol{R})$. Define λ as the following measure of proof length: for $\pi = (B_0, B_1, \ldots, B_N)$, $\lambda(\pi) = N$. Given λ we can define a function φ that for any sentence D in $\boldsymbol{S}$ and any subset $\boldsymbol{B}$ of $\boldsymbol{S}$ asks what is the length of the shortest proof of D within the deductive system $(\boldsymbol{B}, \boldsymbol{R})$ ($\boldsymbol{R}$ stays fixed). Formally, we define the minimum proof length measure as follows:

(4.7.9) $\varphi(\mathrm{D} \mid \mathbf{B}) =_{\mathrm{def}} \inf\{\lambda(\pi) \mid \pi \textit{ is a proof of } \mathrm{D} \textit{ within } (\mathbf{B}, \mathbf{R})\}$.

Several observations are now in order. First, it is straightforward to show that φ satisfies (4.7.1)–(4.7.3) *and is therefore an RMS measure.*[20] *Second, φ is finite precisely on the deductive closure $\boldsymbol{R}(\boldsymbol{B})$, and infinite elsewhere.*[21] *Third, what we defined in Section* 4.4 *as the axioms/implicit resources of the complexity system $(\boldsymbol{S}, \varphi)$ (i.e., $Ax_\varphi(\boldsymbol{S}) = \{A \in \boldsymbol{S} : \varphi(A) = 0\}$) is none other than the axioms of the deductive system $(\boldsymbol{B}, \boldsymbol{R})$ (i.e., $\boldsymbol{B}$). In other words, $Ax_\varphi(\boldsymbol{S}) = \boldsymbol{B}$.*[22] *Fourth, consistency of the deductive system $(\boldsymbol{B}, \boldsymbol{R})$ has a natural characterization in terms of φ: $(\boldsymbol{B}, \boldsymbol{R})$ is inconsistent just in case there is some $A \in \boldsymbol{S}$ such that $\varphi(A \mid \boldsymbol{B}) < \infty$ and $\varphi(\sim A \mid \boldsymbol{B}) < \infty$. Fifth, φ takes values only in $N \cup \{\infty\}$ – the natural numbers plus infinity (in particular, $\varphi(D \mid \boldsymbol{B}) = \infty$ means D has no proof from $\boldsymbol{B}$ via $\boldsymbol{R}$).*

[20] (4.7.1) and (4.7.2) are trivial to prove. (4.7.3) is a matter of concatenating proofs, but does require a little bookkeeping: to show $\varphi(\mathrm{B} \mid \mathrm{C}) \leq \varphi(\mathrm{A} \mid \mathrm{C}) + \varphi(\mathrm{B} \mid \mathrm{C} \cup \{\mathrm{A}\})$, let $(\mathrm{D}_0, \mathrm{D}_1, \ldots, \mathrm{D}_M)$ be a minimal length proof of A from C ($\mathrm{D}_M = \mathrm{A}$) and let $(\mathrm{E}_0, \mathrm{E}_1, \ldots, \mathrm{E}_N)$ be a minimal length proof of B from $\mathrm{C} \cup \{\mathrm{A}\}$ ($\mathrm{E}_N = \mathrm{B}$). If A does not appear in $(\mathrm{E}_0, \mathrm{E}_1, \ldots, \mathrm{E}_N)$, then $\varphi(\mathrm{B} \mid \mathrm{C}) = \mathrm{N}$ and we are done. If A does appear in $(\mathrm{E}_0, \mathrm{E}_1, \ldots, \mathrm{E}_N)$, then because the proof is with respect to resources $\mathrm{C} \cup \{\mathrm{A}\}$, we may assume that $\mathrm{E}_0 = \mathrm{A}$ (if not, simply move A to the E_0 position – this rearrangement is still a proof of B from $\mathrm{C} \cup \{\mathrm{A}\}$). It follows that $(\mathrm{D}_0, \mathrm{D}_1, \ldots, \mathrm{D}_{M-1}, \mathrm{A}, \mathrm{E}_1, \ldots, \mathrm{E}_N)$ is a proof of length $\mathrm{M} + \mathrm{N} = \varphi(\mathrm{A} \mid \mathrm{C}) + \varphi(\mathrm{B} \mid \mathrm{C} \cup \{\mathrm{A}\})$ from resources C. Hence $\varphi(\mathrm{B} \mid \mathrm{C}) \leq \varphi(\mathrm{A} \mid \mathrm{C}) + \varphi(\mathrm{B} \mid \mathrm{C} \cup \{\mathrm{A}\})$.

[21] If D has a proof π within the deductive system (**B**, **R**), then $\lambda(\pi)$ is finite and so the infimum in (4.7.8) is finite as well. If on the other hand D has no proof within the deductive system (**B**, **R**), then the infimum in (4.7.8) is over an empty set. The convention among mathematicians is to identify infima over empty sets with positive infinity, i.e., ∞.

[22] The only 1-tuple proofs of the form (B_0) are those where B_0 belongs to **B**. This follows because the inference rules **R** were defined to operate only on nonempty collections of sentences.

Sixth, for φ to be well-defined, φ must be limited to proofs of minimum length since proofs can always be padded with irrelevant steps.

The minimum proof length measure φ allows for a more precise analysis of proof–theoretic notions like entailment and consistency than is possible within ordinary proof theory, which is concerned not with the length/difficulty of proofs, but simply with whether or not proofs exist. Consider, for instance, Harvey Friedman's notion of n-consistency.[23] *According to Friedman a deductive system* $(\boldsymbol{B}, \boldsymbol{R})$ *is* n-*consistent if there is no proof of a contradiction in the system whose length is less than or equal to n steps. Since an inconsistent system entails a contradiction in finitely many steps, we may ask, What is the fewest number of steps required to reach a contradiction? If the fewest number of steps is* $n + 1 (= \varphi(A \ \& \ {\sim}A \mid \boldsymbol{B})$ *for some* $A)$*, then this inconsistent system is nevertheless n-consistent. An inconsistent system is therefore always n-consistent for some n. Now if n is very large, it may be utterly beyond our resources to discover a contradiction that renders the system logically inconsistent. Thus, for practical purposes, an n-consistent system for which n is so large that we cannot hope in practice to discover a contradiction from it may be indistinguishable from a logically consistent system (i.e., an* ∞*-consistent system in Friedman's terminology).*

Arend Heyting has taken up this topic, though in a somewhat different guise, in his book Intuitionism. *There he presents a delightful dialogue in which proponents of various philosophical positions concerning the nature of mathematics argue their views. In this dialogue Heyting places the pragmatic view of consistency I have just described in the mouth of an interlocutor named Letter. A Hilbertian formalist known as Form has just demanded of Letter that Letter provide "some modes of reasoning to prove the consistency of your formal system" (Heyting, 1971, p. 7). Letter's response, particularly in light of Gödel's second theorem, seems entirely appropriate (Heyting, 1971, p. 7): "Why should I want to prove* [*consistency*]*? You must not forget that our formal systems are constructed with the aim towards applications and that in general*

[23] Friedman's seminal work on this topic, though never published, was circulated as an unpublished preprint entitled "On the Consistency, Completeness and Correctness Problems" (Columbus, Ohio, Ohio State University, 1979). More recently, Friedman's ideas about n-consistency have been revived in the work of Krajicek and Pudlak (1989).

they prove useful; this fact would be difficult to explain if every formula were deducible in them. Thereby we get a practical conviction of consistency which suffices for our work." The minimum proof length measure along with Friedman's notion of n-consistency helps to make sense of what Heyting calls "a practical conviction of consistency."

Like consistency, entailment too receives a more precise analysis at the hands of the minimum proof length measure. Instead of simply determining whether a certain set of premises permits the proof of a certain claim, and thereby entails that claim, by asking instead for the minimum length of proof, the minimum proof length measure offers some insight into how difficult it is to prove things. To be sure, there is no exact correspondence between difficulty of proving something and minimum length of proof: Some things are tedious to prove (i.e., require many steps), but are perfectly straightforward whereas others can be proved very quickly (i.e., require very few steps) once one grasps a crucial insight, though achieving the crucial insight may be exceedingly difficult. Indeed, complexity theory may be utterly irrelevant to characterizing the difficulty of achieving the crucial insights that lead to groundbreaking conceptual advances (take for instance the introduction of the zero or the complex numbers in mathematics). Nevertheless, other things being equal (and I stress this qualification), it seems reasonable to think that the difficulty of proving something increases as the minimum number of inferential steps needed to prove it increases.

Something like this certainly seems to be the case in mathematics. As Bradley and Swartz (1979, pp. 147–9) observe,

> *There are ... some propositions the knowledge of whose truth, if it is humanly possible at all, can be acquired only by an enormous investment in inferential reasoning [cf. expenditure of effort]. The proofs of many theorems in formal logic and pure mathematics certainly call for a great deal more than simple analytical understanding of the concepts involved. And in some cases the amount of investment in analysis and inference that seems to be called for, in order that we should know whether a proposition is true or false, may turn out to be entirely beyond the intellectual resources of mere human beings.*
>
> *As a case in point consider the famous, but as yet unproved, proposition of arithmetic known as Goldbach's Conjecture, viz., Every even number greater than two is the sum of two primes.... Goldbach's Conjecture is*

> easily understood. In fact we understand it well enough to be able to test it on the first few of an infinite number of cases. . . . [But] for all we know, it may turn out to be unprovable by any being having the capacities for knowledge-acquisition which we human beings have. Of course, we do not now know whether or not it will eventually succumb to our attempts to prove it. Maybe it will. In this case it will be known ratiocinatively. But then, again, maybe it will not. In that case it may well be one of those propositions whose truth is not known because its truth is unknowable. At present we simply do not know which.

The "enormous investment in inferential reasoning," the "intellectual resources of mere human beings," and "the capacities for knowledge-acquisition which we human beings have" can all be unpacked in terms of the effort mathematicians expend trying to prove things. An infinitely powerful problem solver is able to settle the Goldbach Conjecture, either by providing a counterexample (i.e., an even integer greater than 2 that is not the sum of two primes), or by running through all the even integers greater than 2 and in each case finding a pair of primes that sums to it (this is of course a brute force approach, unlikely to win any prizes for elegance; but then again this is the virtue of an infinitely powerful problem solver – the ability to solve everything by albeit inelegant means).

Once the problem solver is limited, however, the question about resources and their optimal use cannot be avoided. The solutions to mathematical problems are widely held to be noncontingent since mathematical propositions are regarded as necessarily true or false. Nevertheless, the capacity of rational agents to solve mathematical problems is contingent, depending on the resources available to these agents. Their capacity to solve mathematical problems is therefore inextricably tied to the complexity of the problems under consideration and the amount of effort they can expend to try to solve the problems. Since the solutions to mathematical problems are typically proofs of theorems, and since the difficulty of coming up with these solutions correlates to some extent with minimum length of proof, the minimum proof length measure provides one way of studying what mathematical problems are solvable, and by extension what mathematical knowledge is attainable. Thus, in the minimum proof length measure we have an RMS measure that figures centrally into such key questions in the philosophy of mathematics as consistency, *entailment, and the practical limitations to mathematical knowledge. This ends the examples.*

In concluding this section, I want to state two useful results about constructing new RMS measures from old.

Proposition 4.7.1. *Let RMS(S) denote all the RMS measures defined on S* $\times$*pow(S). Then RMS(S) satisfies the following properties:*

(i) For every $\varphi \in RMS(S)$ *and every nonnegative real number c,* $c\varphi \in RMS(S)$.
(ii) For all φ *and* $\psi \in RMS(S)$, $\varphi + \psi \in RMS(S)$.
(iii) For every collection of RMS measures $\mathbf{F} \subset RMS(S)$, $\xi(C \mid A) = sup\{\varphi(C \mid A) \mid \varphi \in \mathbf{F}\}$ *defines an RMS measure in RMS(S).*

Remarks. (i) and (ii) together imply that RMS(**S**) is what mathematicians call a cone. Since by convention the supremum of an empty set of nonnegative numbers is 0, if **F** is empty, ξ is just the RMS measure that is identically zero.

Proof. (i) and (ii) are obvious. (iii) follows immediately by noting that $\sup\{\varphi(C \mid A) + \varphi(D \mid \mathbf{A} \cup \{C\}) \mid \varphi \in \mathbf{F}\} \leq \sup\{\varphi(C \mid A) \mid \varphi \in \mathbf{F}\} + \sup\{\varphi(D \mid \mathbf{A} \cup \{C\}) \mid \varphi \in \mathbf{F}\}$.

Proposition 4.7.2. *If* ψ *is any RMS measure on* **S** *and* **A** *is a fixed subset of* **S***, then the function* φ *defined by* $\varphi(C \mid \mathbf{B}) =_{\text{def}} \psi(C \mid \mathbf{A} \cup \mathbf{B})$ *is also an RMS measure on* **S** *(*$C \in S$ *and* $\mathbf{B} \subset S$*).*

The proof is trivial and is therefore omitted. Proposition 4.7.2 allows us to incorporate resources directly into RMS measures. Thus, for an RMS-measure/resource pair $(\psi, \mathbf{A})$, we can define the derived RMS measure $\varphi(C \mid \mathbf{B})$ as $\psi(C \mid \mathbf{A} \cup \mathbf{B})$ by incorporating the resources **A** directly into ψ. There is an analogy here with conditional probabilities: Just as probability measures conditioned on fixed background information are still probability measures, so RMS measures conditioned on fixed resources are still RMS measures. Further technical results regarding RMS measures can be found in Section 4.8.

Since an RMS measure φ is a partial function from $\mathbf{S} \times pow(\mathbf{S}) \to [0, \infty]$, one may ask what happens when the second argument of φ varies over subcollections of the axioms. Recall that in Section 4.4 the axioms of the complexity system $(\mathbf{S}, \varphi)$ were defined as $Ax_\varphi = \{A \in \mathbf{S} \mid \varphi(A \mid \emptyset) = 0\}$, that is, the sentences in $\mathbf{S}$ representing problems that can be solved without additional resources and without any difficulty. Since for RMS measures we construe difficulty in terms of effort/work (cf. Section 4.7), we can think of the axioms of φ as those problems that even without additional resources require no effort to solve.

Given an RMS measure φ on a collection of sentences $\mathbf{S}$, and a collection of axioms $\mathbf{A}$ that is a subset of Ax_φ, let us therefore start by asking, How does $\varphi(C \mid \emptyset)$ compare to $\varphi(C \mid \mathbf{A})$ for an arbitrary element C of $\mathbf{S}$? Alternatively, How does conditioning on the axioms affect the RMS measure φ? The following proposition offers a partial answer.

Proposition 4.8.1. *For any finite collection of axioms* $\mathbf{A}$ *from the RMS system* $(\mathbf{S},\varphi)$*, conditioning on* $\mathbf{A}$ *does not change the RMS measure* φ*, that is, for all* C *in* $\mathbf{S}$*,* $\varphi(C \mid \emptyset) = \varphi(C \mid \mathbf{A})$.

Proof. It is enough to note that for an arbitrary RMS measure ψ and any element A such that $\psi(A) = 0$, $\psi(C \mid A) = \psi(C)$: $\psi(C \mid A) \leq \psi(C)$ is always true (additional resources, in this case $\mathbf{A}$, can only decrease complexity), whereas by subadditivity $\psi(C) = \psi(C \mid \emptyset) \leq \psi(A \mid \emptyset) + \psi(C \mid A) = \psi(C \mid A)$. Thus, for the case in question, if $\mathbf{A} = \{A_1, A_2, \ldots, A_N\}$, for all C in $\mathbf{S}$, $\varphi(C \mid \emptyset) = \varphi(C \mid \{A_1\}) = \varphi(C \mid \{A_1, A_2\}) = \cdots = \varphi(C \mid \{A_1, A_2, \ldots, A_n\}) = \varphi(C \mid \mathbf{A})$.

This result, however, does not generalize to infinite collections of axioms. For consider the metric space of nonnegative reals, $[0, \infty)$, with the usual metric $d(x, y) = |x - y|$. As we saw in Example 4.7.1, $\varphi(x \mid A) = \inf\{d(x, y) \mid y \in A\})$ defines an RMS measure on $[0, \infty)$. By Proposition 4.7.2 $\varphi'(x \mid A) = \varphi(x \mid A \cup \mathbf{N})$ is also an RMS measure ($\mathbf{N}$ is the set of the natural numbers). Next define the following extended real-valued function Φ on the Cartesian product of the extended reals with its powerset (i.e., $\Phi: [0, \infty] \times pow([0, \infty]) \to [0, \infty]$):

$$\Phi(x \mid A) = \begin{cases} \varphi'(x \mid A \cap [0, \infty)) \text{ if } x \neq \infty \\ 0 \text{ if } x = \infty, \text{ and } \infty \in A \text{ or } A \cap \mathbf{N} \text{ is infinite} \\ \infty \text{ if } x = \infty, \ \infty \notin A, \text{ and } A \cap \mathbf{N} \text{ is finite.} \end{cases}$$

(4.8.1)

Φ is an RMS measure. Redundancy and monotonicity are obvious. Subadditivity requires checking a few cases: For A and B in $[0, \infty)$ and C in $pow([0, \infty])$ (n.b., this is the powerset of the extended reals, not just the powerset of the nonnegative reals),

(4.8.2) $$\Phi(b \mid C) \leq \Phi(a \mid C) + \Phi(b \mid C \cup \{a\})$$

follows immediately because φ' is subadditive. If $a = \infty$, $\infty \notin C$, and C contains only finitely many natural numbers, then $\Phi(a \mid C) = \infty$ and (4.8.2) follows immediately. If $a = \infty$ and $\infty \in C$, then equality holds in (4.8.2) (in this case $\Phi(a \mid C) = 0$). If $a = \infty$ and C contains infinitely many natural numbers but not ∞, then equality holds as well (again $\Phi(a \mid C) = 0$). Finally we need to consider the case where $a < \infty$ and $b = \infty$. Since $a < \infty$, both C and $C \cup \{a\}$ contain ∞, or both contain finitely many natural numbers, or both contain infinitely many natural numbers. In each case therefore $\Phi(b \mid C) = \Phi(b \mid C \cup \{a\})$. Having exhausted all possibilities, we see that Φ is subadditive.

To see now that Proposition 4.8.1 doesn't extend to infinite collections of axioms, observe that the axioms of Φ are precisely the natural numbers **N** and that $\Phi(\infty \mid A) = \infty$ implies $\Phi(\infty \mid A \cup \{u\}) = \infty$ for any $0 \leq u < \infty$ (for $\Phi(\infty \mid A) = \infty$ to hold, A has to have finite intersection with **N**; but in this case $A \cup \{u\}$ also has finite intersection with **N**, and since $u \neq \infty$, $\Phi(\infty \mid A \cup \{u\}) = \infty$). Hence, while it's true that for any finite subcollection C of **N**

$$\Phi(\infty \mid C) = \Phi(\infty) = \infty,$$

for any infinite subcollection D of **N**

$$0 = \Phi(\infty \mid D) \neq \Phi(\infty) = \infty.$$

Since therefore conditioning on infinitely many axioms of Φ changes Φ, it follows that Proposition 4.8.1 can't be extended to infinite collections of axioms.

Because such counterexamples to infinitary versions of Proposition 4.8.1 exist, the following definition is nonvacuous: An RMS measure φ on **S** is *regular* if conditioning on any subcollection of axioms does not change φ, that is, if for any subcollection of axioms **A** (finite or infinite) associated with φ, $\varphi(\mathrm{C}) = \varphi(\mathrm{C} \mid \mathbf{A})$ regardless of C. Observe that by Proposition 4.8.1 every RMS measure with only finitely many axioms is regular. Observe also that by the monotonicity of RMS measures it is enough to check the regularity of φ on the entire set of axioms Ax_φ: Since any subcollection of axioms **A** satisfies $\varphi(\mathrm{C}) \geq \varphi(\mathrm{C} \mid \mathbf{A}) \geq \varphi(\mathrm{C} \mid Ax_\varphi)$, as soon as it is determined that $\varphi(\mathrm{C}) = \varphi(\mathrm{C} \mid Ax_\varphi)$, it follows that $\varphi(\mathrm{C}) = \varphi(\mathrm{C} \mid \mathbf{A})$ for all subcollections of axioms **A**.

It's possible to define a stronger version of regularity. Since by Proposition 4.7.2 any RMS measure φ with respect to **S** and any fixed subset **A** of **S** yield a new RMS measure ψ defined by $\psi(\mathrm{C} \mid \mathbf{B}) = \varphi(\mathrm{C} \mid \mathbf{A} \cup \mathbf{B})$ ($\mathrm{C} \in \mathbf{S}$ and $\mathbf{B} \subset \mathbf{S}$ both arbitrary), one may ask if all such ψ are regular (**A** is here arbitrary). If all such derived RMS measures are regular, then we say φ is completely regular. This can be stated formally as follows: An RMS measure φ on **S** is *completely regular* if for any subset **A** of **S**, the derived RMS measure ψ defined by $\psi(\mathrm{C} \mid \mathbf{B}) = \varphi(\mathrm{C} \mid \mathbf{A} \cup \mathbf{B})$ is regular ($\mathrm{C} \in \mathbf{S}$ and $\mathbf{B} \subset \mathbf{S}$ both arbitrary). Note that in this case the derived RMS measure ψ includes **A** among its axioms. The following counterexample demonstrates that not all regular RMS measures are completely regular.

Let d be the ordinary absolute-value distance defined on $[0, \infty)$, the same as in the previous counterexample. We then define the RMS measure Ψ on $[0, \infty]$ as follows:

$$\Psi(\mathrm{x} \mid \mathrm{A}) = \begin{cases} \inf\{\mathrm{d}(\mathrm{x}, \mathrm{y}) \mid \mathrm{y} \in \mathrm{A} \cap [0, \infty)\} \text{ if } \mathrm{x} \neq \infty \\ 0 \text{ if } \mathrm{x} = \infty, \text{ and } \infty \in \mathrm{A} \text{ or } \mathrm{A} \cap \mathbf{N} \text{ is infinite} \\ \infty \text{ if } \mathrm{x} = \infty, \ \infty \notin \mathrm{A}, \text{ and } \mathrm{A} \cap \mathbf{N} \text{ is finite.} \end{cases}$$

(4.8.3)

The proof that Ψ is an RMS measure is virtually identical with the proof that Φ as defined in (4.8.1) is an RMS measure. Now because Ψ has no axioms (and thus a fortiori only finitely many axioms), by Proposition 4.8.1 Ψ is regular. Moreover, $\Phi(\mathrm{x} \mid \mathrm{A})$ is identical with $\Psi(\mathrm{x} \mid \mathrm{A} \cup \mathbf{N})$. Hence, since Φ fails to be regular, it follows that Ψ

cannot be completely regular. Thus, not all regular RMS measures are completely regular.

It's also possible to define a weaker version or regularity: An RMS measure φ on **S** is *semiregular* if for any subcollection of axioms **A** associated with φ, $\varphi(C) = \varphi(C \mid \mathbf{A})$ for all the absolutely intractable elements C of φ, that is, $\varphi(C \mid \mathbf{A}) = \infty$ whenever $\varphi(C) = \infty$ (the absolutely intractable elements are those with infinite complexity, like C here). We then define an RMS measure φ on **S** as *completely semiregular* if for any subset **A** of **S**, the derived RMS measure ψ defined by $\psi(C \mid \mathbf{B}) = \varphi(C \mid \mathbf{A} \cup \mathbf{B})$ is semiregular ($C \in \mathbf{S}$ and $\mathbf{B} \subset \mathbf{S}$ both arbitrary). Slight modifications of the two preceding counterexamples show that not all RMS measures are semiregular and that not all semiregular RMS measures are completely semiregular.

The next two propositions show that the RMS measures associated with metrics and deductive systems are completely regular.

Proposition 4.8.2 (cf. Example 4.7.1). *Let (M, D) be a metric space. Then the RMS measure $\varphi(x \mid A) =_{\text{def}} d(x, A)$ is completely regular.*

Proof. For an arbitrary subset A of M define the derived RMS measure $\psi(x \mid B) = \varphi(x \mid A \cup B)$ for all x in M and all B in *pow*(M). The axioms of ψ are those xs that satisfy $\psi(x) = \psi(x \mid \emptyset) = \varphi(x \mid A \cup \emptyset) = \varphi(x \mid A) = d(x, A) = 0$, which is none other than those xs in the closure of A (i.e., the set A together with its limit points; we'll denote the closure of A by $cl(A)$). In short, $Ax_\psi = cl(A)$. This follows from an easily verified result in point-set topology asserting that $d(x, A)$ and $d(x, cl(A))$ are identical functions of x. Given now that $Ax_\psi = cl(A) \supset A$, and thus that for all x in M $\psi(x) = \psi(x \mid \emptyset) = \varphi(x \mid A \cup \emptyset) = \varphi(x \mid A) = d(x, A) = d(x, cl(A)) = \varphi(x \mid cl(A)) = \varphi(x \mid A \cup cl(A)) = \psi(x \mid cl(A))$, it follows that for all x in M, $\psi(x) = \psi(x \mid Ax_\psi)$. All such ψ derived from φ are therefore regular. It follows that φ is completely regular.

Proposition 4.8.3 (cf. Example 4.7.5). *Let **R** comprise a set of inference rules defined with respect to the collection of sentences **S**. Let $\varphi(C \mid A)$ define the length of the shortest proof of C from premises A*

*($\subset$ S) via the inference rules **R**. Then φ is a completely regular RMS measure – that is, the minimum proof length measure is completely regular.*

Proof. Suppose $\mathbf{A} \subset \mathbf{S}$ and define $\psi(\mathrm{C} \mid \mathbf{B}) = \varphi(\mathrm{C} \mid \mathbf{A} \cup \mathbf{B})$, $\mathrm{C} \in \mathbf{S}$ and $\mathbf{B} \subset \mathbf{S}$ arbitrary. Then the axioms Ax_ψ associated with ψ comprise none other than the members of $\mathbf{A}$, that is, $Ax_\psi = \mathbf{A}$. Hence $\psi(\mathrm{C}) = \varphi(\mathrm{C} \mid \mathbf{A}) = \varphi(\mathrm{C} \mid \mathbf{A} \cup \mathbf{A}) = \psi(\mathrm{C} \mid \mathbf{A}) = \psi(\mathrm{C} \mid Ax_\psi)$, which establishes the regularity of ψ and thus the complete regularity of φ.

The two preceding propositions are special cases of a more general result. Call an RMS measure φ on **S** *finitely generated* if for every $\mathrm{C} \in \mathbf{S}$ and every $\mathbf{A} \subset \mathbf{S}$, **A** has a finite subset $\mathbf{A}_{\mathrm{fin}}$ (depending on C) such that $\varphi(\mathrm{C} \mid \mathbf{A}) = \varphi(\mathrm{C} \mid \mathbf{A}_{\mathrm{fin}})$. Call an RMS measure φ on **S** *singular* if for every $\mathrm{C} \in \mathbf{S}$ and every nonempty $\mathbf{A} \subset \mathbf{S}$, $\varphi(\mathrm{C} \mid \mathbf{A}) = \inf\{\varphi(\mathrm{C} \mid \mathrm{A}) \mid \mathrm{A} \in \mathbf{A}\}$.[24] It's clear that RMS measures derived from deductive systems are finitely generated, and that RMS measures derived from metrics are singular.

Proposition 4.8.4. *Finitely generated and singular RMS measures are completely regular.*

Proof. The proof for finitely generated RMS measures follows immediately from Proposition 4.8.1. The proof for singular RMS measures is virtually identical with the proof of Proposition 4.8.2 once we define the closure of **A** with respect to φ as $cl_\varphi(\mathbf{A}) = \{\mathrm{C} \in \mathbf{S} \mid \varphi(\mathrm{C} \mid \mathbf{A}) = 0\}$ ($\supset$ **A**), and note that the triangle inequality for RMS measures together with the assumption of singularity implies $\varphi(\mathrm{C} \mid \mathbf{A}) = \varphi(\mathrm{C} \mid cl_\varphi(\mathbf{A}))$ for all $\mathrm{C} \in \mathbf{S}$.

From a complexity–theoretic point of view individual axioms are null resources – conditioning on them individually or in finite batches is equivalent to conditioning on nothing at all (i.e., conditioning on the null set). What's more, for completely regular RMS measures infinite collections of axioms offer no advantage over finite collections

[24] Since the infimum of an empty collection of nonnegative real numbers is by convention ∞, we could have defined an RMS measure φ on **S** as *singular* if for every $\mathrm{C} \in \mathbf{S}$ and for every $\mathbf{A} \subset \mathbf{S}$ (empty or nonempty) $\varphi(\mathrm{C} \mid \mathbf{A}) = \inf\{\varphi(\mathrm{C} \mid \mathrm{A}) \mid \mathrm{A} \in \mathbf{A}\}$, thereby identifying $\varphi(\mathrm{C} \mid \emptyset)$ with ∞.

of axioms – conditioning on infinite collections is in this case also equivalent to conditioning on nothing at all. Conditioning on non-axioms, however, is a different story. Indeed, since $\varphi(\text{A} \mid \{\text{A}\}) = 0$ regardless of A, strict diminution of complexity is always a possibility when conditioning on nonaxioms. Subsets **D** of **S** that uniformly decrease $\varphi(\text{A} \mid \mathbf{D})$ as a function of A are of interest in this regard. Two such **D** are worth mentioning: (1) a subset **D** of **S** is *dense* in the RMS system $(\mathbf{S}, \varphi)$ if $\varphi(\text{C} \mid \mathbf{D}) = 0$ regardless of C; (2) a subset **D** of **S** is ε-*dense* in the RMS system $(\mathbf{S}, \varphi)$ if $\varphi(\text{C} \mid \mathbf{D}) < \varepsilon$ regardless of C $(\varepsilon > 0)$.

Finally, I want to point out an equivalence relation and a partial ordering that are naturally associated with RMS systems: Given $(\mathbf{S}, \varphi)$, it is natural to say that A is *equivalent* to B (written $\text{A} \Leftrightarrow \text{B}$) iff $\varphi(\text{A} \mid \text{B}) = \varphi(\text{B} \mid \text{A}) = 0$ and that A *dominates* B (written $\text{A} \Rightarrow \text{B}$ or alternatively $\text{B} \Leftarrow \text{A}$) iff $\varphi(\text{B} \mid \text{A}) = 0$ (A and B $\in$ **S**). Given the defining conditions for RMS measures (i.e., redundancy, monotonicity, and subadditivity) and the resulting triangle inequality (i.e., $\varphi(\text{C} \mid \text{A}) \leq \varphi(\text{B} \mid \text{A}) + \varphi(\text{C} \mid \text{B})$ for all A, B, and C in **S**), it follows that $\Leftrightarrow$ is an equivalence relation and that $\Rightarrow$ is a preorder that induces a partial order on the equivalence classes of $\Leftrightarrow$. Dominance can be used to examine such questions in the logic of conditionals as strengthening the antecedent and weakening the consequent (e.g., when does $\text{A} \Rightarrow \text{B}$ entail $(\text{A} \,\&\, \text{C}) \Rightarrow \text{B}$ and $\text{A} \Rightarrow (\text{B} \vee \text{C})$).

5

Specification

5.1 PATTERNS

Our aim remains to explicate and justify the Law of Small Probability. With the theoretical underpinnings for this law now in place (i.e., probability theory and complexity theory), our next task is to explicate the dual notions of specification and small probability. We treat specification in this chapter, small probability in the next. Specifications are those patterns that in combination with small probability warrant the elimination of chance. In explicating specification, we therefore need first to be clear what we mean by a pattern.

By a *pattern* we shall mean any description that corresponds uniquely to some prescribed event. Formally, a pattern may be defined as a description–correspondence pair $\langle D, * \rangle$ where the description D belongs to a descriptive language **D** and the correspondence * is a partial function between **D** and a collection of events **E** so that * includes D in its domain of definition (as a partial function * need not be defined on all of **D**). $\langle D, * \rangle$ is therefore a pattern relative to the descriptive language **D** and the collection of events **E**. Formally, **D** can be any nonempty set,[1] and **E** any collection of actual or possible events. What makes **D** "descriptive" is that the correspondence * maps descriptions to events. Since for any pattern $\langle D, * \rangle$ the correspondence * includes D in its domain of definition, we let D^* denote the event that corresponds to D under *. To simplify notation we frequently represent patterns simply by the letter D (the correspondence * in this case being implicit). It will be clear from context whether D signifies the full pattern $\langle D, * \rangle$ or merely its first component.

Given this definition of pattern we then say that an event E *conforms* to a pattern D just in case the occurrence of E entails the occurrence

[1] In this respect the descriptive language associated with a pattern is like the sentences on which a complexity measure is defined (cf. Section 4.4).

of D^* (or, in the notation of Section 3.6, $E \Rightarrow D^*$).[2] Alternatively, we say that a pattern D *delimits* an event E just in case E conforms to D. In case D^* and E are identical events (i.e., $D^* = E$, which holds just in case $E \Rightarrow D^*$ and $D^* \Rightarrow E$), we say that the pattern D *matches* the event E, or equivalently that the event E *matches* the pattern D. We also say that D^* *subsumes* E just in case E entails D^*. These definitions help us relate patterns to events while preserving their differences.

To illustrate these definitions, consider the following example. A pair of dice is rolled. There are thirty-six possible ways the dice might land. These possibilities comprise events of the form (x, y), where x and y are numbers between 1 and 6. Thus, we take (3, 5) as the event of rolling a 3 with the first die and a 5 with the second. Although the elementary outcomes for rolling the pair of dice all take this form, we typically expand our conception of events to include more general events, like rolling the dice so that the sum of their faces is at least 11, an event we can represent as (5, 6)(6, 5)(6, 6). The relevant collection of events **E** from rolling two dice then comprises all strings of the form

$$(x_1, y_1)(x_2, y_2) \cdots (x_n, y_n)$$

where the x_is and y_is are between 1 and 6.

As for the descriptive language **D** that describes the events in **E**, we can take **D** to comprise ordinary English sentences. The correspondence * that maps **D** into **E** can then be taken as our ordinary way of interpreting English sentences. Thus, the pattern ⟨"the sum of faces equals seven," *⟩ matches the event

$$E = (6, 1)(5, 2)(4, 3)(3, 4)(2, 5)(1, 6).$$

Moreover, this pattern delimits (without matching) the more highly constrained events like (6, 1) and (2, 5)(1, 6) and (4, 3)(3, 4)(2, 5), all of which entail E.

5.2 THE REQUISITE PRECONDITION

As noted in Chapter 1, patterns come in two varieties, *specifications* and *fabrications*. Specifications are the good patterns, the ones that

[2] To say that the event E entails the event D^* is to say that if the event E should happen, then the event D^* must happen as well. For instance, buying a Rolls Royce entails buying a car.

legitimately warrant eliminating chance, whereas fabrications are the bad patterns, the ones that are ad hoc and do not warrant eliminating chance. What, then, marks the difference between specifications and fabrications? Before we can even begin to answer this question, certain items need to be in place. We shall refer to these items collectively as the *requisite precondition.*

To decide whether a pattern is suitable for eliminating chance, we are never just given an event E and a pattern D to which E conforms. In addition we are given a descriptive language **D** and a collection of events **E**. Moreover, since D is shorthand for ⟨D, *⟩, we are also given a partial function * that maps **D** to **E**, and in particular associates D with the event D*. Like the sentences to which complexity measures apply (cf. Section 4.4), the collection **D** can be any nonempty set. The important thing is that D be in the set and that the correspondence * map D to an event D* entailed by E. As for the collection of events **E**, it may comprise any collection of actual or possible events so long as it contains E. Moreover, since we are trying to decide whether to attribute E to chance, we also need background information **H** (i.e., a chance hypothesis) characterizing how E might have occurred by chance. **H** in turn calls for a probability measure **P** to estimate likelihoods conditional on **H**, and in particular the likelihood of E conditional on **H**.

Still more is needed to establish a pattern's suitability for eliminating chance. To see this, consider the following event E_ψ, an event that to all appearances was obtained by flipping a fair coin 100 times:

(E_ψ) THTTTHHTHHTTTTTHTHTTHHHTT
HTHHHTHHHTTTTTTTHTTHTTTHH
THTTTHTHTHHTTHHHHTTTHTTHH
THTHTHHHHTTHHTHHHHTHHHHTT.

Is E_ψ due to chance or not? A standard trick of statistics professors in teaching introductory statistics is to have half the students in a class each flip a coin 100 times, recording the sequence of heads and tails on a slip of paper, and then have each student in the other half as a purely mental act mimic a sequence of 100 coin tosses, also recording the sequence of heads and tails on a slip of paper. When the students then hand in their slips of paper, it is the professor's job to sort the papers into two piles, those generated by flipping a fair coin, and those concocted in the students' heads. To the amazement of the students, the statistics professor is typically able to sort the papers with 100 percent accuracy.

There is no mystery here. The statistics professor simply looks for a repetition of six or seven heads or tails in a row to distinguish the truly random from the pseudo-random sequences (the truly random sequences being those derived from flipping a fair coin, the pseudo-random sequences being those concocted in the students' heads). In a hundred coin flips one is quite likely to see six or seven such repetitions. On the other hand, people concocting pseudo-random sequences with their minds tend to alternate between heads and tails much too frequently. Whereas with a truly random sequence of coin tosses there is a fifty percent chance that one toss will differ from the next, as a matter of human psychology people expect that one toss will differ from the next around seventy percent of the time.

How then will our statistics professor fare when confronted with E_ψ? Will she attribute E_ψ to chance or to the musings of someone trying to mimic chance? According to the professor's crude randomness checker, E_ψ would be considered truly random, for E_ψ contains a repetition of seven tails in a row. Everything that at first blush would lead us to regard E_ψ as truly random checks out. There are exactly fifty alternations between heads and tails (as opposed to the seventy that would be expected from humans trying to mimic chance). What's more, the relative frequencies of heads and tails check out: There were forty-nine heads and fifty-one tails. Thus, it's not as though the coin supposedly responsible for E_ψ was heavily biased in favor of one side versus the other.

Suppose, however, our statistics professor suspects she is not up against a neophyte statistics student, but is instead up against a fellow statistician trying to put one over on her. To help organize her problem, study it more carefully, submit it to computational analysis, and ultimately determine whether E_ψ occurred by chance, our statistics professor will find it convenient to let strings of 0s and 1s represent the outcomes of coin flips, say with 1 corresponding to heads and 0 to tails. In this case the following description D_ψ will describe the event E_ψ:

(D_ψ)   0100011011000001010011100
        1011101110000000100100011
    0100010101100111100010011
        0101011110011011110111100.

The collection of events **E** can then be defined as the powerset of the 100-fold Cartesian product $\{H, T\}^{100}$($\{H, T\}^{100}$ = the Cartesian product of {H, T} with itself 100 times), the descriptive language **D** can similarly be defined as the powerset of the 100-fold Cartesian product $\{0, 1\}^{100}$, and the correspondence * between **D** and **E** can be defined as the canonical extension of the function between {0, 1} and {H, T} that takes 1 to H and 0 to T. In this way D_ψ becomes a pattern that matches the event E_ψ, that is, $D_\psi^* = E_\psi$.

In trying to uncover whether E_ψ occurred by chance, our statistics professor needs next to identify the chance hypothesis **H** under which E_ψ might have occurred by chance, as well as the probability measure **P** that estimates the likelihood that an event in **E** occurs by chance under the chance hypothesis **H** (in particular, $\mathbf{P}(\cdot \mid \mathbf{H})$ will assign a probability to E_ψ). Since we are assuming that if E_ψ occurred by chance, then E_ψ occurred through the flipping of a fair coin, the chance hypothesis **H** characterizes a fair coin flipped 100 times under stochastically independent conditions where at each flip heads and tails have probability 1/2 (or, as statisticians would say, the coin flips are independent and identically distributed with heads and tails equiprobable). Thus, conditional on **H**, the probability measure **P** assigns to E_ψ the probability $\mathbf{P}(E_\psi \mid \mathbf{H}) = \mathbf{P}(D_\psi^* \mid \mathbf{H}) = 2^{-100}$, which is approximately 10^{-30}.

To organize her problem, and thereby help determine whether E_ψ occurred by chance, our statistics professor will therefore lay out the following six items: **E**, **D**, E_ψ, D_ψ, **H**, and **P**. Though necessary, by themselves these items are hardly sufficient to determine whether E_ψ occurred by chance. The reason is not difficult to see. Indeed, there is nothing in these six items to prevent D_ψ simply from being read off E_ψ. Thus far, instead of deriving D_ψ independently of E_ψ, our statistics professor has simply formulated D_ψ in response to E_ψ. These six items therefore provide no support one way or the other for determining whether E_ψ occurred by chance.

We are back to the question of how to explain an arrow stuck in a bull's-eye (cf. Section 1.2). Whether we attribute an arrow sticking in a bull's-eye to chance depends on whether the bull's-eye was in place before the arrow was shot, or whether it was painted around the arrow only after the arrow landed. The bull's-eye (= pattern) certainly delimits where the arrow landed (= event). But if the bull's-eye was painted around the arrow only after the arrow landed, there is no reason to think that where the arrow landed was dictated by something other than chance. Only if the position of the target (= pattern) is in some sense independent of whatever chance process might be responsible for the arrow's flight can this pattern preclude chance. Now specification is all about the type of independence that must obtain between a target's position (= pattern) and an arrow's flight (= event) to preclude chance.

What then is missing from **E**, **D**, E_ψ, D_ψ, **H**, and **P** that would render D_ψ independent of E_ψ, and thereby preclude E_ψ from occurring by chance? To answer this question let us rewrite D_ψ as follows:

(D_ψ)

```
   0
   1
  00
  01
  10
  11
 000
 001
 010
 011
 100
 101
 110
 111
0000
0001
0010
0011
0100
0101
0110
0111
1000
```

1001
1010
1011
1100
1101
1110
1111
00

By viewing D_ψ this way, anyone with the least exposure to binary arithmetic immediately recognizes that D_ψ was formulated simply by writing the binary numbers in ascending order, starting with the one-digit binary numbers (i.e., 0 and 1), proceeding to the two-digit binary numbers (i.e., 00, 01, 10, and 11), and continuing on up until 100 digits were recorded. It's therefore intuitively clear that D_ψ cannot describe a truly random event (i.e., an event gotten by tossing a fair coin), but instead describes a pseudo-random event (hence the subscript ψ), concocted by doing a little binary arithmetic.

Although it's now intuitively clear why chance cannot properly explain E_ψ, let us consider more closely why this is the case. In trying to unravel whether E_ψ occurred by chance, we initially laid out **E**, **D**, E_ψ, D_ψ, **H**, and **P**. By themselves these items proved insufficient to eliminate chance. Rather, to eliminate chance we had also to recognize that D_ψ could be readily obtained by performing some simple arithmetic operations with binary numbers. Thus, to eliminate chance we needed to supplement these items with side information **I** comprising our knowledge of binary arithmetic. For many of us **I** will include the following items of information:

I_1: Binary numbers are ordinarily represented with the symbols "0" and "1."

I_2: The binary number $b_n \cdots b_2 b_1 b_0$ equals the decimal number $b_0 2^0 + b_1 2^1 + b_2 2^2 + \cdots + b_n 2^n$.

I_3: 0 is a natural place to begin counting binary numbers.

I_4: In counting (binary) numbers, one lists them in order of increasing magnitude.

I_5: Binary numbers are naturally grouped according to the number of digits they employy.

I_6: When representing a k digit binary number with an m digit binary number where $m > k$, it is customary to prefix m–k 0s in front of the k digit number.

If we now equate **I** with the conjunction I_1 & I_2 & I_3 & I_4 & I_5 & I_6, **I** satisfies two conditions that enable it to eliminate chance. The first is that **I** is conditionally independent of E_ψ given **H**: Without knowing whether E_ψ happened and given only **H** and **I**, we know no more about what event happened than if we had simply been given **H** alone. Alternatively, supplementing **H** by **I** does not affect the probability we assign to E_ψ. This is certainly the case here since our knowledge of binary arithmetic does not affect the estimates of likelihood we assign to coin tosses. Conditional independence is the standard probabilistic way of unpacking epistemic independence. Two things are epistemically independent if knowledge about one thing (in this case **I**) does not affect knowledge about the other (in this case E_ψ's occurrence conditional on **H**). Conditional independence formalizes this relation by requiring that probabilities conditioned on additional knowledge remain unchanged.

The other condition **I** must satisfy to eliminate chance is that it has to be sufficient to formulate D_ψ. Treating E_ψ as an indeterminate event that occurred according to the chance hypothesis **H**, we pretend **I** contains crucial information about this event and try to formulate a pattern that delimits it. For the second condition to be satisfied D_ψ has to be among the patterns formulated this way. The logic underlying this second condition is important. Because the first condition established that **I** is conditionally independent of E_ψ given **H**, any knowledge we have of **I** ought to give us no knowledge about E_ψ so long as – and this is the crucial assumption – E_ψ occurred according to the chance hypothesis **H**. Hence any pattern formulated strictly on the basis of **I** ought not to give us any knowledge about E_ψ either. Yet the fact that it does, inasmuch as D_ψ delimits E_ψ, means that **I** is after all giving us knowledge about E_ψ. The assumption that E_ψ occurred according to the chance hypothesis **H** is therefore thrown into question.[3]

A precise formulation of these conditions will be given in the next section. For the moment, however, I want simply to observe that while

[3] The two conditions here described are necessary for side information to eliminate chance. These conditions become sufficient to eliminate chance only when the relevant probabilities become small enough. How small is small enough is the point of Chapter 6.

the first of these conditions is probabilistic, the second is complexity–theoretic. That the first is probabilistic is obvious. That the second is complexity-theoretic follows once we conceive pattern-formulation as problem-solving where the problem is formulating a given pattern (here D_ψ) and the resources constitute the given side information (here **I**). If we now let a complexity measure φ estimate the difficulty of such problems, to say that a pattern can be formulated on the basis of side information is to say that its complexity is less than some relevant tractability bound λ (recall that we consider those problems solvable whose complexity is less than a relevant tractability bound – cf. Section 4.5).

Formalizing this second condition requires the notion of a bounded complexity measure. A *bounded complexity measure* is any ordered pair $\Phi = (\varphi, \lambda)$ where φ is a complexity measure and λ is a tractability bound (cf. Sections 4.4 and 4.5 respectively). Bounded complexity measures not only estimate the difficulty required to solve problems (by means of the complexity measure φ), but also identify the problems that can actually be solved (by means of the tractability bound λ). Thus, for a bounded complexity measure $\Phi = (\varphi, \lambda)$ and a problem–resource pair (Q, R), Q is considered solvable provided that the complexity of Q given R is less than λ, that is, $\varphi(Q \mid R) < \lambda$. By utilizing both a complexity measure and a tractability bound, bounded complexity measures tell us what problems we can reasonably expect to solve in practice.

We can now list all the items that have to be in place for determining whether a pattern is suitable to eliminate chance:

(1) A collection of events **E**.
(2) A descriptive language **D**.
(3) An event E belonging to **E**.
(4) A pattern D whose correspondence * maps **D** to **E**.
(5) A chance hypothesis **H**.
(6) A probability measure **P** where $\mathbf{P}(\cdot \mid \mathbf{H})$ estimates the likelihood of events in **E** given **H**.
(7) Side information **I**.
(8) A bounded complexity measure $\Phi = (\varphi, \lambda)$ where $\varphi(\cdot \mid \mathbf{I})$ estimates the difficulty of formulating patterns in **D** given **I**, and λ fixes the level of complexity at which formulating such patterns is feasible.

If, therefore, we are already given an event E and a pattern D (and thus by implication a collection of events **E** and a descriptive language **D**), determining whether D is a suitable pattern for eliminating chance requires that the last four items on this list be in place, namely, a chance hypothesis **H**, a probability measure **P**, side information **I**, and a bounded complexity measure $\Phi = (\varphi, \lambda)$. We refer to these last four items collectively as a *requisite precondition*, and denote it by $\Sigma = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Phi)$.

5.3 DETACHABILITY

With this last piece of formal machinery in place, we are now finally in a position to define the type of independence that must obtain between a pattern and an event if that pattern is going to be suitable for eliminating chance:

Definition. *Given an event E, a pattern D (which may or may not delimit E), and a requisite precondition* $\Sigma = (\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{I}, \Phi = (\varphi, \lambda))$*, we say D is* **detachable** *from E relative to* Σ *if and only if the following conditions are satisfied:*

CINDE $\boldsymbol{P}(E \mid \boldsymbol{H} \& \boldsymbol{J}) = \boldsymbol{P}(E \mid \boldsymbol{H})$ *for any information* $\boldsymbol{J}$ *generated by* $\boldsymbol{I}$.
TRACT $\varphi(D \mid \boldsymbol{I}) < \lambda$.

If Σ is clear from context, we say simply that D is detachable from E. CINDE is short for the *conditional independence condition* and TRACT for the *tractability condition*.[4] Let us now turn to these conditions in detail.

First, to motivate these conditions, let us see how they arise within chance elimination arguments generally. The generic chance elimination argument starts with a rational agent, a subject S, who learns that an event E has occurred. By examining the circumstances under

[4] Computational complexity theorists familiar with zero-knowledge proofs will observe a similarity between the protocols of zero-knowledge proofs and the side information that renders a pattern detachable: Side information **I** which satisfies CINDE and TRACT is analogous to the protocol of a zero-knowledge proof in that **I** tells us nothing about the circumstances under which the event in question occurred (cf. the protocol in a zero-knowledge proof telling us nothing about how actually to prove the theorem in question) while still enabling us to delimit the event by means of a pattern (cf. the protocol nevertheless convincing us that the theorem is true). For the basics of zero-knowledge proofs consult Goldwasser, Micali, and Rackoff (1985) and Goldreich (1988).

which E occurred, S finds that a chance process characterized by the chance hypothesis **H** and the probability measure **P** could have been operating to produce E. S wants therefore to determine whether E occurred according to the chance hypothesis **H**. Now unless S can refer E to a pattern, S will be stuck attributing E to **H** since chance is always the default option in explanation.[5]

S therefore identifies a pattern D that delimits E. How S arrives at D is immaterial. S may simply read D off the event E, or S may propose D without any knowledge of E. The logic of discovery is immaterial. The important thing is that once D is in hand, S be able to determine whether D is the right sort of pattern to eliminate chance. How, then, is S going to show this? Specifically, how is S going to show D capable of eliminating **H** as the explanation of E? To eliminate **H**, S will have to identify certain side information **I** wherewith S can formulate D apart from any knowledge of E. As we saw in Section 5.2, such side information must satisfy two conditions. The first is that the side information **I** has to be conditionally independent of E given **H**: Without knowing whether E happened and given **H** and **I**, S must know no more about what event happened than if S had simply been given **H** alone. CINDE captures and makes precise this condition.

The second condition builds on the first. Having determined that **I** is conditionally independent of E given **H**, S sets aside any knowledge of E and assumes that whatever happened, happened according to the chance hypothesis **H**. Treating E as an indeterminate event, S investigates whether **I** may nonetheless contain crucial information about E. Admitting only that some event compatible with **H** has occurred, S attempts nonetheless to formulate a pattern that delimits E. But S already possesses such a pattern in D. S's task is therefore not strictly speaking to formulate D, but rather to confirm that D could have been formulated on the basis of **I**. And this is a matter of confirming that the problem of formulating D on the basis of **I** is tractable for S. With respect to a bounded complexity measure $\Phi = (\varphi, \lambda)$ that characterizes S's problem-solving capability (i.e., φ characterizes the degree of

[5] Cf. the notion of explanatory priority connected with the Explanatory Filter in Section 2.1. In that section we gave regularity explanations priority over chance explanations, and chance explanations priority over design explanations. Technically, therefore, the primary default option in explanation is regularity, with chance becoming the default option once regularity has been eliminated. But since regularity can be assimilated to chance as the special case of probabilities collapsing to zero and one, in a broader sense we may say that chance is the primary default option in explanation.

difficulty S faces in solving a problem, and λ the degree of difficulty below which S can still expect to solve problems), this is a matter of confirming that the degree of difficulty φ assigns to the problem of formulating D given **I** is strictly less than λ. TRACT coincides with this last condition.

The interrelation between CINDE and TRACT is important. Because **I** is conditionally independent of E given **H**, any knowledge S has about **I** ought to give S no knowledge about E so long as – and this is the crucial assumption – E occurred according to the chance hypothesis **H**. Hence any pattern formulated on the basis of **I** ought not to give S any knowledge about E either. Yet the fact that it does in case D delimits E means that **I** is after all giving S knowledge about E. The assumption that E occurred according to the chance hypothesis **H**, though not quite refuted, is therefore called into question. To actually refute this assumption, and thereby eliminate chance, S will have to do one more thing, namely, show that the probability $\mathbf{P}(D^* \mid \mathbf{H})$, that is, the probability of the event described by the pattern D, is small enough. Determining just how small this probability has to be to eliminate **H** is the subject of Chapter 6.

Three clarifications are worth appending to this account of chance elimination arguments. First, the subject S, the event E, and the side information **I** can just as well be counterfactual as actual. A subject S learns that an event E has occurred and comes up with side information **I** to determine whether E occurred by chance. On the one hand, S can be a nonfictional human subject learning of an event E that actually occurred and employing side information **I** that is all true. On the other hand, S may again be nonfictional, E may again have actually occurred, but this time S considers the counterfactual information **I**, and how it would affect S's determination of whether E occurred by chance. For instance, S may flip a coin a thousand times, thereby generating an actual event E, and then consider whether attributing E to chance would be tenable if a psychic had predicted this event in advance (the psychic's prediction here constituting counterfactual information). Any combination of factual and counterfactual Ss, Es, and **I**s is possible. Even the subject S can be fictional – everything from a character in a novel to a computational agent with unprecedented powers to a Peircean ideal community of rational agents.

Second, at the heart of any chance elimination argument is always the subject S who endeavors to eliminate chance. Everything depends on what S knows, believes, determines, and provisionally accepts. S considers an actual or possible event E. S posits a chance process **H** responsible for E. S introduces a probability measure **P** that together with **H** estimates likelihoods for events related to E. S exhibits the pattern D. S identifies the side information **I** that S will then use to detach D from E. S determines that **I** is conditionally independent of E given **H**. S determines that D could be formulated solely on the basis of **I** without recourse to E, that is, that formulating D on the basis of **I** constitutes a tractable problem. The complexity measure φ characterizes how S estimates the difficulty of problems like formulating D on the basis of **I**; moreover, the tractability bound characterizes where S locates the highest degree of difficulty that S can still handle. Thus, whereas **P** and **H** characterize S's knowledge of estimating likelihood, $\Phi = (\varphi, \lambda)$ and **I** characterizes S's knowledge of estimating problem-difficulty. And finally, S must establish that the relevant probabilities are small enough so that eliminating **H** as the explanation of E is warranted (cf. Chapter 6).

Third and last, identifying suitable patterns and side information for eliminating chance requires of S insight. This contrasts with how S identifies the other elements in a chance elimination argument. A subject S will as a matter of course confront all sorts of events that require explanation, among them E. As for how S estimates likelihood and problem difficulty, this can be referred to the community of discourse to which S belongs (see Sections 3.5 and 4.3). Thus, the conditional probability $\mathbf{P}(\cdot \mid \mathbf{H})$ that characterizes S's knowledge of estimating likelihoods and the bounded complexity measure $\Phi = (\varphi, \lambda)$ that characterizes S's knowledge of estimating problem difficulty can be referred to S's community of discourse – specifically the norms and practices developed by that community to estimate likelihood and difficulty. But how S identifies the patterns and side information that eliminate chance is less obvious. In Section 5.2, for instance, how does one see that the pattern D_ψ consists of an ascending sequence of binary numbers? Unless one knows what to look for, D_ψ will appear random. What's needed is insight, and insight admits no hard and fast rules (cf. Bourne et al., 1979, pp. 5–7). We have such insights all the time, and use them to identify patterns and side information that eliminate chance. But the logic of discovery by which we identify such patterns and side information is largely a mystery.

Having motivated the conditions defining detachability, let us now unpack them. First CINDE. The intuition underlying CINDE is that **I** should provide no clue about E's occurrence (i.e., **I** and E are epistemically independent). At the very least **I** therefore needs to be conditionally independent of E given **H** (i.e., $\mathbf{P}(E \mid \mathbf{H} \,\&\, \mathbf{I}) = \mathbf{P}(E \mid \mathbf{H})$). This isn't enough, however, since **I** may contain information that is not conditionally independent of E given **H** even when **I** taken as a whole is conditionally independent of E given **H** (cf. Section 5.6).[6] Side information always constitutes background information, and as such decomposes into elementary items of information (cf. Section 3.3). **I** can therefore be constructed from items of information $I_1, I_2, \ldots, I_N$ by means of the logical connectives $\sim$, &, and $\vee$.[7] Now the whole point of CINDE is that any information **J** constructed from these same $I_1, I_2, \ldots, I_N$ by means of the logical connectives must itself be conditionally independent of E given **H** (i.e., $\mathbf{P}(E \mid \mathbf{H} \,\&\, \mathbf{J}) = \mathbf{P}(E \mid \mathbf{H})$). For **J** to be constructible by means of logical connectives from the same items of information out of which **I** is constructed is what we mean by saying **I** *generates* **J**, or equivalently, **J** is *generated by* **I**.[8]

Next consider TRACT. The intuition underlying TRACT is that simply by using **I** it should be possible to reconstruct D. It's as though we had failed to note D, but were still able to recapture D by means of **I**. To be sure, we have noted D already. TRACT has us veil D and then asks whether **I** is sufficient to unveil D. Whether **I** is up to this task then gets unpacked as whether the problem of formulating D given **I** is tractable, which is what $\varphi(D \mid \mathbf{I}) < \lambda$, the inequality that defines TRACT, signifies. This inequality commits a mild abuse of notation. $\varphi(D \mid \mathbf{I})$ is meant to signify the estimated difficulty of formulating D given **I**. What $\varphi(D \mid \mathbf{I})$ literally expresses, though, is the estimated

[6] This is a common occurrence when working with conditional independence, and stochastic independence more generally. See Bauer (1981, p. 150, Example 1).

[7] For simplicity I am assuming that **I** is finitely generated by such "elementary items of information." By making this assumption I sidestep certain technical questions in probability theory. For all practical applications it is enough to assume that **I** is finitely generated. Nevertheless, a fully general treatment of the conditional independence condition requires the mathematical machinery alluded to in the following note.

[8] For those who prefer to do their probability theory with σ- or Boolean algebras, side information **I** can be viewed as isomorphic to a subalgebra of the algebra over which the probability measure **P** is defined. CINDE then says that for any element **J** in the subalgebra defined by **I**, **J** is conditionally independent of E given **H** (see Bauer, 1981, chs. 1, 5, and 10). This approach to CINDE through subalgebras is more general and powerful than the one presented here, though also less perspicuous.

difficulty of a pattern given side information. Whereas formulating a pattern constitutes a problem, a pattern by itself does not constitute a problem. Within the expression $\varphi(\mathrm{D} \mid \mathbf{I})$, D must therefore be construed as the problem of formulating the pattern D. It would therefore have been more accurate (and more awkward) to write something like $\varphi(\mathit{formulate}\,\mathrm{D} \mid \mathbf{I})$.[9]

Verifying TRACT is not a matter of formulating D from scratch. Typically, a subject S explicitly identifies an event E (whose explanation is in question), as well as a pattern D that delimits E. In particular, D will already be in S's hands. S's task, therefore, is not so much to formulate D, as to convince oneself that S could have formulated D knowing only that D delimits some indeterminate event compatible with **H**, and that **I** provides information that may be relevant to this indeterminate event. Having identified E and D, S pretends E and D are still unidentified, and then determines whether the side information **I** is enough to formulate D.

If S were literally to formulate D by means of **I**, S's strategy would be to generate as many patterns as possible, hoping that D was among them. Thus, to solve the problem of formulating D, S might have to exhibit other patterns as well. It's therefore not as though S has one, and only one, opportunity to hit the right pattern. What's crucial, however, is that the target pattern D be among those patterns exhibited. Thus, in the literal solution to the problem of formulating D by means of **I**, we imagine S generating a list of patterns: $D_1, D_2, D_3, \ldots$. The problem of formulating D is then literally solved so long as this list

[9] A fully formalized version of the tractability condition might therefore look as follows: given a descriptive language **D** and a collection of events **E**, let **PATT** denote the set of all description–correspondence pairs $\langle D, * \rangle$ where D is a member of **D** and * is a partial function from **D** to **E** such that the description D is in the domain of definition of * (i.e., **PATT** is the set of all patterns based on **D** and **E**). Next, given a subject S and the chance hypothesis **H**, let **INFO** be the set of all items of information that S might conceivably use when examining an event conceivably due to **H**. Next, take **PATT**, and using S and **H** transform each pattern D in **PATT** into the problem $\mathit{formulate}_{S,\mathbf{H}}\mathrm{D}$, i.e., the problem of S formulating D knowing only that D is a pattern in **PATT** and that D delimits some event compatible with **H**. Let **PROB** denote the set of all such $\mathit{formulate}_{S,\mathbf{H}}\mathrm{D}$. Next, form the set $\mathbf{S} = \mathbf{PROB} \cup \mathbf{INFO}$ and define a bounded complexity measure $\Phi = \langle \varphi, \lambda \rangle$ on **S** so that $\varphi(A \mid B)$ is defined only where A is in **PROB** and B is in the powerset of **INFO**. (Because complexity measures are partial functions, this is legitimate. Note that **S** constitutes the "sentences" on which the complexity measure φ is defined – see Section 4.4). Then $\varphi(\mathit{formulate}_{S,\mathbf{H}}\mathrm{D} \mid \mathbf{I})$ gives S's estimate of the difficulty S faces formulating D, knowing that D is a pattern in **PATT** and that D delimits some event compatible with **H**, as well as additionally knowing the side information **I**. Moreover, λ gives the degree of difficulty below which S can still expect to solve this problem. $\varphi(\mathrm{D} \mid \mathbf{I})$ is then a shorthand for $\varphi(\mathit{formulate}_{S,\mathbf{H}}\mathrm{D} \mid \mathbf{I})$, and $\varphi(\mathrm{D} \mid \mathbf{I}) < \lambda$, the tractability condition, signifies that S can indeed expect to solve the problem $\mathit{formulate}_{S,\mathbf{H}}\mathrm{D}$ by means of **I**.

contains D. The more patterns S is able to generate this way, the greater S's capacity for eliminating chance.

In concluding this section, I want to describe a convention that streamlines the tractability condition. In the tractability condition the inequality $\varphi(D \mid \mathbf{I}) < \lambda$ signifies that a subject S, simply by employing **I**, is able to formulate D. Now typically the side information **I** will not signify the totality of S's problem-solving capabilities, but only those items of information specifically relevant to formulating D (cf. Section 5.2 where **I** comprises six items of information, each about binary arithmetic). We therefore distinguish S's generic problem-solving capabilities, a resource we may call **G**, from the side information specifically relevant to formulating D, namely **I**. Since the resources in **G** are generic, any information generated by **G** & **I** will be conditionally independent of E given **H**, and so there's no cause for concern that **G** will upset the conditional independence condition. Still, it may seem disingenuous to write $\varphi(D \mid \mathbf{I}) < \lambda$ when in fact S is using not just **I**, but **G** & **I** to formulate D. As a matter of convention we therefore agree to incorporate the generic resources **G** directly into φ. Letting ψ denote a complexity measure that estimates S's difficulty of formulating patterns solely in relation to resources explicitly mentioned, we then define φ as follows: $\varphi(D \mid \mathbf{I}) =_{def} \psi(D \mid \mathbf{G} \,\&\, \mathbf{I})$ (cf. Proposition 4.7.2 where conditioning a complexity measure on fixed resources yields another complexity measure). This convention streamlines the tractability condition, focusing attention on those resources specifically relevant to formulating D, namely **I**.

5.4 SPECIFICATION DEFINED

In relating a pattern D to an event E, detachability places no restrictions on what we might call the "Venn diagrammatic relationship" between E and D* (recall that D* is the event described by D). Detachability simply has nothing to say about whether D* is compatible with E, overlaps with E, entails E, is entailed by E, coincides with E, or is inconsistent with E. For instance, D* might denote an arrow hitting a target, and E might denote that same arrow landing a thousand miles away from the target. Even if D and E are detachable (which in this case they may very well be), by doing nothing to locate E, D becomes useless for determining whether E occurred by chance. Indeed, a target (= pattern) that is detachable from an arrow's flight

(= event) can preclude chance only if the arrow hits the target. More generally, for a pattern to preclude the chance occurrence of an event, not only must it be detachable from the event, but it must also delimit the event. Only such patterns are specifications.

We therefore define specification as the following relation between patterns and events:

Definition. *Given an event E, a pattern D, and a requisite precondition* $\Sigma = (\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{I}, \Phi = (\varphi, \lambda))$, *we say that D is a* **specification** *of E relative to* Σ *(or equivalently that D* **specifies** *E relative to* Σ*) if and only if the following conditions are satisfied:*

CINDE $\boldsymbol{P}(E \mid \boldsymbol{H} \,\&\, \boldsymbol{J}) = \boldsymbol{P}(E \mid \boldsymbol{H})$ *for any information* $\boldsymbol{J}$ *generated by* $\boldsymbol{I}$.
TRACT $\varphi(D \mid \boldsymbol{I}) < \lambda$.
DELIM *D delimits E.*

DELIM is short for the *delimiter condition.* As we noted in Section 5.1, to say that D delimits E (or equivalently that E conforms to D) means that E entails D* (i.e., that the occurrence of E guarantees the occurrence of D*). Except for DELIM, this definition coincides with the definition of detachability. We can therefore shorten this definition by saying D is a specification of E relative to Σ (or equivalently that D specifies E relative to Σ) if and only if D is detachable from E relative to Σ and D delimits E. If Σ is clear from context, we simply say that D is a specification of E (or equivalently that D specifies E).

So long as Σ is clear from context, we can define the following four predicates.

detach(D, E) $=_{\text{def}}$ D is detachable from E.
sp(D, E) $=_{\text{def}}$ D is detachable from E and D delimits E.
sp(D) $=_{\text{def}}$ D specifies D*.[10]
sp(E) $=_{\text{def}}$ There is a pattern D such that D specifies E and D* = E.

These predicates will prove useful for explicating the Law of Small Probability in Chapter 6. Note that *sp* serves triple-duty, once as a two-place predicate, twice as a one-place predicate, in one instance

[10] Since D automatically delimits D*, this definition is equivalent to saying that D is detachable from D*.

applied to patterns, in the other to events. As a two-place predicate *sp* coincides with the definition of specification. As a one-place predicate *sp* answers either of two questions: (1) What does it mean for a pattern to be a *specification*? or (2) What does it mean for an event to be *specified*?

Specification is fundamentally a relational notion, relating patterns and events. To ask whether a pattern is a specification, or whether an event is specified is therefore like using a transitive verb intransitively. Strictly speaking, using a transitive verb intransitively is a mistake. Nevertheless, if the omitted direct object is implied, the mistake is avoided, and the intransitive use of a transitive verb becomes an elliptical expression that makes perfect sense. So too, specification, though fundamentally a relational notion, makes perfect sense as a monadic predicate so long as we fill in the missing term. Thus, if we want to speak of a pattern D constituting a specification without reference to some fixed event, we must let D* play the role of what otherwise would be the fixed event. On the other hand, if we want to speak of an event E as specified without reference to some fixed pattern, then we must be able to find a pattern D that not only specifies E, but for which D* equals E.[11] $sp(\mathrm{D})$ is logically equivalent to both $detach(\mathrm{D}, \mathrm{D}^*)$ and $sp(\mathrm{D}, \mathrm{D}^*)$. $sp(\mathrm{E})$ is logically equivalent to $\exists \mathrm{D}[sp(\mathrm{D}, \mathrm{E}) \ \& \ \mathrm{D}^* = \mathrm{E}]$, where the existential quantifier $\exists\mathrm{D}$ ranges over patterns.

In concluding this section, I want to consider what it is for a pattern not to be a specification (i.e., for a pattern to be what we've called a fabrication). Take therefore a pattern D that delimits an event E. What must fail for D not to specify E? Clearly D must not be detachable from E. But what is it for D not to be detachable from E? Detachability is always defined with respect to a requisite precondition $\Sigma = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Phi = (\varphi, \lambda))$. What's more, $\mathbf{H}$, $\mathbf{P}$, and Φ are usually fixed by context (usually $\mathbf{H}$ and $\mathbf{P}$ characterize the chance process supposedly responsible for E, and Φ characterizes the problem-solving ability of a subject S). It follows that once $\mathbf{H}$, $\mathbf{P}$, and Φ are fixed, to show that D is detachable from E, it is enough to produce side information $\mathbf{I}$ satisfying the conditional independence and tractability conditions. Thus conversely, to show that D is not detachable from E one must demonstrate that no such side information exists.

[11] After all, any event can be specified by a tautology. Thus, for $sp(\mathrm{E})$ to be an interesting predicate, the pattern that specifies E must delimit E as tightly as possible. This is of course best accomplished by finding a pattern D that actually matches E, i.e., $\mathrm{D}^* = \mathrm{E}$.

We now face a fundamental asymmetry. We can know with assurance that a pattern is detachable from an event – we simply have to exhibit side information that satisfies the conditional independence and tractability conditions. But what if we are unable to exhibit such side information? Suppose we are given a pattern D that delimits an event E (**H**, **P**, and Φ are fixed), and are unable to discover any side information satisfying the conditional independence and tractability conditions. Does it therefore follow that D is incapable of being detached from E? No, for it's always possible such side information exists, but has simply eluded us. Short of an explicit proof demonstrating that no side information is capable of detaching D from E, we remain undecided whether a pattern we have as yet failed to detach from an event might actually be detachable.

This asymmetry between knowing with assurance when a pattern is detachable and never quite knowing for sure whether an as-yet undetached pattern might in fact prove detachable is unavoidable. This asymmetry falls under the general truth that once a problem is solved, we can know with assurance that it is indeed solved, but until a problem is solved, short of an in-principle proof demonstrating that no solution exists, we must leave open the possibility that it has a solution. This asymmetry was implicit throughout the examples of Chapter 1, where we saw that design inferences could be reliably drawn, but not that design inferences could be definitively precluded. This asymmetry is especially apparent in the study of randomness (cf. Sections 1.6 and 5.10), where nonrandom bit-strings can be reliably identified, but random bit-strings are always in danger of being redesignated as nonrandom (the nonrandom strings being ascribed to design, the random strings being ascribed to chance). Or, as Persi Diaconis once remarked, "We know what randomness isn't, not what it is."[12] The next six sections form a series of case studies elaborating detachability and specification.

5.5 PYRAMIDS AND PRESIDENTS[13]

Suppose that at 10:00 a.m. Monday morning the president of the United States is assassinated. The resident psychics at the *National*

[12] Persi Diaconis made this remark at the Interdisciplinary Conference on Randomness, Ohio State University, 11–16 April 1988.
[13] I'm indebted to Philip Quinn for this example.

Enquirer are caught completely off guard. In trying to make sense of the president's assassination, the psychics recall last year's trip taken at company expense to Egypt and the pyramids. One of the psychics recalls an unusual inscription on one of the slabs that make up the Great Pyramid. Without further ado the psychics announce that this inscription predicted the president's assassination. Their announcement forms this week's headline in the *National Enquirer*.

Although we are accustomed to this sort of silliness from the *National Enquirer*, this example does point up the importance for the tractability condition of simultaneously formulating both a description D as well as a correspondence * mapping descriptions to events (i.e., of formulating a full-blown pattern ⟨D, *⟩ as opposed to merely a description D), if detachability is going to serve as a reliable tool for inquiry. Given all the information **I** that the psychics gathered on their visit to the pyramids – information we may assume is conditionally independent of the event E (i.e., the president's assassination) given the chance hypothesis **H** (i.e., all the information available to the Secret Service for gauging the president's safety) – the question we need to ask is this: Is **I** adequate for formulating not merely some description D that according to an arbitrary convention can then in post hoc fashion be connected to the event E, but rather for formulating simultaneously both a description D and a map * that maps D to E (i.e., a pattern ⟨D, *⟩ that delimits E) so that there is no question that **I** is genuinely informative about E?

Note that in this example no logical impossibility precludes **I** from being genuinely informative about E. If for instance certain hieroglyphics on the Great Pyramid should be found which, when translated, describe in minute detail the circumstances under which the president of the United States happened to die, it would be possible for enough information to be contained in these hieroglyphic inscriptions to leave no doubt that the president's assassination had been predicted. Though prophecies are typically ambiguous and fuzzy, there is nothing in principle to prevent a given prophecy from constituting a univocal and accurate prediction.

But of course, the inscription found by the psychics from the *National Enquirer* is nothing of the sort. All that the psychics have is some unusual markings which they have arbitrarily connected to the assassination of the president. The psychics have focused on certain features of the inscription and interpreted them as predicting the assassination of the president. Their interpretation is forced. They are putting on airs. They are claiming supernatural insights that they don't possess. What's more, they are claiming these insights simply to boost the circulation of their paper.

In sum, it is not enough for the side information **I** to yield a description D, which in conjunction with information beyond the scope of **I** can then be used to formulate a correspondence * such that D and * together then identify E. (Language after all is conventional. If we permit ourselves free rein with the correspondence *, we can fix any description we want – say something as simple as the single binary digit "0" – and make it refer to anything else we want, say Shakespeare's *Hamlet*. That * should make "0" refer to *Hamlet*, however, is totally arbitrary.) Rather **I** by itself must be genuinely informative about E, making possible the simultaneous formulation of a description D as well as a map * which together clearly identify E. In short, **I** needs to provide us with a full-blown pattern, and not merely with a description.

5.6 INFORMATION TUCKED WITHIN INFORMATION

For detachability to serve as a reliable criterion for eliminating chance, it is not enough that side information **I** be conditionally independent of E given **H**; rather, what's also needed is that any information **J** generated by **I** have this same property. To see that this latter, stronger condition is needed in formulating the conditional independence condition, suppose CINDE were formulated simply in terms of the conditional independence of **I** from E given **H**, that is, suppose CINDE simply read "**I** satisfies $\mathbf{P}(E \mid \mathbf{H} \,\&\, \mathbf{I}) = \mathbf{P}(E \mid \mathbf{H})$," and not as it does "**I** satisfies $\mathbf{P}(E \mid \mathbf{H} \,\&\, \mathbf{J}) = \mathbf{P}(E \mid \mathbf{H})$ for any information **J** generated by **I**." This weaker version of CINDE would then underwrite the elimination of chance even in cases where chance is the correct explanation.

To see this consider the following sequence of 100 coin tosses:

(E_R) HHTTTTHHTHTHHTTTHHTHHHHHH
HTHTTTHHTTTHHTHHTTHHHTHHH
TTTHHTTHTTTTHTHHHHTHHHTHH
TTHHHHHTHTTHTHTTHTHTHHHHT.

This is a genuinely random sequence of coin tosses gotten by tossing a fair coin 100 times. We now situate E_R within a requisite precondition. We therefore identify a descriptive language **D** (made up of strings of 0s and 1s where 0 represents tails and 1 represents heads); a collection of events **E** (events that consist of 100 coin tosses); the pattern D_R consisting of a correspondence * between **D** and **E** that is the canonical extension of the map that takes 0 to **T** and 1 to H, together with the description

(D_R)
1100001101011000110111111
1010001100011011001110111
0001100100001011110111011
0011111010010100101011110

which under * maps precisely onto the event E_R; the chance hypothesis **H** that treats coin tosses as stochastically independent, and assigns a probability of 1/2 to heads and tails each; and a probability measure **P** that assigns probabilities according to this chance hypothesis **H**. What's more, we let the Greek letter α represent the first 99 tosses of the sequence E_R, that is,

(α)
HHTTTTHHTHTHHTTTHHTHHHHHH
HTHTTTHHTTTHHTHHTTHHHTHHH
TTTHHTTHTTTTHTHHHHTHHHTHH
TTHHHHHTHTTHTHTTHTHTHHHH_.

E_R can therefore be represented by the composite αT.

Consider now the following two items of information:

I_1: The first 99 coin tosses of the event that happened are without a shadow of doubt known to be α.

I_2: The last toss of the event that happened is with probability $1–2^{-100}$ known to be H (in direct contradiction to E_R, whose last toss is T).

How might we have obtained these two items of information? We can imagine that we actually did observe the first 99 coin tosses (this gets us I_1), but then got distracted and missed seeing the last toss. Nevertheless, a friend of ours who happened to witness the last toss was kind enough to transmit what occurred on the last toss to us across a communication channel. This communication channel, however, has some noise, so that there is a small probability, namely 2^{-100}, that an H will come across the channel as a T and vice versa.[14] By some fluke, this is what happened – the T that actually happened has come to us as an H.

Suppose now that the side information **I** equals I_1 & I_2. Then it's clear from the way I_1 and I_2 have been defined that $\mathbf{P}(E_R \mid \mathbf{H} \,\&\, \mathbf{I}) = \mathbf{P}(E_R \mid \mathbf{H}) = 2^{-100}$. The side information **I** is therefore conditionally independent of E_R given **H**. Moreover, since from I_1 alone it follows that E_R has to be either αT or αH, **I** too constrains E_R to be either αT or αH. With only two events compatible with **I**, the problem of formulating patterns that match either of these events is now solvable in short order. In particular, the problem of formulating D_R on the basis of **I** is tractable. A relevant bounded complexity measure $\Phi = (\varphi, \lambda)$ is one that measures the time it takes for an ordinary human being to record bit strings of length 100, and sets λ at, say, ten minutes. Indeed, all that needs to be done is that sequence of 0s and 1s corresponding to αT and αH be written down (the correspondence * connecting descriptions to events is already given).

It follows that for the weaker version of CINDE (i.e., the version requiring only that the side information **I** be conditionally independent of the event given the chance hypothesis, and not that all additional side information **J** generated by **I** be conditionally independent in this way), the pattern D_R would be detachable from E_R, and thus constitute a specification. If, therefore, detachability is the key for eliminating chance, by substituting the weaker version of CINDE in the definition of detachability, we end up eliminating chance as the explanation of E_R even though chance is the right explanation (in this case we know the causal story behind E_R – that it was obtained by flipping a fair coin; the chance hypothesis **H** therefore has to be the right explanation of E_R).

On the other hand, with the original version of CINDE as it appears in the definition of detachability in Section 5.3, D_R won't be detachable through side information like $\mathbf{I} = I_1 \,\&\, I_2$. Yes, $\mathbf{P}(E_R \mid \mathbf{H} \,\&\, \mathbf{I}) = \mathbf{P}(E_R \mid \mathbf{H}) = 2^{-100}$, but $1/2 = \mathbf{P}(E_R \mid \mathbf{H} \,\&\, I_1) \neq \mathbf{P}(E_R \mid \mathbf{H}) = 2^{-100}$ and $2^{-199} = \mathbf{P}(E_R \mid \mathbf{H} \,\&\, I_2) \neq \mathbf{P}(E_R \mid \mathbf{H}) = 2^{-100}$. It follows that even

[14] Note that here and throughout the examples of this section I use the phrase "small probability" in a pretheoretic, intuitive sense. "Small probability" receives a precise sense in the next chapter.

though **I** is conditionally independent of E_R given **H**, there is information tucked inside of **I** (i.e., information generated by **I**) that is not conditionally independent of E_R given **H**. **I** therefore does not satisfy CINDE, and thus rightly fails to eliminate chance as the explanation of E_R.

5.7 PREDICTION

Although the general account of specification given in this chapter is new, a special case of specification is well-known to us all, namely, prediction. Consider the following argument:

Premise: Someone predicts that an event is going to happen.
Premise: Given the chance hypothesis **H**, the event has extremely small probability of happening.
Premise: The event happens (subsequent to the prediction).
Conclusion: The event did not happen by chance (i.e., the chance hypothesis **H** was not responsible for the event).

This argument is entirely unexceptional and receives the full endorsement of the statistics community. It is the classic chance-elimination argument (cf. Day, 1961, pp. 230–41). These days statisticians refer to predictions as *rejection regions*. C. S. Peirce (1883 [1955], pp. 207–10) referred to them as *predesignations*. Peirce himself employed predesignations to underwrite chance-elimination arguments within scientific reasoning.

As a general rule, if E is any event that occurs by chance according to the chance hypothesis **H** and probability measure **P**, and if **I** is any information identified prior to the occurrence of E that is compatible with **H**, then the conditional independence condition will be satisfied. **I** fails to be compatible with **H** if instead of supplementing the chance hypothesis **H** with further information, **I** substitutes an entirely new chance hypothesis. To see how this might happen, imagine the chance hypothesis **H** describes flipping a fair coin. Imagine further that right before E occurs, we learn the coin is two-headed. The side information we just learned is therefore incompatible with our original chance hypothesis **H**. Initially we thought we were going to flip a fair coin; next we discover we are going to flip a two-headed coin. Learning **I** forces us to discard **H** and replace it with a new chance hypothesis.

Such incompatibilities between **I** and **H** never arise for predictions. Predictions are a special form of side information. For **I** to be a prediction, **I** must assume the following form:

A subject S has exhibited a pattern D at time t_1.

What makes **I** a prediction of E is that E occurs at some time t_2 subsequent to time t_1 (i.e., $t_1 < t_2$). From this it's obvious why successful predictions are always specifications. Tractability is immediate because **I** contains an explicit reference to D – the problem of formulating D on the basis of **I** is therefore trivial. Conditional independence is immediate as well because the patterns a subject S formulates prior to a chance event cannot be based on any knowledge of the event, nor can such patterns in any way constrain what will happen – if this were not the case, casinos would quickly go out of business. To be sure, with side information identified after an event, we must always be on guard that we haven't smuggled in knowledge about the event's actual occurrence. But with side information identified before an event, our only concern is that the side information **I** be compatible with the chance hypothesis **H**. Since this is always the case with predictions, it follows that D is detachable from E. Moreover, for the prediction to be successful, D must delimit E, in which case D is a specification of E as well.

5.8 INCREASING THE POWER OF A COMPLEXITY MEASURE

The dual notions of detachability and specification presuppose a probabilistic set-up (i.e., a chance hypothesis **H** and a probability measure **P**) and a complexity–theoretic set-up (i.e., a bounded complexity measure $\Phi = (\varphi, \lambda)$). Suppose now that we are given a pattern D and an event E delimited by D, and that we fix a probabilistic set-up incorporating this pattern and event (i.e., we fix **H** and **P**), but let the complexity–theoretic set-up vary. The question I wish now to consider is this: As we let the complexity-theoretic set-up vary, how does the detachability of D from E vary? More specifically, what sort of relationship must exist between a pair of bounded complexity measures so that detachability with respect to one entails detachability with respect to the other?

To see what's at stake, suppose we are given two bounded complexity measures, $\Phi = (\varphi, \lambda)$ and $\Psi = (\psi, \mu)$. We assume that φ and ψ are complexity measures defined for the same problem–resource pairs (i.e., for any problem–resource pair (Q, R), $\varphi(Q \mid R)$ is defined iff $\psi(Q \mid R)$ is defined). Our question can then be reformulated as follows: What has to be true about the relation between Φ and Ψ so that if D is detachable from E relative to Φ, D will continue to be detachable from E relative to Ψ? The simplest way for this to occur is for the following condition to be satisfied: Whenever (Q, R) is a problem–resource pair that is tractable for Φ, (Q, R) is also tractable for Ψ, that is, $\varphi(Q \mid R) < \lambda$ entails $\psi(Q \mid R) < \mu$ for all problem–resource pairs (Q, R) in the mutual domain of definition of φ and ψ. If this relation holds between Φ and Ψ, we shall say that Ψ is more powerful than Φ.[15] Thus, whenever D is a pattern that is detachable from an event E relative to the requisite precondition $\Sigma = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Phi)$, if we then substitute for Φ a bounded complexity measure Ψ that is more powerful than Φ, D will be detachable from E relative to $\Sigma' = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Psi)$.

In practice, the power of complexity measures varies with technological advance. As technology advances, problems that were previously intractable often become tractable.[16] This is certainly the case in computer science where any computational problem solvable on one computer remains solvable on computers with faster processors and more memory. Thus, we may imagine someone given a pattern D, an event E, and a requisite precondition $\Sigma = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Phi)$, where Φ measures and bounds the difficulty of problems strictly in terms of the computational power available in 1960. To assess whether with side information **I** it is feasible to formulate the pattern D is then to assess whether computational resources as they stood in 1960 are adequate for solving this problem. Now it is plain that computational resources for solving computational problems have increased many orders of magnitude since 1960. Thus, any bounded complexity measure Ψ that measures and bounds the difficulty of problems in terms of the computational power available currently will be more powerful than Φ, with the result that past specifications will continue to be specifications, though past fabrications (i.e., patterns that in the past failed to count as specifications) may because of

[15] This relation is transitive: if Ξ is more powerful than Ψ and Ψ is more powerful than Φ, then Ξ is more powerful than Φ.

[16] There is no hard and fast rule here, however. Technological advance may also raise new problems and exacerbate old problems. For instance, tracking down computer criminals may be an intractable problem whereas tracking down good old-fashioned criminals may not. I am indebted to Dorothy Grover for this observation.

improvements in technology now become specifications. In short, increasing the power of a bounded complexity measure leaves intact what patterns were detachable previously, and may add some new ones as well.

5.9 CAPUTO REVISITED

I want next to reconsider the case of Nicholas Caputo (cf. Section 1.2). Recall that Nicholas Caputo was the Democratic clerk from Essex County, New Jersey, who in forty out of forty-one times assigned the Democrats the top ballot line in his county. When taken before the New Jersey Supreme Court on account of this anomaly, Caputo denied any wrongdoing. The court, however, remained unconvinced. Given the court's finding that "the chances of picking the same name 40 out of 41 times were less than 1 in 50 billion," the court suggested that Caputo institute changes in the way future ballot-line drawings would be conducted, changes that were to stem "further loss of public confidence in the integrity of the electoral process." By making this suggestion, the court in effect eliminated chance as the proper explanation for how Caputo chose his ballot lines. What I propose to do in this section is to rationally reconstruct (in terms of the technical apparatus developed in this and the last two chapters) the court's logic whereby it eliminated chance in explaining Caputo's anomalous ballot line selections.

As has been stressed throughout this monograph, extreme improbability by itself is not enough to preclude an event from having occurred by chance. Something else is needed. What else was there in the Caputo case? Caputo claimed to have made his ballot line selections using an urn model. If we now suppose that all selections were strictly between Democrats and Republicans, the chance hypothesis **H** can be conceived as an urn model with two balls, one labeled "D" for Democrat, the other "R" for Republican. With respect to **H** any sequence of Ds and Rs of length forty-one is equiprobable.

For concreteness, let us suppose Caputo's actual selection of ballot lines was as follows:

(E_G) DDDDDDDDDDDDDDDDDDDDDDRDDDDDDDDDDDDDDDDDD.

Thus, we suppose that for the initial twenty-two times Caputo chose the Democrats to head the ballot line; then at the twenty-third time he chose the Republicans; after which, for the remaining times, he chose the Democrats. The subscript G reminds us of Caputo's reputation as "the man with the *g*olden arm").[17]

Consider next a second possible course Caputo's career as county clerk might have taken. Let us therefore suppose that Caputo has again had forty-one occasions on which to select ballot lines in Essex County, but that this time Caputo chose both Democrats and Republicans to head the ballots pretty evenly – let us say in the following order:

(E_W) DRRDRDRRDDDRDRDDRDRRDRRDRRRDRRRDRDDDRDRDD.

In this instance the Democrats came out on top only twenty times, and the Republicans twenty-one times. The sequence of Ds and Rs that represent E_W was constructed just now by flipping a coin forty-one times and writing down "D" for each occurrence of heads, and "R" for each occurrence of tails. Since in this instance Caputo gave the Republicans the top ballot line more frequently than his own Democratic party, the subscript W signifies that Caputo is now "the man with the *w*ooden arm."

Clearly, had Caputo's ballot line assignments coincided with E_W instead of E_G, he would not have had to face the New Jersey Supreme Court. What then discriminated these events in the eyes of the court? Since $\mathbf{P}(E_W \mid \mathbf{H}) = \mathbf{P}(E_G \mid \mathbf{H}) = 2^{-41}$ (i.e., approximately one in two trillion), probabilities alone cannot discriminate between the events E_W and E_G. What besides sheer improbability, then, made the court unwilling to attribute E_G to chance? And why would the court have been willing to attribute E_W to chance?

To discriminate between E_W and E_G, the court needed something more than sheer improbability. What it needed was something like the following items of information:

I_1: Nicholas Caputo is a Democrat.

I_2: Nicholas Caputo would like to see the Democrats appear first on the ballot since having the first place on the ballot line significantly increases one's chances of winning an election.

[17] The *New York Times* article in which the Caputo case was discussed and which was cited in Section 1.2 did not explicitly mention the one place where Caputo gave the top ballot line to the Republicans. For the sake of this discussion I'll therefore assume E_G is what actually happened.

I_3: Nicholas Caputo, as election commissioner of Essex County, has full control over who appears first on the ballots in Essex County.

I_4: Election commissioners in the past have been guilty of all manner of fraud, including unfair assignments of ballot lines, especially when these assignments favored their own political party.

I_5: If Caputo were assigning ballot lines fairly, then both Democrats and Republicans should receive priority roughly the same number of times.

Given the side information $\mathbf{I} = I_1 \,\&\, I_2 \,\&\, I_3 \,\&\, I_4 \,\&\, I_5$, it is a simple matter for the court to formulate various "cheating patterns" by which Caputo might give the Democrats priority on the ballot. The most blatant cheat is, of course, to have the Democrats always attain priority. Next most blatant is to have the Republicans attain priority exactly one time (there are forty-one ways the Republicans can attain priority exactly one time – E_G being a case in point). Slightly less blatant – though still blatant – is to allow the Republicans to attain priority exactly two times (there are 820 ways the Republicans can attain priority exactly two times). This line of reasoning can be extended a bit further, with the Republicans being thrown a few additional sops, though still getting the short end of the stick. The point is, given this side information **I**, the court is easily able (possibly with the aid of a personal computer) to formulate and exhibit ways Caputo could cheat, one of which would surely include E_G.

What do these "cheating patterns" look like in terms of the formal apparatus developed earlier in this chapter? If we let our descriptive language **D** comprise ordinary English sentences, and let the correspondence * between descriptions **D** and events **E** simply be the ordinary way we interpret sentences in the English language, then we obtain the following cheating patterns:

D_{41}: ⟨"The Democrats got the top ballot line all 41 times," *⟩
D_{40}: ⟨"The Democrats got the top ballot line at least 40 times," *⟩
D_{39}: ⟨"The Democrats got the top ballot line at least 39 times," *⟩
etc.

These patterns capture the different ways Caputo might have cheated.

Of these patterns D_{40} is the one that particularly interests us. Indeed, this was the pattern implicitly used by the New Jersey Supreme Court to defeat chance as the explanation of Caputo's ballot line selections. D_{40} corresponds to an event consisting of forty-two possible outcomes: the Democrats attaining priority all forty-one times (one possible outcome) and the Democrats attaining priority exactly forty times (forty-one possible outcomes). Included among these possible outcomes is of course the event E_G, which for the sake of this discussion we are assuming is the event that actually occurred. It follows that E_G entails D^*_{40} (i.e., $E_G \Rightarrow D^*_{40}$) and therefore that D_{40} is a pattern that delimits E_G.

What, then, is it about the pattern D_{40} that warrants eliminating chance in the explanation of E_G? As usual we need a small probability event. Note, however, that for a generic pattern D and a generic event E delimited by D, the event that needs to have small probability to eliminate chance is not E but D*. Since D delimits E, E entails D* and hence $\mathbf{P}(E \mid \mathbf{H}) \leq \mathbf{P}(D^* \mid \mathbf{H})$ (cf. Section 3.6). Thus, for D* to have a small probability guarantees that E will have small probability as well. The reverse, however, is not the case. For a chance-elimination argument based on small probabilities to carry through, it is not enough for E to have small probability.

To see this, consider the tautological pattern T, whose associated event T* is the certain event (i.e., $\mathbf{P}(T^* \mid \mathbf{H}) = 1$). E can now be any event at all, even one of minuscule probability, yet T will delimit E. What's more, since no side information whatsoever is needed to formulate a tautology, T will be detachable from E as well (tautologies automatically satisfy CINDE and TRACT). T is therefore a specification. It follows that for specifications to be usefully employed in eliminating chance, it is not enough to constrain the probability of E. Rather, any pattern D that delimits E must correspond to an event that itself has small probability. In other words, D* – and not simply E – needs to have small probability.

These observations now apply to the Caputo case as follows. For the pattern D_{40} to warrant eliminating chance in the explanation of E_G, it is therefore necessary that $\mathbf{P}(D^*_{40} \mid \mathbf{H})$ have small probability. And indeed, this was precisely the probability that the New Jersey Supreme Court computed when it "noted that the chances of picking the same name 40 out of 41 times were less than 1 in 50 billion": $\mathbf{P}(D^*_{40} \mid \mathbf{H}) = 42 \times 2^{-41} \approx 0.00000000002$, or 1 in 50 billion. Equally important for eliminating chance in the explanation of E_G is that the pattern D_{40} avoid being ad hoc, which is to say that the pattern D_{40} must be detachable from E_G.

To show that D_{40} is detachable from E_G, it is enough to show that the side information **I** (= I_1 & I_2 & I_3 & I_4 & I_5) satisfies the conditional independence and tractability conditions. Tractability is straightforward. As noted earlier, from **I** it is easy to formulate the description D_{40} together with the correspondence * (which is simply our ordinary way of interpreting English sentences) so that D_{40}^* subsumes E_G. Indeed, once **I** is in hand, it is perfectly obvious which "cheating patterns" Caputo might have employed to circumvent chance, one of which is clearly D_{40}. Thus, given a complexity measure φ that models how difficult it is for the court to formulate possible sequences of ballot line selections from various types of side information, $\varphi(D_{40} \mid \mathbf{I})$ will be quite small, certainly less than any tractability bound λ reflecting the court's ability to formulate such sequences. Thus, for the pattern D_{40} (which was the pattern used by the New Jersey Supreme Court to assess the claim that Caputo's golden arm was simply a matter of luck), $\mathbf{P}(D_{40}^* \mid \mathbf{H})$ assumes a tiny probability whereas $\varphi(D_{40} \mid \mathbf{I})$ assumes a manageable level of complexity (i.e., $\varphi(D_{40} \mid \mathbf{I}) < \lambda$). This balancing of small probabilities with manageable complexities is typical of chance elimination arguments. Of course, for $\varphi(D_{40} \mid \mathbf{I})$ to assume a manageable level of complexity is just another way of saying the tractability condition is satisfied.

To show that **I** satisfies the conditional independence condition is equally straightforward. We have to show that E_G is conditionally independent of **H** given **I**. To see what could go wrong here, suppose that instead of simply the side information **I** consisting of the items of information I_1 through I_5, we had considered the side information **I′** consisting of the items of information I_1 through I_5 together with the following item of information, I_6:

I_6: Nicholas Caputo, as Essex County clerk, on forty-one occasions selected the ballot lines in his county as follows: on the first twenty-two occasions the Democrats got the top ballot line; on the twenty-third occasion the Republicans got the top ballot line; and on the remaining eighteen occasions the Democrats got the top ballot line.

Inserting I_6 into the side information for detaching D_{40} from E_G is clearly illegitimate. To eliminate chance in explaining E_G, we need to satisfy the tractability condition, and thus need to identify side information that in fairly short order yields a pattern that delimits E_G.

I_6, by explicitly identifying the event E_G, certainly fits the bill. But I_6 is an exercise in overkill. Incorporating I_6 into $\mathbf{I}'$, to be sure, enables us to satisfy the tractability condition (D_{40} can be directly read off I_6), but at the expense of the conditional independence condition. Indeed, precisely because it incorporates I_6, side information like $\mathbf{I}'$ is incapable of satisfying the conditional independence condition.

The conditional independence condition prevents bogus items of information like I_6 from insinuating themselves into the side information that's going to be used to eliminate chance. In the Caputo case all the items of information that make up **I**, namely I_1 through I_5, as well as any side information **J** generated from these items, will be conditionally independent of E_G given **H**. This is clear because each of I_1 through I_5 can be known and exhibited irrespective of Caputo's actual ballot line selections. Side information like $\mathbf{I}'$ that incorporates I_6, however, is a different story. I_6 explicitly states the precise sequence of Caputo's ballot line selections – I_6 tells us flat out that E_G happened. I_6 is therefore not conditionally independent of E_G given **H** ($2^{-41} = \mathbf{P}(E_G \mid \mathbf{H}) \neq \mathbf{P}(E_G \mid \mathbf{H} \,\&\, I_6) = 1$). It follows that $\mathbf{I}'$ does not satisfy the conditional independence condition. **I**, on the other hand, satisfies both it and the tractability condition, thereby detaching D_{40} from E_G.

5.10 RANDOMNESS REVISITED

Persi Diaconis once remarked, "We know what randomness isn't, not what it is."[18] In uttering this remark, Diaconis hoped that someday a positive theory of randomness would be put forward, one that would tell us precisely what randomness is, and not simply what it isn't. Given our discussion of detachability and specification we can now understand why this hope is unsustainable, and why randomness must always be approached through the back door of what in the first instance is nonrandom. Diaconis's remark, instead of expressing a distant hope about some future theory of randomness, thus expresses what actually is the fundamental truth about randomness, namely, that

[18] This remark formed the broad conclusion of the Interdisciplinary Conference on Randomness held at Ohio State University, 11–16 April 1988. Persi Diaconis and Harvey Friedman organized this conference during the height of the "chaos theory" enthusiasm. Although no proceedings were ever published, the conference remains significant for assembling philosophers, mathematicians, psychologists, computer scientists, physicists, and statisticians to compare notes specifically on the topic of randomness.

randomness can only be understood in reference to what is nonrandom (cf. Dembski, 1991).

In this section I want to show how this back-door approach to randomness has been inherent in the study of randomness all along. Andrei Kolmogorov's (1965) seminal work on randomness is representative here, and I shall focus on it. Specifically, I shall rationally reconstruct his approach to randomness in terms of the formal apparatus of probability, complexity, and specification developed in this and the last two chapters. Recall that Kolmogorov's problem in formulating a theory of randomness was that even in the simplest case of flipping a fair coin, Kolmogorov had no way of discriminating probabilistically between random and nonrandom sequences of coin flips (cf. Section 1.7).[19]

For instance, consider the following two events, each conceivably derived from flipping a fair coin:

```
(E_R)        HHTTTTHHTHTHHTTTHHTHHHHHH
                 HTHTTTHHTTTHHTHHTTHHHTHHH
             TTTHHTTHTTTTHTHHHHTHHHTHH
                 TTHHHHHTHTTHTHTTHTHTHHHHT

(E_N)        HHHHHHHHHHHHHHHHHHHHHHHHH
                 HHHHHHHHHHHHHHHHHHHHHHHHH
             HHHHHHHHHHHHHHHHHHHHHHHHH
                 HHHHHHHHHHHHHHHHHHHHHHHHH.
```

The causal story underlying these events differs: E_R was obtained just now by flipping a fair coin 100 times, whereas E_N was produced artificially. Now the problem facing Kolmogorov was this: Short of knowing the causal story underlying these events, Kolmogorov had no way on strictly probabilistic grounds for claiming that one, but not the other, of these events could reasonably be expected by chance. E_R looks like it could have happened by chance; E_N does not. Thus, we call E_R random and E_N nonrandom.

Although this way of distinguishing E_R from E_N is intuitively compelling, it needs to be made precise. To say that a sequence of heads and tails is the result of chance is to say it is the output of a certain type

[19] Kolmogorov's theory of randomness is nowadays referred to as *algorithmic information theory* (cf. van Lambalgen, 1989).

of causal process, namely, a chance or stochastic process. To attribute a sequence to chance is thus to tell a certain type of causal story (in this case that the sequence derives from flipping a fair coin). Thus, even a prototypically nonrandom sequence like E_N would be a chance sequence if the right sort of causal story underlies it (for instance, if it derived from flipping a fair coin). On the other hand, to say that a sequence of heads and tails is random is to say that it is representative of the type of sequence we could reasonably expect to occur by chance. Now "representativeness" is always a matter of degree, depending on how closely what's doing the representing matches what's being represented. It follows that randomness is a matter of degree, with sequences of heads and tails more or less random according to whether they are more or less representative of what we expect to occur by chance. In contrast, chance is all-or-nothing, with an event either having the right causal story or not.

Now it is a curious fact that randomness qua "representativeness of chance" cannot be cashed out in terms of the mathematical theory that properly characterizes chance, namely, probability theory. Even in as simple a case as tossing a fair coin, Kolmogorov found ordinary probability theory, with its usual way of computing probabilities, utterly incapable of distinguishing two events, like E_R and E_N, according to their degree of randomness. On strictly probabilistic grounds he could say nothing about which of these two events was *more random* (E_R and E_N have been suggestively subscripted, "R" for random, "N" for nonrandom). Kolmogorov wanted to say that E_R was "more random" than E_N – in fact, he wanted to say that whereas E_R could readily have happened by chance, there was no way that E_N could have happened by chance. But given nothing more than ordinary probability theory, Kolmogorov could at most say that each of these events had the same small probability of occurring, namely 1 in 2^{100}, or approximately 1 in 10^{30}. Indeed, every sequence of 100 coin tosses has exactly this same small probability of occurring.

Since probabilities alone could not discriminate E_R from E_N, Kolmogorov looked elsewhere. Where he looked was computational complexity theory. Identifying coin tosses with sequences of binary digits, and employing as his measure of computational complexity the length of the shortest program needed to generate a given binary sequence, what Kolmogorov said was that a sequence of 0s and 1s becomes increasingly random as the shortest computer program capable of generating the sequence becomes increasingly long (Kolmogorov, 1965).

Since such sequences of 0s and 1s can be construed as members of a descriptive language **D**, the events E_R and E_N correspond respectively to the following binary sequences:

```
(DR)        1100001101011000110111111
                1010001100011011001110111
            0001100100001011110111011
                0011111010010100101011110
```

and

```
(DN)        1111111111111111111111111
                1111111111111111111111111
            1111111111111111111111111
                1111111111111111111111111
```

where "1" corresponds to the occurrence of heads and "0" corresponds to the occurrence of tails. If we now let * be the canonical extension of the correspondence that takes 0 to T and 1 to H, then * connects the descriptive language **D** to the collection of possible coin tosses, which we denote by **E**. $\langle D_R, *\rangle$ and $\langle D_N, *\rangle$ are then patterns that match the events E_R and E_N respectively (as usual, we abbreviate these patterns by D_R and D_N respectively). What's more, given a chance hypothesis **H** that characterizes the flipping of a fair coin, $\mathbf{P}(\cdot \mid \mathbf{H})$ estimates the likelihood of events in **E**. In particular, $\mathbf{P}(E_R \mid \mathbf{H}) = \mathbf{P}(E_N \mid \mathbf{H}) = 2^{-100} \approx 10^{-30}$.

We now cash out randomness in terms of detachability: E_N is nonrandom if D_N can be detached from E_N; E_R is random if D_R cannot be detached from E_R. Accordingly, we need a bounded complexity measure $\Phi = (\varphi, \lambda)$ and side information **I** with respect to which D_N can be detached from E_N (thus rendering E_N nonrandom), but with respect to which D_R remains undetached from E_R (thus keeping E_R random). Moreover, once Φ and **I** are in place, we can cash out the degree to which E_N is nonrandom in terms of the magnitude of the complexity $\varphi(D_N \mid \mathbf{I})$ (the smaller $\varphi(D_N \mid \mathbf{I})$, the less random E_N).

Kolmogorov's own approach to randomness (see Kolmogorov, 1965) paralleled this approach to randomness. In particular, Kolmogorov proposed a computational complexity measure φ that for a given programming environment **I** computes the length of the shortest program needed to generate a given binary sequence (a programming environment **I** is a computer together with whatever software is running on that computer; the connection with our formal apparatus arises because **I** can also be represented as side information). To be sure, Kolmogorov used different notation. But this complexity measure is presupposed throughout his discussion, as is an underlying programming environment. Indeed, the length-of-shortest-program measure figured explicitly in Kolmogorov's approach to randomness.

Tractability bounds also figured in Kolmogorov's approach to randomness. Such bounds were implicit in the way Kolmogorov discriminated between random and nonrandom binary sequences. According to Kolmogorov, binary sequences become increasingly nonrandom as their shortest generating programs diminish in length. In designating a given binary sequence as nonrandom (as opposed to assigning it a *degree* of nonrandomness), Kolmogorov's approach therefore requires a cut-off according to which binary sequences generated by programs of length less than the cut-off are by definition nonrandom. But such a cut-off is none other than a tractability bound λ, with the nonrandom sequences then being the ones generated by programs of length less than λ. It's therefore no great leap to view Kolmogorov as putting forward a bounded complexity measure $\Phi = (\varphi, \lambda)$.

Within Kolmogorov's approach to randomness, the bounded complexity measure $\Phi = (\varphi, \lambda)$ can be conceived as follows. For a binary sequence D from the descriptive language **D** together with a given programming environment **I**, $\varphi(\mathrm{D} \mid \mathbf{I})$ yields the length of the shortest program that, when run within the programming environment **I**, outputs the binary sequence D. Since programs can always be represented as binary sequences, we can think of the programs within **I** that generate binary sequences as themselves binary sequences, and thus measure program length as the length of the binary sequences that represent the programs. We can, for instance, imagine **I** as designating a certain Macintosh personal computer running a given operating system and having installed on it a given Fortran compiler. $\varphi(\mathrm{D} \mid \mathbf{I})$ can then be taken as calculating the length of the shortest Fortran program (let us say the uncompiled version in ASCII code) that within this programming environment, when compiled and run, outputs D.

Now it is intuitively obvious that in terms of any programming environment **I** that we are likely to employ, the binary sequence $\mathrm{D_N}$ is going to require a much shorter program to generate than the binary sequence D_R. Indeed, any programming environment we are likely to employ will permit loops, and thus allow D_N to be compressed into a program that functionally is equivalent to "repeat '1' 100 times." D_R, on the other hand, precisely because it does represent a chance sequence generated by coin tosses, is highly unlikely to be compressible in this way (there is a simple combinatorial reason for this: When one enumerates all the possible binary sequences, there turn out to be a lot more incompressible sequences than compressible sequences – see Kolmogorov, 1965; compare also Barrow, 1991, pp. 14–6, 44–5, 269–71).

Thus, in terms of the complexity–theoretic formalism developed in Chapter 4, given a programming environment **I** capable of generating the binary sequences D_R and D_N, and a complexity measure φ that for a given programming environment measures the minimum program length needed to generate a given binary sequence, $\varphi(D_R \mid \mathbf{I})$ is going to be a lot bigger than $\varphi(D_N \mid \mathbf{I})$, or as it is sometimes written, $\varphi(D_R \mid \mathbf{I}) \gg \varphi(D_N \mid \mathbf{I})$. What's more, to designate D_R as random and D_N as nonrandom, the tractability bound λ will have to sit squarely between these two complexities, that is, $\varphi(D_N \mid \mathbf{I}) < \lambda \leq \varphi(D_R \mid \mathbf{I})$.

There is thus an asymmetry between what complexity tells us about the binary sequences D_R and D_N, and what probability tells us about the corresponding events E_R and E_N. The asymmetry consists in this: Whereas $\mathbf{P}(E_R \mid \mathbf{H})$ and $\mathbf{P}(E_N \mid \mathbf{H})$ are equal, $\varphi(D_R \mid \mathbf{I})$ and $\varphi(D_N \mid \mathbf{I})$ are sharply disparate with $\varphi(D_R \mid \mathbf{I}) \gg \varphi(D_N \mid \mathbf{I})$. Kolmogorov's approach to randomness hinges precisely on this asymmetry. Accordingly, in his approach to randomness a binary sequence D becomes increasingly nonrandom as the complexity $\varphi(D \mid \mathbf{I})$ diminishes – the smaller $\varphi(D \mid \mathbf{I})$, the more nonrandom D. In particular, by fixing a tractability bound λ, one designates as nonrandom all D satisfying $\varphi(D \mid \mathbf{I}) < \lambda$.

Kolmogorov's approach to randomness therefore confirms Persi Diaconis's dictum ("We know what randomness isn't, not what it is"); for what is random is determined strictly in relation to what is nonrandom, and not vice versa. Once we discover a short program that generates a given binary sequence, we know that its complexity is small and that it is nonrandom (cf. D_N). On the other hand, if our only programs for generating a given binary sequence are long, unless we have exhausted all programs of shorter length (something which in all but the simplest cases itself constitutes an intractable problem), we shall not know whether a significantly shorter program exists (cf. D_R). Thus, we can definitely know when a binary sequence is nonrandom; but in all but the simplest cases we can't be sure whether a sequence is random – we may simply have missed the trick that reveals the sequence to be nonrandom.[20]

This, by the way, is why the "random" numbers outputted by random number generators are properly called "pseudo-random." Indeed, as soon as we find the trick that shows how a long binary sequence can be generated by a short program, the binary sequence can no longer properly be regarded as random; on the other hand, until we do find such a trick, we give the binary sequence the benefit of the doubt and provisionally assume that it is random.[21] Randomness is thus always a provisional designation whereas nonrandomness is not. Thus, the sequence

(D_ψ)
```
0100011011000001010011100
        1011101110000000100100011
0100010101100111100010011
        0101011110011011110111100
```

which was introduced in Section 5.2 and initially appeared to be random, was seen later to be nonrandom once we saw how this sequence could be generated, namely by counting forward in base two (i.e., 0, 1, 00, 01, 10, 11, 000, 001, ...).

Except for one loose end, this summarizes Kolmogorov's approach to randomness, and recasts it in terms of the formal apparatus developed in this and the last two chapters. The loose end is this. Since programming environments are concrete computer systems, a programming environment **I** needs within our formal apparatus to be construed as side information providing a complete description of the computer system in question. Thus **I**, instead of being, say, a Macintosh computer that's running various assorted software, now becomes a complete description of this Macintosh's computer architecture together with the source code of any software running on it.

The connection between detachability and Kolmogorov's approach to randomness is now straightforward. Since any programming environment **I**, now construed as side information, can be formulated

[20] The mathematician Alan Turing used to boast that he could produce a long sequence of digits by means of a very compact program which no one with access simply to the sequence could reconstruct.

[21] See Dembski (1991, p. 87), and compare Knuth (1981, p. 27).

without reference to coin tossing, **I** will automatically satisfy the conditional independence condition. In particular, for any item of information **J** generated by **I**, $\mathbf{P}(E_R \mid \mathbf{H} \& \mathbf{J}) = \mathbf{P}(E_R \mid \mathbf{H}) = 2^{-100}$ and $\mathbf{P}(E_N \mid \mathbf{H} \& \mathbf{J}) = \mathbf{P}(E_N \mid \mathbf{H}) = 2^{-100}$. On the other hand, because of $\varphi(D_N \mid \mathbf{I}) < \lambda \leq \varphi(D_R \mid \mathbf{I})$, **I** satisfies the tractability condition only for D_N. D_N is therefore detachable from E_N, whereas D_R, at least for now, remains undetached from E_R. This makes precise what our pretheoretic intuitions have told us all along, namely, that E_N is nonrandom and E_R is random.

6

Small probability

6.1 PROBABILISTIC RESOURCES

To explicate and justify the Law of Small Probability, one key concept remains to be introduced, namely, probabilistic resources. To illustrate probabilistic resources, imagine that a massive revision of the criminal justice system has just taken place. Henceforth a convicted criminal is sentenced to serve time in prison until he flips n heads in a row, where n is selected according to the severity of the offense (we assume all coin flips are fair and duly recorded – no cheating is possible).

Thus, for a ten-year prison sentence, if we assume the prisoner can flip a coin once every five seconds (this seems reasonable), eight hours a day, six days a week, and given that the average streak of heads has length 2 ($= \Sigma_{1 \leq i < \infty} i2^{-i}$), then the prisoner will on average attempt a streak of n heads once every 10 seconds, or 6 attempts a minute, or 360 attempts an hour, or 2,880 attempts in an eight-hour work day, or 901,440 attempts a year (assuming a six-day work week), or approximately 9 million attempts in ten years. Nine million is approximately 2^{23}. Thus, if we required a prisoner to flip twenty-three heads in a row before being released, we would half the time expect to see him released within approximately ten years. Of course specific instances will vary – some prisoners being released after only a short stay, others never recording the elusive twenty-three heads!

Now consider a prisoner's reaction when after approximately ten years he finally flips twenty-three heads in a row. Is he shocked? Does he think a miracle has occurred? Certainly not. His reaction is more likely to be *It's about time!* Given the number of opportunities for observing twenty-three heads in a row, he has about an even chance of getting out of prison within ten years. There is thus nothing improbable about his getting out of prison within this span of time. It is improbable that on any given occasion he will flip twenty-three heads in a row. But when all these occasions are considered jointly, it becomes quite probable that he will be out of prison within ten years.

The underlying probability theory is straightforward. Given an event E that might occur in at least one of n trials, let E(n) denote the event that E occurs in at least one of these n trials. Suppose now that E has positive probability p, that is, $\mathbf{P}(E) = p$,[1] and that E might occur in any one of n trials. If these trials are stochastically independent (as will be the case for prisoners flipping coins), and the underlying chance mechanism is stable in the sense that probabilities don't change over time, then from elementary probability theory it follows that

$$\begin{aligned}\mathbf{P}(E(n)) &= \mathbf{P}(\text{E happens at least once in n trials}) \\ &= \sum_{1\leq i\leq n} \mathbf{P}(\text{E happens for the first time at trial i}) \\ &= \sum_{1\leq i\leq n} (1-p)^{i-1}p \\ &= 1-(1-p)^n.\end{aligned}$$[2]

Thus, given E and given n opportunities for E to occur, the question is not whether E taken by itself is improbable, but whether E remains improbable once all the opportunities for E to occur have been factored in, that is, whether E(n) – and not E – is improbable. In the case of our coin-flipping prisoner, if we let E denote the event of flipping twenty-three heads in a row, even though the probability $\mathbf{P}(E) = 2^{-23}$ appears quite small, it is $\mathbf{P}(E(9{,}000{,}000)) \approx 1/2$, which clearly is not small, that gives the prisoner hope of actually getting out of prison within his lifetime. If the prisoner's life expectancy is better than ten years, he stands a reasonably good chance of getting out of prison.

Within our revamped criminal justice system, a prisoner's probabilistic resources comprise the number of occasions (=n) he has to produce a given number of heads in a row (=k). If the prisoner must flip k heads in a row to get out of prison, and has n occasions on which to flip k heads in a row, then his probability of getting out of prison is

[1] Here, and wherever convenient in this chapter, we suppress the chance hypothesis **H** on which probabilities are conditioned. $\mathbf{P}(\cdot)$ therefore abbreviates $\mathbf{P}(\cdot \mid \mathbf{H})$.

[2] Cf. Grimmett and Stirzaker (1982, p. 38) for their discussion of the geometric distribution. Note that $\mathbf{P}(E(n))$ is p when $n = 1$ and increases to 1 as n goes to infinity.

$p(k, n) = 1 - (1-2^{-k})^n$ (cf. the preceding calculation). What's significant about this formula is that for fixed k, as n increases $p(k, n)$ converges to 1, whereas for fixed n, as k increases $p(k, n)$ converges to 0.

Probabilistic resources n and number of heads k thus offset each other. If the prisoner's probabilistic resources (= n) are big enough, it doesn't matter how many heads in a row (= k) he must obtain to get out of prison – his probabilistic resources will then be adequate for getting him out of prison. If, on the other hand, the number of heads (= k) is exorbitant, the prisoner's probabilistic resources (= n) stand little chance of getting him out of prison. Consider, for instance, a prisoner condemned to flip 100 heads in a row (= k). The probability of getting 100 heads in a row on a given trial is so small that the prisoner has no practical hope of getting out of prison, even if his life expectancy and coin-tossing ability were dramatically increased. If he could, for instance, make 10 billion attempts each year to obtain 100 heads in a row (that's coin-flipping at a rate of over 500 flips per second, twenty-four hours a day, seven days a week, for a full year), then he stands only an even chance of getting out of prison in 10^{20} years. His probabilistic resources are so inadequate for obtaining the desired 100 heads that it's pointless to entertain hopes of freedom.

In the prisoner example probabilistic resources consisted of the number of opportunities for an event to occur. We'll call this type of probabilistic resource a *replicational resource* (cf. the number of trials or replications for an event to occur). Replicational resources are not the only type of probabilistic resource. Probabilistic resources can also assume another form in which the key question is not how many opportunities there are for a given event to *occur*, but rather, how many opportunities there are to *specify* an as yet undetermined event. We'll call this other type of probabilistic resource a *specificational resource.*[3] Because lotteries illustrate specificational resources perfectly, we consider next a lottery example.

Imagine, therefore, that in the interest of eliminating the national debt, the federal government decides to hold a national lottery in

[3] The distinction between replicational resources and specificational resources will be particularly important in Section 6.5 where we set a universal probability bound that depends on the total number of specifications possible in the universe.

which the grand prize is to be dictator of the United States for a single day. That is, for twenty-four hours the winner will have full power over every aspect of government. By winning this lottery, pacifists can order the wholesale destruction of all our nuclear weapons, racists can order the mass execution of whatever races they despise, porn kings can turn our cities into giant orgies, etc. Since moderates will clearly want to prevent extremists from winning, moderates will be inclined to invest heavily in this lottery.

The following consideration, however, mitigates this natural inclination: The federal government has so constructed the lottery that the probability of any one ticket winning is 1 in 2^{100}, or approximately 1 in 10^{30}. Indeed, the lottery has been so constructed that to buy a ticket, the lottery player pays a fixed price, in this case ten dollars, and then records a bit string of length 100 – whichever string he or she chooses so long as it does not match a string that has already been chosen. Players are permitted to purchase as many tickets as they wish, subject only to their financial resources and the time it takes to record bit strings of length 100. The lottery is to be drawn at a special meeting of the United States Senate: In alphabetical order each senator is to flip a single coin once and record the resulting coin toss.

Suppose now the fateful day has arrived. A trillion tickets have been sold at ten dollars a piece. To prevent cheating, Congress has enlisted the National Academy of Sciences. In accord with the NAS's recommendation, each ticket holder's name is duly entered onto a secure database, together with the tickets purchased and the ticket numbers (i.e., the bit strings relevant to deciding the winner). All this information is now in place. After much fanfare the senators start flipping their coins, beginning with Senator Amy Aaron and ending with Senator Zig Zygmund. As soon as Senator Zygmund announces his toss, the database is consulted to determine whether the lottery had a winner. Lo and behold, the lottery did indeed have a winner – Sam Slayer, leader of the White Trash Nation. Sam's first act as dictator is to raise a swastika over the Capitol.

From a probabilist's perspective there is one overriding implausibility to this example. The implausibility rests not with the federal government sponsoring a lottery to eliminate the national debt, nor with the fascistic prize of dictator for a day, nor with the way the lottery is decided at a special meeting of the Senate, nor even with the fantastically poor odds of anyone winning the lottery. The implausibility rests with the lottery having a winner. Indeed, as a probabilist, I would encourage the federal government to sponsor such a lottery provided it could be successfully marketed to the American public, for it is obvious that if the lottery is run fairly, there will be no winner – the odds are simply too much against it. Suppose, for instance, that a trillion tickets are sold at ten dollars apiece (this would cover the national debt as it stands at the time of this writing). What is the probability that one of these tickets (qua specifications) will match the winning string of 0s and 1s drawn by the Senate? An elementary calculation shows that this probability cannot exceed 1 in 10^{18}. Even if we increase the number of lottery tickets sold by a few orders of magnitude, there still won't be sufficiently many tickets sold for the lottery to stand a reasonable chance of having a winner.

The relevant calculation for this type of probabilistic resource is as follows. Given an event E and n opportunities to specify E, let $D_1, D_2, \ldots, D_n$ be patterns that could conceivably be specifications of E (cf. Sections 5.1 to 5.4). Without loss of generality we may assume that each of these patterns derives from a single descriptive language **D**, where **D** is the formal language recursively generated from the n descriptions D_1 to D_n via the logical connectives $\sim$, &, and $\vee$. Given this assumption we may extend the correspondence * from $D_1, D_2, \ldots, D_n$ to all of **D** so that * is a homomorphism of Boolean algebras (i.e., disjunctions of descriptions become unions of the corresponding events, conjunctions become intersections, and negations become complements).

Thus, the probability that one of these patterns $D_1, D_2, \ldots, D_n$ delimits E cannot exceed the probability of the event signified by the disjunction $D_1 \vee D_2 \vee \cdots \vee D_n$, which is just the union of the events $D_1^*, D_2^*, \ldots, D_n^*$, that is,

$$(D_1 \vee D_2 \vee \cdots \vee D_n)^* = D_1^* \cup D_2^* \cup \cdots \cup D_n^*.$$

Hence the probability that one of the patterns $D_1, D_2, \ldots, D_n$ delimits E cannot exceed

$$\begin{aligned} \mathbf{P}(D_1^* \cup D_2^* \cup \cdots \cup D_n^*) &\leq \sum_{1 \leq i \leq n} \mathbf{P}(D_i^*) \\ &\leq n \left[\max_{1 \leq i \leq n} \mathbf{P}(D_i^*) \right]. \end{aligned}$$

If all the $\mathbf{P}(D_i^*)$s are the same and equal to $\mathbf{P}(E) = p$, this last term is just np.[4]

Sometimes it is necessary to consider both types of probabilistic resources in tandem, those depending on the number of opportunities for an event to occur (i.e., replicational) as well as those depending on the number of opportunities to specify a given event (i.e., specificational). Suppose, for instance, that in the preceding lottery the Senate will hold up to a thousand drawings to determine a winner. Thus, instead of having Senators Aaron through Zygmund flip their pennies in succession just once, we have them repeat this process up to a thousand times, stopping short of the thousand repetitions in case there happens to be a winner. If we now assume as before that a trillion tickets have been sold, then for this probabilistic set-up the probabilistic resources include both a trillion specifications and a thousand possible replications. An elementary calculation now shows that the probability of this modified lottery having a winner is no greater than 1 in 10^{15}. It therefore remains highly unlikely that this modified lottery, despite the increase in probabilistic resources, will have a winner.

To say that replicational resources consist of the number of separate occasions for an event to occur is not to say that the separate occasions have to be stochastically independent or identically distributed (as was the case with the coin-flipping prisoners). Consider the following example. An engineer is responsible for maintaining a certain machine. Suppose k is the number of days since the machine was last fixed, and that the probability of the machine's failure given that it's been k days since it was last fixed is $1-2^{-k}$. Thus, if the machine was fixed today ($k = 0$), there is a zero probability that the machine will break down today; if the machine was fixed yesterday ($k = 1$), there is a probability of $1/2$ that the machine will break down today; if the machine was last fixed two days ago ($k = 2$), there is a probability of $3/4$ that the machine will break down today; and so on. We assume the engineer fixes the machine as soon as it fails. If

[4] Probability measures can be defined both over Boolean algebras of events or over sentences that describe events. The formulations are not only logically equivalent, but actually isomorphic via the map *. In Chapter 3 probability measures were formulated over sentences. Here, in this calculation, it is more convenient to employ probability measures defined over Boolean algebras of events.

we now let E denote the machine breaking down on a single day and consider a block of eleven days that begins with a day on which the machine was fixed, then n = 11 constitutes a replicational resource for E, and E(11) denotes the occurrence of at least one breakdown within those days. The probability of E(11) can then be computed as follows:

$$\begin{aligned}
\mathbf{P}(E(11)) &= \mathbf{P}(\text{a failure occurs at least once in those 11 days}) \\
&= 1 - \mathbf{P}(\text{no failure occurs in those 11 days}) \\
&= 1 - (1)(1/2)(1/4)(1/8)(1/16)\cdots(1/1024) \\
&= 1 - \left[\prod_{0\le i\le 10} (2^{-i})\right] \\
&= 1 - 2^{-55}.
\end{aligned}$$

It is evident that machine failures on distinct days are not stochastically independent.

Similar complications can arise for specificational resources. Thus, to say that specificational resources consist of the number of separate opportunities to specify an as-yet undetermined event is not to say that the specifications have to specify mutually exclusive events, or have to be obtained in some prescribed manner. For instance, instead of a lottery whose specifications qua tickets are all distinct (as in the dictator-for-a-day example), we can consider a lottery whose tickets can share the same identification number (as do many state lotteries), thereby allowing multiple winners. So too, as with replicational resources, probabilistic dependencies among specificational resources can be allowed to vary.

In sum, probabilistic resources comprise the relevant ways an event can occur (replicational resources) and be specified (specificational resources) within a given context. The important question therefore is not What is the probability of the event in question? but rather What does its probability become after all the relevant probabilistic resources have been factored in? Probabilities can never be considered in isolation, but must always be referred to a relevant class of probabilistic resources. A seemingly improbable event can become quite probable when referred to the relevant probabilistic resources.[5] Anthropic principles that look to multiple universes bank on precisely this point: While the emergence of intelligent life in our universe is vastly improbable (at least by some accounts), when we factor in all the possible universes that might have given rise to us, the emergence of intelligent life is rendered a virtual certainty. Although the status of possible universes other than our own remains a matter of controversy, their role as probabilistic resources is perfectly clear (cf. Barrow and Tipler, 1986).

Similar considerations apply to the origin of life on earth. If life on earth originated spontaneously through the undirected activity of prebiotic chemicals, the relevant probability is not the probability that life could have originated spontaneously on the earth, but rather the probability that life could have originated spontaneously on at least one of the planets in the universe. From an evolutionary point of view, life on earth is a historical contingency, and need not have occurred – indeed, it might have been highly unlikely to occur on the earth as such.[6] As human beings residing on planet earth, we succumb too easily to a selection effect, mistaking the improbability of life originating on earth for the bigger probability of life originating somewhere in the universe. The earth was fortunate. Yet if the earth hadn't been so

[5] It's important not to confuse an event's probabilistic resources with its reference class. To understand the difference, consider the following example. Suppose a peasant in the days of Robin Hood hears that Sherwood Forest is teeming with deer and would like to hunt there. Unfortunately, Sherwood Forest belongs to the nobility, and so, the peasant is prohibited from hunting there. Moreover, in the public forests where the peasant may hunt, hunters have so thoroughly decimated the deer population that only a few stragglers remain. Suppose, now, the peasant goes hunting in one of the public forests and happens to shoot a deer with bow and arrow. A nobleman, who sees the peasant dragging the dead deer, claims it was felled simply by shooting an arrow at random. This seems absurd to the peasant. Nevertheless, the nobleman, who regularly hunts in Sherwood Forest, has felled many a deer with random arrows.

Who's right and who's wrong? The peasant is clearly right, and the nobleman is clearly wrong. The peasant is confined to public forests containing few deer. These forests constitute the peasant's reference class (indeed, the nobleman's private forests are irrelevant). But note, once the reference class is fixed – public forests for peasants, private forests for nobility – it is a new question entirely to ask how many arrows the peasant may end up shooting. These, of course, constitute the peasant's probabilistic resources. If the peasant could shoot a billion arrows at random, even in forests with severely depleted deer populations, it may be quite likely that one of these arrows will fell a deer. To sum up, the forests to which the peasant is confined constitute the peasant's reference class, whereas the number of arrows the peasant can shoot constitutes the peasant's probabilistic resources. In practice this difference shows up as follows: reference classes comprise what we're given (e.g., public forests in the case of peasants) whereas probabilistic resources comprise what we can do with what we're given (e.g., shoot as many arrows as possible).

[6] Though compare Kauffman (1993, ch. 7), who thinks the origin of life by undirected means, even when localized to the earth, constitutes a virtual certainty.

fortunate, fortune would have smiled on some other planet (which is not to preclude that fortune hasn't smiled on some other planet).

The number of planets therefore constitutes a probabilistic (replicational) resource for the spontaneous origin of life in the universe. Each planet provides an independent trial for life to originate spontaneously. As Richard Dawkins (1987, pp. 139, 141, 143–4) puts it,

> We can accept a certain amount of luck in our explanations, but not too much.... What is the largest single event of sheer naked coincidence, sheer unadulterated miraculous luck, that we are allowed to get away with in our theories, and still say that we have a satisfactory explanation of life?... How much luck are we allowed to assume in a theory of the origin of life on Earth?... Begin by giving a name to the probability, however low it is, that life will originate on any randomly designated planet of some particular type. Call this number the spontaneous generation probability or SGP. It is the SGP that we shall arrive at if we sit down with our chemistry textbooks, or strike sparks through plausible mixtures at atmospheric gases in our laboratory, and calculate the odds of replicating molecules springing spontaneously into existence in a typical planetary atmosphere.... [Even] if we assume ... that life has originated only once in the universe, it follows that we are *allowed* to postulate a very large amount of luck in a theory, because there are so many planets in the universe where life *could* have originated.... To conclude this argument, the maximum amount of luck that we are allowed to assume, before we reject a particular theory of the origin of life, has odds of one in N, where N is the number of suitable planets in the universe.

Dawkins's N constitutes a probabilistic resource.

Given an event E of probability p, we now let Ω denote the probabilistic resources relevant to E's occurrence. Ω then induces an event E_Ω and a probability p_Ω. The induced event E_Ω occurs just in case the original event E occurs at least once among the probabilistic resources that comprise Ω. We think of E_Ω as factoring in all the ways E might occur and be specified in relation to Ω. The induced probability p_Ω is then simply the probability of E_Ω. E_Ω (resp. p_Ω) will be called the *saturation* of E (resp. p) with respect to the probabilistic resources Ω. Alternatively, we shall speak of E and p being *saturated* by Ω to form E_Ω and p_Ω respectively, and speak generically of E_Ω as a *saturated event* and p_Ω as a *saturated probability*. If we want to draw special attention to the probabilistic resources Ω, we shall also speak in terms of *Ω-saturations* (e.g., E_Ω is an *Ω-saturated event*).

The probabilistic resources Ω can therefore be construed as a function mapping events to events and probabilities to probabilities: Ω

inputs an event E together with its probability p, and outputs the saturated event E_Ω together with its saturated probability p_Ω. Although the precise mapping from events and probabilities to their saturations is rarely made explicit, in practice there is no difficulty assigning an exact sense to the saturated event E_Ω and the saturated probability p_Ω. So long as E_Ω is an event that can be clearly identified, and has a probability p_Ω that can be accurately computed, or at least accurately bounded, using probabilistic resources to form saturations is unproblematic. Note that whenever the probabilistic resources Ω are not empty, the saturated probability p_Ω will exceed the probability p, that is, $p \leq p_\Omega$. This is because p is the probability of E happening in exactly one circumstance whereas p_Ω is the probability of E happening in that circumstance and possibly others.

6.2 THE GENERIC CHANCE ELIMINATION ARGUMENT

With the notion of a probabilistic resource in hand, we can now delineate the common pattern of reasoning that underlies chance-elimination arguments generally:

(1) A subject S learns that an event E has occurred.
(2) By examining the circumstances under which E occurred, S finds that a chance process characterized by the chance hypothesis **H** and the probability measure **P** could have been operating to produce E.
(3) S identifies a pattern D that delimits the event E.
(4) S calculates the probability of the event D* given the chance hypothesis **H**, that is, $\mathbf{P}(D^* \mid \mathbf{H}) = p$.
(5) In accord with how important it is for S to avoid a "false positive" (i.e., attributing E to something other than the chance hypothesis **H** in case **H** actually was responsible for E), S fixes a set of probabilistic resources Ω characterizing the relevant ways D* (and by implication E) might have occurred and been specified given the chance hypothesis **H**.
(6) Using the probabilistic resources Ω, S identifies the saturated event D^*_Ω and calculates (or approximates) the associated saturated probability $p_\Omega (= \mathbf{P}(D^*_\Omega \mid \mathbf{H}))$.
(7) S finds that the saturated probability p_Ω is sufficiently small.

(8) S identifies side information **I** and confirms that **I** satisfies the conditional independence condition, that is, that for any subinformation **J** generated by **I**, **J** is conditionally independent of E given **H**, that is, $\mathbf{P}(E \mid \mathbf{H} \,\&\, \mathbf{J}) = \mathbf{P}(E \mid \mathbf{H})$.

(9) With respect to a bounded complexity measure $\Phi = (\varphi, \lambda)$ that characterizes S's problem-solving capability, S confirms that D and **I** together satisfy the tractability condition, that is, that the problem of formulating the pattern D from the side information **I** is tractable, or equivalently, $\varphi(D \mid \mathbf{I}) < \lambda$.

(10) S is warranted inferring that E did not occur according to the chance hypothesis **H**.

We refer to this pattern of reasoning as the Generic Chance Elimination Argument, or GCEA for short. We unpack the GCEA in this section, and justify it in the next. Key to its justification is making precise what it means for a saturated probability to be "sufficiently small" (cf. (7)).

Since the Generic Chance Elimination Argument is explicated just as well by example as in the abstract, let us consider an illustration due to Richard Swinburne. Swinburne (1979, p. 138), in critiquing the anthropic principle,[7] relates the following story about a mad kidnapper:

> Suppose that a madman kidnaps a victim and shuts him in a room with a cardshuffling machine. The machine shuffles ten packs of cards simultaneously and then draws a card from each pack and exhibits simultaneously the ten cards. The kidnapper tells the victim that he will shortly set the machine to work and it will exhibit its first draw, but that unless the draw consists of an ace of hearts from each pack, the machine will simultaneously set off an explosion which will kill the victim, in consequence of which he will not see which cards the machine drew. The machine is then set to work, and to the amazement and relief of the victim the machine exhibits an ace of hearts drawn from each pack. The victim thinks that this extraordinary fact needs an explanation in terms of the machine having been rigged in some way. But the kidnapper, who now reappears, casts doubt on this suggestion. "It is hardly surprising," he says, "that the machine [drew] only aces of hearts. You could not possibly see anything else. For you would not be here to see anything at all, if any other cards had been drawn." But of course the victim is right and the kidnapper is wrong. There is indeed something extraordinary in need of explanation in ten aces of hearts being drawn. The

[7] For accounts of the anthropic principle see Swinburne (1979), Barrow and Tipler (1986), Hawking (1988), Leslie (1989), and Bertola and Curi (1993).

fact that this peculiar order is a necessary condition of the draw being perceived at all makes what is perceived no less extraordinary and in need of explanation.

The kidnapper explains the kidnap victim's survival by appealing to chance. The kidnap victim thinks this is absurd. Who is right? To settle this question let us recast this example as a Generic Chance Elimination Argument. We begin by identifying the kidnap victim with a subject S intent on eliminating chance. The event E, whose explanatory status is in question, is then the drawing of an ace of hearts from each of the ten decks of playing cards by the cardshuffling machine. To S's relief, S learns that E did indeed happen and thus that S's life has been spared. (Hence (1).) The mad kidnapper, however, wants to convince S that E happened by chance (contrary to (10)).

To refute the kidnapper's claim, S proceeds step by step through the Generic Chance Elimination Argument. By inspecting the cardshuffling machine, S determines that so long as there is no tampering, each of the ten decks is shuffled independently (i.e., shuffling one deck does not affect the others), and that within a deck, each card has the same probability of being drawn (we assume a standard deck of playing cards consisting of fifty-two cards, so that the probability of the ace of hearts being drawn from any one deck is 1/52). The cardshuffling device is thus a chance mechanism that operates according to a chance hypothesis **H** for which the ten decks are stochastically independent and each card from a given deck is equiprobable. (Hence (2).)

By informing S in advance precisely which event from the cardshuffling machine will save S's life (namely, the ten aces of hearts), the kidnapper provides S with a pattern that matches E. S has therefore identified a pattern D that not only delimits E, but for which $D^* = E$. (Hence (3).) Next S calculates the probability of D^* by means of the probability measure **P** and the chance hypothesis **H**, where **H** and **P** together characterize the probabilistic behavior of the cardshuffling machine. Since $D^* = E$, it follows that $\mathbf{P}(D^* \mid \mathbf{H}) = \mathbf{P}(E \mid \mathbf{H}) = (1/52)^{10}$, which is a number between 10^{-18} and 10^{-17}. (Hence (4).)

Even though S distrusts the kidnapper's claim that E happened by chance, S doesn't want to be hasty. After all, highly improbable events happen by chance all the time. What if S is not the only victim ever kidnapped by the madman. Suppose that prior to S being kidnapped, the madman kidnapped billions and billions of other hapless victims, placing them all inside identical cardshuffling machines, and that in each case the cardshuffling machine failed to deliver ten aces of hearts, thereby exploding and killing the victims. S might therefore be an incredibly lucky survivor amidst a stream of carnage. And let's not forget all the other mad kidnappers out there who didn't hook up their victims to exploding cardshuffling machines, but still subjected them to probabilistic experiments in which only an incredibly unlikely event would spare the victim. When all these other victims are factored in, might not S's luck be properly explained as a selection effect, like what happens at a lottery: Even though it's highly unlikely any one individual will win a lottery, that the lottery should have a winner is typically assured. Lottery winners are invariably surprised by their own good fortune at winning a lottery, but their sense of surprise is hardly a reason to doubt that the lottery was run fairly (i.e., that the lottery's outcome was due to chance).

Appealing to a selection effect to explain why the cardshuffling machine didn't kill S wears thin very quickly. S is a human being. Throughout recorded history, the number of humans has not exceeded a trillion, that is, 10^{12}. Even if every one of these human beings was kidnapped by mad kidnappers, placed inside exploding cardshuffling machines, and subjected to the same cardshuffling experiment as S, even with this many kidnap victims, it would still be highly unlikely for any victim to survive. Thus, even if S is supremely generous in assigning probabilistic resources for the occurrence of $D^* = E$, identifying 10^{12} possible replications of the mad kidnapper's experiment with the relevant probabilistic resources Ω – even then it remains highly unlikely that any victims will survive. As far as S is concerned, Ω easily comprises all the probabilistic resources needed to avoid a false positive – that is, attributing D^* to something other than **H** when in fact **H** was responsible for D^*. (Hence (5).)

Because separate card shuffles, whether from the same or from distinct cardshuffling machines, are stochastically independent,[8] the 10^{12} possible replications of the mad kidnapper's experiment (= S's probabilistic resources Ω) are themselves stochastically independent.

[8] By definition, a shuffle is not properly a shuffle unless it is stochastically independent from other shuffles. For the mathematics of cardshuffling see Diaconis (1988). Diaconis approaches cardshuffling through group actions.

The saturated event D^*_Ω therefore consists of at least one occurrence of ten aces of hearts by chance in the 10^{12} possible replications of the mad kidnapper's experiment. Given that $p = \mathbf{P}(D^* \mid \mathbf{H}) = (1/52)^{10} < 10^{-17}$, the saturated probability p_Ω of D^*_Ω is therefore bounded above as follows:[9]

$$\begin{aligned} p_\Omega &= \mathbf{P}(D^*_\Omega \mid \mathbf{H}) \\ &= \mathbf{P}(D^* \text{ happens at least once in } 10^{12} \text{ trials} \mid \mathbf{H}) \\ &= \sum_{1 \leq i \leq 10^{12}} \mathbf{P}(D^* \text{ happens for the first time at trial } i \mid \mathbf{H}) \\ &= \sum_{1 \leq i \leq 10^{12}} (1-p)^{i-1} p \\ &= 1 - (1-p)^{10^{12}} \\ &\approx 10^{12} \times p \quad [\text{because } \log(1-p) \approx -p \text{ for } p \text{ close to zero}] \\ &< 10^{12} \times 10^{-17} \\ &= 10^{-5}. \end{aligned}$$

At the same time

$$\begin{aligned} p_\Omega &\approx 10^{12} \times p \\ &> 10^{12} \times 10^{-18} \\ &= 10^{-6}. \end{aligned}$$

It follows therefore that $10^{-6} < p_\Omega < 10^{-5}$. This turns out to be a close enough approximation of p_Ω for S's purposes. (Hence (6).)

Is p_Ω "sufficiently small" so that if the conditional independence and tractability conditions are also satisfied, S will be warranted inferring that E did not happen according to the chance hypothesis **H**? S started with a pattern D, a pattern that matches E, and found that the probability p of D^* given **H** was bounded above by 10^{-17}, which certainly appears to be a small probability. But until the probabilistic resources relevant to an event are factored in, it is not possible to tell whether what seems like a small probability isn't actually a much larger probability that only seems small because S overlooked a selection effect. S will therefore be generous and factor in whatever probabilistic resources Ω might be relevant to the occurrence

[9] Cf. the calculation at the beginning of Section 6.1 for the event E(n), the event that E happens in at least one of n independent and identically distributed replications.

of D*. Thus, S incorporates into Ω 10^{12} stochastically independent replications of the cardshuffling experiment. But even with probabilistic resources this generous, the resulting saturated probability p_Ω remains small, that is, less than 10^{-5}. Is this small enough for S to eliminate chance in case the conditional independence and tractability conditions are also satisfied? For the moment let us say yes. A full account of what it means for a saturated probability to be "sufficiently small" will be given in the next section. (Hence – provisionally – (7).)

Although S has identified a pattern D that matches E, the question remains whether S is able to detach D from E. As it is, S has already identified the side information **I** needed to detach D from E. Indeed, before shutting S in the room with the cardshuffling machine, the mad kidnapper explicitly informed S of the precise outcome from the cardshuffling machine that prevents the machine from exploding and killing S, namely, the occurrence of ten aces of hearts. In informing S of this possible outcome, the mad kidnapper therefore supplies S with side information **I** that explicitly identifies D. **I** is therefore a prediction. Moreover, as we saw in Section 5.7, successful prediction is always a special case of specification. It follows that the conditional independence and tractability conditions are satisfied.[10] (Hence (8) and (9), respectively.)

Having successfully traversed (1)–(9), S need no longer take seriously the kidnapper's appeal to chance. Rather, S will with Richard Swinburne conclude that the kidnapper is mistaken, and that chance was not responsible for E. As Swinburne (1979, p. 138) put it, "There is indeed something extraordinary in need of explanation in ten aces of hearts being drawn. The fact that this peculiar order is a necessary condition of the draw being perceived at all makes what is perceived no less extraordinary and in need of explanation." Having satisfied (1)–(9), and in particular having fixed a stringent set of probabilistic

[10] In a sense the mad kidnapper has made things too easy for S. Even if the kidnapper had said nothing about which card draw would save S's life, simply by examining the structure and dynamics of the cardshuffling machine, S could learn that one and only one set of card draws from the cardshuffling machine will save S's life, namely, the ten aces of hearts. Having discovered that this is the only set of card draws that could save his life, S would essentially have constructed **I** from scratch, instead of simply being handed **I** by the mad kidnapper. S can therefore successfully traverse (1)–(10) of the Generic Chance Elimination Argument without depending on the mad kidnapper supplying **I**.

resources Ω, S is warranted inferring that E did not occur according to the chance hypothesis **H**. (Hence (10).)

6.3 THE MAGIC NUMBER 1/2

In unpacking the Generic Chance Elimination Argument in Section 6.2, we left a key question unanswered, namely, How small does a saturated probability p_Ω have to be for p_Ω to be "sufficiently small" (cf. (7) of the GCEA)? The aim of this section is to answer this question, and thereby justify the GCEA. In answering this question, we need first to be clear that the question itself is well-posed. For at first blush, the very idea of a saturated probability being sufficiently small seems incoherent. Indeed, in regarding a saturated probability as sufficiently small, a subject S seems to walk right into an infinite regress.

For consider a subject S proceeding step by step through the GCEA. Starting with an event E, S identifies a pattern D that delimits E, and thereupon computes $p = \mathbf{P}(D^* \mid \mathbf{H})$. Having fixed a set of probabilistic resources Ω, S next identifies the corresponding saturated event D^*_Ω and calculates the corresponding saturated probability $p_\Omega = \mathbf{P}(D^*_\Omega \mid \mathbf{H})$. In determining whether to eliminate the chance hypothesis **H**, S uses the probabilistic resources Ω to induce a "second-order probability" p_Ω which – if sufficiently small – would serve to eliminate chance in explaining the "first-order events" E and D^*. But why should S stop with second-order probabilities and second-order events like p_Ω and D^*_Ω? Having employed the probabilistic resources Ω to induce second-order probabilities and events, what is to prevent S from introducing still further probabilistic resources that induce still higher order probabilities and events?

The basic intuition underlying probabilistic resources is that an event has small probability only if its probability remains small after all the probabilistic resources relevant to its occurrence have been factored in. But this way of stating the basic intuition seems to make an infinite regress unavoidable, for why not treat D^*_Ω as an event that in turn needs probabilistic resources relevant to its occurrence factored in, say the probabilistic resources Ω', thereby inducing a "third-order event" $[D^*_\Omega]_{\Omega'}$ together with a "third-order probability" $[p_\Omega]_{\Omega'}$? Unless this regress can be stopped – and stopped quickly – the GCEA will founder and the concept of a small probability will remain elusive, depending on ever higher orders of probability.

Fortunately, this regress can always be avoided by collapsing successive saturations by multiple sets of probabilistic resources into a single saturation by a grand set of probabilistic resources. To see this, consider a human subject S who possesses a fair coin and wants to know whether by tossing this coin, a hundred heads in a row might reasonably be expected by chance. S therefore starts by fixing as probabilistic resources Ω a trillion replications on which S can attempt to obtain a hundred heads in a row. Having factored in these trillion replications, S notes that obtaining a hundred heads in a row still seems incredibly unlikely. S therefore takes the (higher order) event of getting a hundred heads in a row over the course of these trillion replications, and factors in the further probabilistic resources Ω' consisting, let us say, of a billion persons each with the ability to produce a trillion replications.

Now the point to realize is that factoring in first Ω and then Ω' into the occurrence of a hundred heads in a row (i.e., first a trillion replications and then a billion persons each responsible for a trillion replications), is equivalent to simply factoring in a billion trillion replications from the start. This maneuver is entirely typical of probabilistic resources. To factor a set of probabilistic resources Ω into an event, thereby obtaining an event saturated by Ω, and then to factor in still a further set of probabilistic resources Ω', thereby obtaining an event successively saturated by Ω and Ω', is equivalent to factoring in from the start a grand set of probabilistic resources that incorporate both Ω and Ω'.

Thus, rather than cumulating probabilistic resources piecemeal, S should from the start factor in all the probabilistic resources that according to S's interests and needs are going to be relevant to eliminating chance. The precise form and extent of these probabilistic resources will be determined contextually, admitting no hard and fast rules, and depending principally on how stringently S needs to control for false positives, that is, to control for mistakenly eliminating chance when in fact chance was operating. When someone's life is in the balance, as in a criminal case involving a capital offense, S will want to control false positives much more stringently than, say, in social sciences research, where the danger of false positives is always mitigated because, at least in principle, the scientific enterprise is self-correcting. In the case of capital punishment, there is no opportunity to rectify a false positive once the sentence is carried out.

The scientific literature, on the other hand, can always be challenged, critiqued, and corrected.

S therefore has to think carefully about what set of probabilistic resources to use in eliminating chance. Moreover, having fixed a set of probabilistic resources Ω, S better not introduce a further set of probabilistic resources Ω' later down the line. Rather, any such Ω' should be incorporated into the initial set of probabilistic resources Ω right from the start. Once S fixes Ω, all the relevant probabilistic resources have to be there. Any regress beyond a first-order saturated event and first-order saturated probability indicates poor planning on S's part.

But once S has fixed a relevant set of probabilistic resources Ω, how is S to assess whether the saturated probability in (7) of the GCEA (i.e., p_Ω) is indeed "sufficiently small" to warrant the elimination of chance in (10) of the GCEA (provided of course that the rest of the GCEA is satisfied)? First off, let us be clear that the elimination of chance in (10) is never a logical deduction from (1)–(9) unless p_Ω equals zero, in which we case we don't need the GCEA to tell us that the elimination of chance is warranted, for events of zero probability have by definition no chance of happening. But if p_Ω is strictly greater than zero, the possibility always exists that E did happen by chance. The relation between (1)–(9) and (10) in case $p_\Omega > 0$ is therefore never a logical entailment; rather, the possibility always remains that it was a mistake to eliminate chance.

To say that a pattern of reasoning is fallible, however, is not to say it leads us astray. Indeed, a fallible pattern of reasoning may be so constitutive of rationality that to contradict its deliverances would be to go astray – even when its deliverances are mistaken. An example due to C. S. Peirce (1878 [1988], p. 1313) illustrates this point:

> If a man had to choose between drawing a card from a pack containing twenty-five red cards and a black one, or from a pack containing twenty-five black cards and a red one, and if the drawing of a red card were destined to transport him to eternal felicity, and that of a black one to consign him to everlasting woe, it would be folly to deny that he ought to prefer the pack containing the larger portion of red cards.

The "mistake" of picking a black card has a far greater likelihood if the man draws a card at random from the predominantly black deck as opposed to the predominantly red deck. Given a choice of decks and the aim to avoid drawing a black card, the rational thing to do according to Peirce is to choose the predominantly red deck. To choose the predominantly black deck, even if one is lucky enough to draw a red card, is folly. Alternatively, to choose the predominantly red deck, even if one is unlucky enough to draw a black card, remains sound.

Why? Underlying all such claims about what constitutes the rational thing to do within a given probabilistic set-up is a collection of likelihood principles. These principles regulate the rational use of probabilities, apportioning our beliefs and conforming our actions to the estimated likelihood of events. These principles are so fundamental for regulating how finite rational agents like ourselves navigate through life's uncertainties that even basic texts on probability and statistics frequently omit stating them. For instance, the likelihood principle underlying Peirce's example is the following:

L1. *Suppose a subject S desires that an event E happen. Suppose further that to make E happen, S must select one from among n chance processes characterized by the chance hypotheses* $\boldsymbol{H}_1$, $\boldsymbol{H}_2$, ..., $\boldsymbol{H}_n$. *Then S should select a chance process whose chance hypothesis* $\boldsymbol{H}_k$ *renders E most probable, that is,* $\boldsymbol{H}_k$ *satisfies* $\boldsymbol{P}(E \mid \boldsymbol{H}_k) = max_{1 \leq i \leq n} \boldsymbol{P}(E \mid \boldsymbol{H}_i)$.

So too, a likelihood principle underlies the GCEA. This principle regulates when a saturated probability p_Ω is sufficiently small to warrant eliminating chance. If p_Ω in (7) of the GCEA is strictly greater than zero, the possibility of mistakenly eliminating chance in (10) can never be entirely precluded. Nevertheless, this possibility of error diminishes as p_Ω gets smaller. How small is small enough? We can turn this question around: What is the largest positive real number ω such that whenever $p_\Omega < \omega$, p_Ω satisfies (7) of the GCEA (i.e., how big can a saturated probability be, and still be sufficiently small)?

Without an "ω-level" like this to decide when a saturated probability p_Ω is sufficiently small, it will be impossible to justify the GCEA. For without such an ω-level, any subject S who wants on the basis of (1)–(9) to infer that the event E did not happen by chance always faces the charge that p_Ω isn't quite small enough. And how is S to respond to this charge, for S knows that the smaller p_Ω is in (7), the smaller is the possibility of error in (10)? p_Ω may be small enough as far as S is concerned to warrant eliminating chance in (10). But what about convincing someone else? A caviller can always challenge S by claiming that if p_Ω had only been smaller, it would have been legitimate to eliminate chance – but alas, p_Ω wasn't quite small enough.

What, then, is the largest positive real number ω such that whenever $p_\Omega < \omega$, p_Ω satisfies (7) of the GCEA? The answer is $\omega = 1/2$. This value follows from the following likelihood principle:

L2. *Suppose a chance process characterized by the chance hypothesis **H** produces an event E that as yet is unknown to a subject S. Suppose further that S has explicitly identified n events $F_1, F_2, \ldots, F_n$, and is going to conjecture that one of these events subsumes E. Then S should conjecture any F_k having maximal probability, that is, any F_k such that $\mathbf{P}(F_k \mid \mathbf{H}) = max_{1 \leq i \leq n} \mathbf{P}(F_i \mid \mathbf{H})$.*

To see how L2 works, consider a modification of Peirce's card-drawing example. Suppose this time one is given just a single deck of playing cards consisting of twenty-five red cards and a single black card. Suppose further one is forced to conjecture either of two events – either the event F_1 that a red card will be chosen at random, or the event F_2 that a black card will be chosen at random. Suppose that conjecturing the event that happens transports one to eternal felicity, but that conjecturing incorrectly consigns one to everlasting woe. Which event should one conjecture? The answer is clearly F_1, the event that a red card will be drawn (the probability of F_1 equals 25/26 whereas the probability of F_2 equals 1/26).

That ω equals 1/2 is now a corollary of L2. For suppose a subject S is given that a chance process characterized by the chance hypothesis **H** produces an event E that as yet is unknown to S. Suppose further that S has explicitly identified a single event F. Since S has explicitly identified F, S knows the conditions under which not only F, but also its complement F^c occurs. The event F, once explicitly identified, thus induces two events – the original event F, and the complementary event F^c. Because F and F^c are mutually exclusive and exhaustive, it follows from the axioms of probability theory (cf. Section 3.6) that

$$\mathbf{P}(\mathrm{F} \cup \mathrm{F}^c \mid \mathbf{H}) = \mathbf{P}(\mathrm{F} \mid \mathbf{H}) + \mathbf{P}(\mathrm{F}^c \mid \mathbf{H}) = 1,$$

and therefore that $\mathbf{P}(\mathrm{F} \mid \mathbf{H})$ and $\mathbf{P}(\mathrm{F}^c \mid \mathbf{H})$ both equal 1/2, or that one of these probabilities is strictly less than 1/2 and the other strictly greater than 1/2. In particular, $\mathbf{P}(\mathrm{F} \mid \mathbf{H}) < 1/2$ if and only if $\mathbf{P}(\mathrm{F} \mid \mathbf{H}) < \mathbf{P}(\mathrm{F}^c \mid \mathbf{H})$. The following likelihood principle is therefore an immediate corollary of L2:

L3. *Suppose a chance process characterized by the chance hypothesis **H** produces an event E that as yet is unknown to a subject S. Suppose further that S has explicitly identified a single event F for which* $\mathbf{P}(F \mid \mathbf{H}) < 1/2$. *Then in conjecturing that either F or* F^c *subsumes E, S should reject F and conjecture* F^c.

L3 in combination with the GCEA justifies setting ω equal to 1/2. In the GCEA a subject S learns that an event E has occurred and notes that E could conceivably have occurred according to the chance hypothesis **H** (cf. (1) and (2)). In learning of E's occurrence, S explicitly identifies E. Nevertheless, because S has produced a pattern D (cf. (3)) that is detachable from E (cf. (8) and (9)), S can work exclusively with this pattern and bracket the fact that E has already been identified. S can therefore treat E as an as-yet-unknown event produced by **H**.[11] Having produced the pattern D, and for now treating E as a free-floating, unidentified chance event due to **H**, S now proceeds through steps (4)–(7) of the GCEA. Thus, S calculates the probability of the event D* given the chance hypothesis **H**, that is, $\mathbf{P}(\mathrm{D}^* \mid \mathbf{H}) = \mathrm{p}$ (cf. (4)). Next S identifies probabilistic resources Ω that characterize the relevant ways D* (and by implication E) might have occurred and been specified given the chance hypothesis **H** (cf. (5)). Next S uses Ω to identify the saturated event D^*_Ω and calculates the associated saturated probability $\mathrm{p}_\Omega = \mathbf{P}(\mathrm{D}^*_\Omega \mid \mathbf{H})$ (cf. (6)).

Suppose now that S lets D^*_Ω correspond to F in L3 and discovers from (6) that $\mathrm{p}_\Omega = \mathrm{p}(\mathrm{D}^*_\Omega \mid \mathbf{H}) = \mathbf{P}(\mathrm{F} \mid \mathbf{H}) < 1/2$. Because S is still

[11] This move was justified in Section 5.3. To say that D is detachable from E is to say that D can be formulated on the basis of side information **I** that in turn is conditionally independent of E. Since conditional independence models epistemic independence, to say that **I** is conditionally independent of E is to say that **I** can be known independently of E. But since D can be formulated on the basis of **I**, it follows that D can be known independently of E as well. Hence the permission to treat E as an unknown, unidentified, indeterminate event.

treating E as an unknown event, in trying to subsume E under either F or F^c, S is going to choose F^c over F. Without knowing any specifics about E, S will conjecture that F^c rather than F subsumes E. But according to (3), D delimits E, which is equivalent to D^* subsuming E. Moreover, since factoring probabilistic resources into an event can only expand an event, the saturated event $F = D^*_\Omega$ must also subsume E (cf. Section 6.1). S is therefore confronted with two conflicting claims: On the one hand, because D is detachable from E, S is entitled to bracket the identity of E, and infer that if **H** was operating, then it is more likely that F^c rather than F occurred. Thus, as long as the identity of E is bracketed, the rational course to take according to L3 is to suppose that F^c and not F occurred. On the other hand, by not bracketing the identity of E, S finds that F, and not F^c, subsumes the event that actually occurred (i.e., E). The only way out of this conflict is to infer that **H** was never responsible for E in the first place (cf. (10)).

Although the conflict just described does not constitute a logical inconsistency, it does constitute a probabilistic inconsistency – and one that needs to be resolved. What exactly needs to be resolved in a probabilistic inconsistency? Certainly, situations like the following do not engender a probabilistic inconsistency: A subject S decides to flip a fair coin twice, issuing in an event E; before flipping the coin, S explicitly identifies the event F, consisting of two heads in a row; upon flipping the coin S discovers that E coincides with F, that is, that two heads in a row were indeed flipped. Now the fact that F^c has probability 3/4, and therefore exceeds the probability of F, whose probability is 1/4, cuts no ice in eliminating chance here. Until S learns that the event E has happened, S will be guided by L3, preferring to think that F^c will happen. But once S finds that E actually did happen and that E coincides with F, S immediately dispenses with F^c (and rightly so), without ever doubting that E occurred by chance. Yet in the GCEA, S dispenses not with F^c (whose probability is also bigger than the probability of F), but with the chance hypothesis **H**. Why is there a conflict that needs to be resolved in the GCEA (specifically, by eliminating chance), but not in this coin-tossing example? What is the difference?

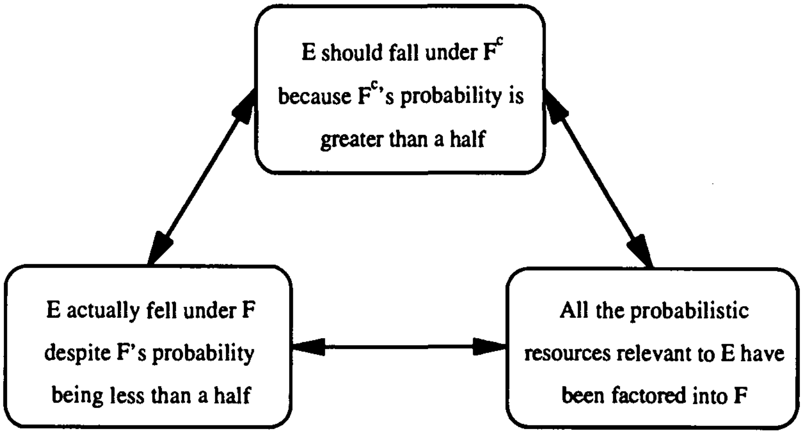
The difference is that in the coin-tossing example S failed to factor in the probabilistic resources relevant for heads to occur twice in a row. Any ordinary human subject S will flip hundreds if not thousands of coins during the course of a lifetime. Once these probabilistic resources are factored in, they render the occurrence of two heads in a row highly probable – certainly better than a probability of 1/2. But suppose instead that a human subject S is going to flip a coin not twice, but two hundred times. The event E that will occur as a result of these two hundred flips is as yet unknown to S. Yet prior to flipping the coin two hundred times, S identifies the event consisting of two hundred heads in a row – call this event R (R corresponds to D^* in the GCEA). S also notes that if all human beings that ever existed were continuously engaged in flipping coins for every moment of their waking existence, the event F that at least one of these human beings flips two hundred heads in a row would have a minuscule probability, certainly less than 1/2 (F equals the saturation of R by all the occasions on which humans might flip coins, and thus corresponds to D^*_Ω in the GCEA). F^c therefore has probability greater than 1/2, and in line with L3 S will think that if chance is responsible for E, then E will fall under F^c rather than F (in particular, E will not be expected to coincide with R).

But what if S now learns that E coincides with two hundred heads in a row (i.e., $E = R$), and thus falls under F? As in the previous example, where the coin was only flipped twice, S can shift gears, repudiate F^c, and continue attributing E to chance. Though this move does not entail a logical inconsistency (unless E has zero probability), it does entail a probabilistic inconsistency. By a probabilistic inconsistency I mean a threefold conflict between

(1) what the likelihood principle L3 tells us to expect if chance were responsible for E, viz., that E should fall under F^c, whose probability is strictly greater than 1/2;
(2) the observation that E did in fact fall under F, and this despite F being identified without knowledge of E; and
(3) the fact that all the probabilistic resources relevant to E's occurrence have been factored into F.

Whereas for two heads in a row only the first two of these conditions were satisfied, for two hundred heads in a row all three conditions were satisfied.

Schematically we can represent a probabilistic inconsistency as follows (in line with L3, this schematic assumes that F was identified without knowledge of E):



Probabilistic inconsistencies are the probabilistic equivalent to logical inconsistencies. Both lead to a reductio ad absurdum. Both demand resolution. Both are resolved by eliminating an offending hypothesis. In the case of probabilistic inconsistencies, the offending hypothesis is chance.

To sum up, a saturated probability p_Ω is sufficiently small just in case it is strictly less than a half. Item 7 of the GCEA can therefore be rewritten as follows:

(7) S finds that the saturated probability p_Ω is strictly less than 1/2.

Given that this is what it means for a saturated probability to be sufficiently small, we can in turn define what it means for an unsaturated probability p to be small relative to a set of probabilistic resources:

Definition. *For an event E and a chance hypothesis* $\mathbf{H}$*, the probability* $p = \mathbf{P}(E \mid \mathbf{H})$ *is* **small** *relative to a set of probabilistic resources* Ω *just in case the saturated probability* $p_\Omega = \mathbf{P}(E_\Omega \mid \mathbf{H})$ *is strictly less than a half.*

A probability is therefore never small *simpliciter*, but small only in relation to a set of probabilistic resources Ω. Given this definition, we can combine (6) and (7) of the GCEA, and rewrite it more compactly as follows:

(6–7) S finds that the probability $p = \mathbf{P}(D^* \mid \mathbf{H})$ is small relative to Ω.

In defining a probability as small whenever its corresponding saturated probability is strictly less than a half, we answer a long-standing question in the logic of statistical decision theory. Given a chance hypothesis **H** and a rejection region R, how small does the probability of R given **H** have to be (i.e., how small **P**(R | **H**)?) so that if an event E falls within R (i.e., if R subsumes E), then the chance hypothesis **H** can be legitimately rejected? This is the key conceptual difficulty that to this day remains unresolved in Ronald Fisher's theory of statistical significance testing. The problem is to justify a "significance level" α (always a positive real number less than one) such that whenever E falls within R and **P**(R | **H**) $< \alpha$, then the chance hypothesis **H** can be legitimately rejected as the explanation of E. To date the problem has been that any proposed value for α has seemed arbitrary, lacking what Howson and Urbach (1993, p. 178) call "a rational foundation."

Specifically, Howson and Urbach (1993, pp. 178–80) direct the following criticism against Fisher's theory:

> Fisher seems to have seen in significance tests some surrogate for the process of refutation, and he frequently went so far as to say that a theory can be "disproved" in a significance test (e.g., Fisher, 1947, p. 16) and that such tests, "when used accurately, are capable of rejecting or invalidating hypotheses, in so far as these are *contradicted by the data*" (Fisher, 1935; italics added). If Fisher intended to imply that tests of significance can demonstrate the falsity of a statistical theory, and it is difficult to see what else he could have meant, then clearly he was wrong.... Fisher seems to be saying that statistical hypotheses may be actually falsified; but the experimental results used in a significance test clearly do not logically contradict [*n.b.*] the null hypothesis.
>
> Fisher was, of course, aware of this, and when he expressed himself more carefully, his justification for significance tests was rather different. The force of a test of significance, Fisher then claimed, "is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution [i.e., the null hypothesis] is not true" (Fisher, 1956, p. 39). But in thus avoiding an unreasonably strong interpretation, Fisher plumped for one that is unhelpfully weak, for the significant or critical results in a test of significance are by definition improbable, relative to the null hypothesis. Inevitably, therefore, the occurrence of a significant result is either a "rare chance" (an improbable event) or the null hypothesis is false, or both. And Fisher's claim amounts to nothing more than this necessary truth. It certainly does not allow one to infer the truth or falsity of any statistical hypothesis from a particular result....

Expositions of Fisherian significance tests typically vacillate over the nature of the conclusions that such tests entitle one to draw. For example, Cramér said that when a hypothesis has been rejected by such a procedure, "we consider the hypothesis is disproved" (1946, p. 334). He quickly pointed out, though, that "this is, of course, by no means equivalent to a *logical* disproof." Cramér contended, however, that although a rejected theory could in fact be true, when the significance level is sufficiently small, "we *feel* practically justified in disregarding this possibility" (original italics altered). No doubt such feelings do often arise; ... but Cramér offered no grounds for thinking that such feelings, when they occur, were generated by the type of reasoning employed in tests of significance, nor was he able to put those feelings onto any systematic or rational basis.

I submit that the probabilistic apparatus developed in the last three sections provides just such a systematic and rational basis for why we can "feel practically justified" eliminating chance in a test of statistical significance. As Howson and Urbach rightly note, there is never a logical contradiction in refusing to eliminate chance in a test of statistical significance – save in the case where the rejection region R has probability zero with respect to **H** and the event E happens to fall within R. Short of this, it is always a logical possibility that E is a chance outcome due to the chance hypothesis **H**, regardless whether E falls within or outside the rejection region R. But precisely because this is a strictly logical point, one does not have to be a statistician to appreciate it. Indeed, it is precisely the statistician's task to tell us how to explain E when the rejection region does not have zero probability.

Is there, then, a systematic and rational basis by which the statistician can properly explain E in reference to **H** when the rejection region R does not have zero probability? The Generic Chance Elimination Argument provides just such a basis. In Fisher's theory, a statistician/subject S is justified eliminating **H** as the explanation of E whenever (1) E falls within a rejection region R, and (2) the probability of R given **H** is sufficiently small (i.e., $\mathbf{P}(R \mid \mathbf{H}) < \alpha$ for some α-level). Although this is certainly part of the story, it is not the whole story. The whole story consists in taking Fisher's account of statistical significance testing, and embedding it in the GCEA, whose systematic and rational basis has been established.

The whole story therefore looks like this. S is about to witness an event E that provisionally will be taken to derive from a chance process characterized by the chance hypothesis **H** (hence (1) and (2) of the GCEA). Prior to witnessing E, S designates a pattern D that will serve as a rejection region (hence (3)). Specifically, the rejection region R will be the event D* corresponding to the pattern D. Because S has identified D prior to the occurrence of E, S is able to detach D from E (hence (8) and (9); recall that rejection regions are predictions – cf. Section 5.7). S now factors in the probabilistic resources Ω that S regards as relevant to S's needs and interests for guarding against mistakenly eliminating **H** in case **H** actually was operating to produce E (hence (4) and (5)). Factoring Ω into the rejection region R induces the saturated event $R_\Omega (= D^*_\Omega)$. S now finds that $\mathbf{P}(R_\Omega \mid \mathbf{H})$ is strictly less than 1/2, and thus that $\mathbf{P}(R \mid \mathbf{H})$ has small probability relative to Ω (hence (6–7); cf. Section 6.3). With (1)–(9) of the GCEA satisfied, S is warranted inferring that E did not occur according to the chance hypothesis **H** (hence (10)).

From this rational reconstruction of statistical significance testing, it's now clear why the significance levels α of 0.05, 0.01, and the like that regularly appear in the applied statistics literature (e.g., the social sciences literature) are arbitrary. α is supposed to bound the probability of the rejection region R, that is, $\mathbf{P}(R \mid \mathbf{H}) < \alpha$. As a rejection region, R is going to be employed to eliminate the chance hypothesis **H** in explaining the observed event E provided E falls within R. The problem is that α-levels like 0.05 and 0.01 are typically instituted without any reference to the probabilistic resources relevant to controlling for false positives. (As usual, a false positive is the error of eliminating **H** as the explanation for E when **H** actually is responsible for E. Statisticians refer to false positives as "type I errors.") Suppose, for instance, an academic journal in the social sciences institutes an α-level of 0.01 to control for false positives. In this case, articles that record what would be an interesting theoretical result so long as the result is not due to chance are accepted for publication only if the result falls inside a rejection region whose probability is 0.01 or less. Any number of journals in experimental psychology, for instance, require an α-level of 0.01 for submission of manuscripts.

But what does such an α-level accomplish? In general, for every hundred experiments in which chance actually was operating, on average one experiment will satisfy an α-level of 0.01 (as is customary, we assume that separate experiments are stochastically independent). Thus, for a journal requiring an α-level of 0.01, for every hundred experiments conducted by researchers in the journal's field of specialization, and in which chance actually was operating, on average one in a hundred of those experiments will slip through the cracks and form the basis of an article acceptable to that journal. Does an α-level of 0.01 therefore provide stringent enough controls on false positives for such a journal? The answer depends on the number of experiments conducted by researchers in the field who submit their findings to the journal. As this number increases, the number of experiments in which chance actually was operating, but for which chance will be (falsely) eliminated, will increase, thus increasing the number of false positives that could conceivably slip into the journal. A journal requiring an α-level of 0.01 will on average allow one in a hundred of such experiments into its pages. More generally, a journal requiring a significance level α will on average allow the proportion α of such experiments into its pages. The more experimental research there is in the journal's field of specialization, the more likely the journal is to include false positives. The relevant probabilistic resource here is therefore the replicational resource consisting of the number N of separate experiments performed by researchers in the journal's field of specialization.

Although the amount of research activity in the journal's field of specialization determines the number of false positives that could conceivably slip into the pages of the journal, in practice the precise α-level a journal sets to control for this error will depend on the needs and interests of the editors of the journal. If the field to which the journal is directed is good about correcting past mistakes, there may be no problem setting the α-level high (e.g., $\alpha = 0.05$). On the other hand, if the field to which the journal is directed is bad about correcting past mistakes, it will probably be a good idea to set the α-level low (e.g., $\alpha = 0.0001$) so that what errors do make it into the journal will be few.

Nevertheless, from the vantage of the GCEA, the choice of any such α-level is not arbitrary, but obtains a definite meaning only in reference to probabilistic resources. Given an event E, a rejection region R, a chance hypothesis **H**, and a significance level α, Fisher's theory does little more than press the fundamental intuition that **H** ought not to be viewed as responsible for E so long as E falls within R and $\mathbf{P}(R \mid \mathbf{H})$ is small (meaning $\mathbf{P}(R \mid \mathbf{H}) < \alpha$). But how small is small enough? The account of small probabilities given in Section 6.3 provides the answer. In the case of a journal that requires a significance level of size α, α will be small provided that the probability of a false positive worming its way into the journal remains strictly less than 1/2 once all the probabilistic resources relevant to E's occurrence are factored in. But since in this case probabilistic resources consist of N stochastically independent experiments governed by the chance hypothesis **H**, for α to be small means that for N stochastically independent events $R_1, R_2, \ldots, R_N$ each with probability α, the following inequality has to be satisfied:

$$\mathbf{P}(R_1 \text{ or } R_2 \text{ or } \cdots \text{ or } R_N \mid \mathbf{H}) < 1/2.$$

An elementary probability calculation now shows that N must be bounded as follows (the logarithm here and throughout is the natural logarithm):

$$N < \frac{\log 1/2}{\log(1-\alpha)} \approx \frac{\log 2}{\alpha}.$$

Thus, from the vantage of the GCEA, an α-level in a test of statistical significance is small enough to eliminate chance for an event falling inside a rejection region so long as the total number N of stochastically independent opportunities for the event to occur (= probabilistic resources) is less than $(1/\alpha)(\log 2)$. Alternatively, once N (= probabilistic resources) is fixed, for an α-level to constitute a small probability, the following inequality must be satisfied:

$$\alpha < \frac{\log 2}{N}.$$

6.5 LOCAL AND UNIVERSAL SMALL PROBABILITIES

There are two types of small probabilities, local and universal. To understand the difference, consider the following story about Alice, Bob, and Cathy. Imagine that Alice is about to become the editor of a new journal in experimental psychology. In setting editorial policies for her journal, Alice must institute an α-level to control for false positives (i.e., type I errors) slipping into her journal. How is Alice going to choose an α-level? Since Alice is schooled in the logic of small probabilities, Alice is not simply going to copy the α-level of some existing journal in her field. Rather, she will select a set of probabilistic resources Ω commensurate with her needs and interests for controlling false positives, and she will then determine the largest probability p such that for a rejection region R, if $\mathbf{P}(R \mid \mathbf{H}) = p$, then $\mathbf{P}(R_\Omega \mid \mathbf{H}) = p_\Omega < 1/2$. The largest such p will then constitute the α-level of her journal.[12]

To select the probabilistic resources Ω relevant to her journal, Alice adopts the following rationale. Because her journal is highly specialized, Alice determines that in a given year no more than 600 researchers are performing experiments that might enter her journal. Moreover, given the rate at which these experiments can be performed, Alice determines that in a given year a given researcher will be unable to conduct more than 20 experiments that might enter her journal. Finally, in reviewing the longevity of most journals historically, Alice determines that her journal will be lucky to survive 100 years without folding. Thus, she reasons that throughout the journal's existence, there will be no more than $N_A = 600 \times 20 \times 100 = 1.2 \times 10^6$ separate experiments that might enter her journal.

To ensure that the probability of a false positive slipping into Alice's journal remains less than a half, Alice therefore employs the inequality at the end of Section 6.4, and calculates a significance level $\alpha_A = (\log 2)/(1.2 \times 10^6) \approx 5.77 \times 10^{-7}$. So long as all the articles submitted to Alice's journal set the probabilities of their rejection regions below 5.77×10^{-7}, the probability that even a single false positive will slip into Alice's journal during its projected 100 years remains less than a half. Alternatively, it's more likely than not that throughout its lifetime Alice's journal will be free from false positives.

To continue our story, imagine that several years have passed since the inception of Alice's journal, and that Bob now comes along while Alice's journal is in full swing. Bob too is an experimental psychologist schooled in the logic of small probabilities. In perusing the editorial policies of Alice's journal, Bob finds that Alice has fixed an α-level of 5.77×10^{-7}. Moreover, from reading the fine print on the inside cover page of Alice's journal, Bob learns the rationale underlying Alice's choice of α. Upon further reflection, however, Bob is troubled. Granted, Alice's choice of α guarantees it's more likely than not that her journal will be free of false positives throughout the journal's lifetime. But what about all the other journals Bob employs in his research? Bob estimates there are about 200 journals he needs to consult professionally. It may be fine for Alice that her journal

[12] More precisely, α will equal the least upper bound of the set of real numbers $\{p$: for all rejection regions R such that $\mathbf{P}(R \mid \mathbf{H}) = p, \mathbf{P}(R_\Omega \mid \mathbf{H}) = p_\Omega < 1/2\}$.

has a better than even chance of excluding false positives throughout its lifetime. Bob, however, wants all the journals he is ever going to consult to have a better than even chance of excluding false positives throughout their lifetimes.

For Bob's purposes, therefore, an α-level of 5.77×10^{-7} is not going to be small enough. What then is going to be an appropriate α-level for Bob? To select probabilistic resources Ω that induce an appropriate α-level, Bob adopts a rationale similar to Alice's. But whereas Alice determined that no more than 600 researchers could conceivably submit articles to her journal in a given year, surveying the field of experimental psychology as a whole, Bob determines that in a given year up to 90,000 researchers could conceivably submit articles to the 200 journals Bob uses for his own research. Whereas Alice determined that the rate at which research is conducted in her field of specialization precludes a researcher from conducting more than 20 experiments in a given year, Bob determines that research in some areas of experimental psychology permits experiments to be conducted at the faster rate of up to 40 experiments per year. Finally, in reviewing the longevity of most journals historically, Bob determines that even though most journals will be lucky to survive a hundred years without folding, a few of the journals he consults are well into their second century. Thus, Bob will want to keep false positives from entering the journals he consults not just for 100 years, as did Alice, but for 200 years.

Bob therefore calculates that no more than $N_B = 90{,}000 \times 40 \times 200 = 7.2 \times 10^8$ separate experiments can conceivably enter the journals he consults. Corresponding to N_B, Bob calculates a significance level $\alpha_B = (\log 2)/(7.2 \times 10^8) \approx 9.63 \times 10^{-10}$ (cf. the inequality at the end of Section 6.4). If it were up to him, Bob would require that all the articles submitted to the journals he consults set the probabilities of their rejection regions below 9.63×10^{-10}. Given this α-level, the probability of even a single false positive slipping into any of these journals would remain strictly less than a half.

Finally, consider Bob's friend Cathy. Cathy has observed Bob's crusade to institute more stringent α-levels for the journals he consults (in Bob's case, journals related to experimental psychology). Cathy is concerned, however, where Bob's crusade will end. Cathy too is schooled in the logic of small probabilities. But though trained as an experimental psychologist, Cathy no longer works in experimental psychology per se. Rather, Cathy belongs to an interdisciplinary think-tank whose researchers revel in crossing disciplinary boundaries. Unlike Bob, Cathy and her think-tank are therefore not limited to a mere 200 journals. Cathy and her think-tank are free to dip into any of the 30,000 or so journals that constitute the scientific literature worldwide.

If Cathy now adopts Bob's rationale for selecting probabilistic resources Ω, to keep false positives from entering the journals available to her, Cathy will need to expand Bob's probabilistic resources considerably. Thus, whereas Bob determined that no more than 90,000 researchers could conceivably submit articles to the journals he consults in a given year, by surveying the totality of scientific research, Cathy determines that in a given year as many as 10,000,000 researchers could conceivably submit articles to the 30,000 journals available to her. Whereas Bob determined that the rate at which research is conducted in his field precludes a researcher from conducting more than 40 experiments in a given year, Cathy determines that research with computer simulations permits experiments to be conducted at much faster rates – with present computational speeds even one per minute is not inconceivable. Thus, to be safe Cathy allows as many as $60 \times 24 \times 365 \approx 500{,}000$ experiments per researcher per year. Finally, despite the relatively recent inception of most journals historically, Cathy decides to do Bob one better and allow that a journal might survive a millennium. Thus, Cathy will try to keep false positives from entering the journals available to her not just for a two centuries, as did Bob, but for 1000 years.

Cathy therefore concludes that $N_C = 10{,}000{,}000 \times 500{,}000 \times 1{,}000 = 5 \times 10^{15}$ sets an upper bound on the number of separate experiments that might enter the journals available to her. It follows that if Cathy were to adopt Alice's and Bob's rationale, Cathy would set her probabilistic resources to $N_C = 5 \times 10^{15}$ experiments, and then compute a significance level $\alpha_C = (\log 2)/(5 \times 10^{15}) \approx 1.38 \times 10^{-16}$. But if Cathy is willing to go this far, why should she stop simply with false positives that threaten to slip into the scientific literature? In the ordinary circumstances of life there are plenty of opportunities to err with false positives, attributing events to something other than chance when they actually did result from chance. What if Cathy decides to incorporate into her α-level these ordinary circumstances of life? Her

α-level seems to be heading off asymptotically to zero. Where will this chain of reasoning end?[13]

In practice this asymptotic vanishing of α-levels does not arise. Indeed, one's need to control false positives is typically balanced by one's desire to get work done, which requires relaxing control over false positives. An overconcern to avoid error almost invariably stifles inquiry. Thus, in practice the tendency to decrease α-levels and increase probabilistic resources to control false positives is checked by a desire to foster inquiry. What's more, false positives can always be corrected through further research. It follows that moderation typically guides the choice of α-levels and probabilistic resources Ω. Moreover, once α and Ω are chosen, they tend to remain stable (unlike the Alice-Bob-Cathy story).

All the same, it is important to understand what happens when we disregard our need to get work done and foster inquiry, and instead focus exclusively on controlling for false positives. If we do this, we do not enter the infinite regress suggested by the Alice-Bob-Cathy story, but rather a finite regress that ends abruptly. Indeed, we find at the end of this regress a privileged set of probabilistic resources Δ that encompasses all the probabilistic resources a finite rational agent operating in the actual world need ever consider. This privileged set of probabilistic resources Δ is such that whenever a probability p is small relative to Δ, p remains small relative to any other set of probabilistic resources Ω that might arise in practice. The startling thing about Δ is that it is well-defined and reasonably compact so that it is useful in practical applications.

Δ derives from the total number of specifications capable of fitting in the actual world. Physics strictly limits this number, not just for the present, but throughout cosmic history. Call this number N. N will be computed shortly, but for now I want simply to show why a certain probability induced by N, namely $1/(2N)$, is the smallest probability we'll ever need. Given that the actual world can contain

[13] Although I won't take this approach, one way to end this chain of reasoning is to look to a Peircean ideal community. According to Peirce (1878 [1988], p. 1315), "logicality inexorably requires that our interests shall *not* be limited. They must not stop at our own fate, but must embrace the whole community. This community, again, must not be limited, but must extend to all races of beings with whom we can come into immediate or mediate intellectual relation. I must reach . . . beyond this geological epoch, beyond all bounds. He who would not sacrifice his own soul to save the whole world, is, as it seems to me, illogical in all his inferences, collectively. Logic is rooted in the social principle." Instead of looking to an ideal community of rational agents, I will look to the limitations of matter/energy in the universe.

no more than N specifications, consider what it would mean for even one specified event of probability strictly less than 1/(2N) to happen by chance. The total number of specified events is N. The total number of specified events of probability strictly less than 1/(2N) is therefore some number K that cannot exceed N (i.e., $K \leq N$). The (saturated) probability that even one specified event of probability strictly less than 1/(2N) ever happens by chance is therefore strictly less than K/(2N), which in turn is less than or equal to 1/2 (since $K \leq N$). It follows there is less than an even chance that even one specified event of probability strictly less than 1/(2N) will happen by chance throughout cosmic history. It's therefore safer to conjecture that none of these specified events ever happens by chance.

The reasoning here is identical to that in a lottery: Given a lottery with k tickets each of which has probability strictly less than 1/(2n) of winning, and given that k does not exceed n, the probability of the lottery having a winner is strictly less than k/(2n), which in turn is less than or equal to 1/2 (since $k \leq n$). Since the probability of the lottery having a winner is therefore strictly less than 1/2, it's safer to conjecture that the lottery will have no winner. Of course in practice, should such a lottery have a winner, we would tend to attribute it to chance. For most lotteries this is innocuous since typically we can expand our probabilistic resources to include other lotteries, thereby rendering chance plausible. But with the collection of all specified events of probability strictly less than 1/(2N), we no longer have this option. This is the lottery to end all lotteries.

Let us now flesh out this argument. We start by considering the following historical table containing every specification within the actual world throughout its entire history:

Subject	Event	Specification
S_1	E_1	D_1
S_2	E_2	D_2
$\vdots$	$\vdots$	$\vdots$
S_{K-1}	E_{K-1}	D_{K-1}
S_K	E_K	D_K
$\vdots$	$\vdots$	$\vdots$
S_N	E_N	D_N

Each row of this table denotes a subject S_i specifying an event E_i with a pattern D_i (hence D_i is detachable from E_i, and D_i^* subsumes E_i). Because this is a historical table, all the specifications in it are actual. In particular, specifications that might have been but never were (i.e., counterfactual specifications) have no place in it. Note that the subjects S_i are not limited to humans. Any finite rational agent instantiated in the physical stuff of the universe will do (e.g., extraterrestrial intelligences or cybernetic computers).

How big is N? Physical constraints strictly limit both the number of subjects that can exist at any one time and the speed with which any subject can generate specifications of events. Specifically, within the known physical universe there are estimated to be no more than 10^{80} elementary particles. Moreover, the properties of matter are such that transitions from one physical state to another cannot occur at a rate faster than 10^{45} times per second.[14] Finally, the universe itself is about a billion times younger than 10^{25} seconds (assuming the universe is around ten to twenty billion years old). If we now assume that any subject that ever specifies an event within the known physical universe must comprise at least one elementary particle, then these cosmological constraints imply that the total number of specified events throughout cosmic history cannot exceed

$$10^{80} \times 10^{45} \times 10^{25} = 10^{150}.$$

This is N.[15]

Note that the units in this equation are as follows: 10^{80} is a pure number – an upper bound on the number of elementary particles in the universe; 10^{45} is in hertz – alterations in the states of matter per second; 10^{25} is in seconds – an upper bound on the number of seconds that the universe can maintain its present integrity (i.e., before collapsing back on itself in "the big crunch" or undergoing heat death by

[14] This universal bound on transitions between physical states is based on the Planck time, which constitutes the smallest physically meaningful unit of time. See Halliday and Resnick (1988, p. 544). Note that universal time bounds for electronic computers have clock speeds between ten and twenty magnitudes slower than the Planck time. See Wegener (1987, p. 2).

[15] We can now see why the distinction between replicational and specificational resources is so crucial. Specifications are easily individuated and numerically manageable. Replications are not. Consider, for instance, the number of distinct ways to individuate 500,000 coin tosses among a total of 1,000,000 coin tosses. This number computes to the combinatorial $(1{,}000{,}000!)/(500{,}000!)^2 \approx 10^{300{,}000}$. It follows there are $10^{300{,}000}$ distinct ways to replicate 500,000 coin tosses among these 1,000,000 coin tosses (note, however, that these replications are not stochastically independent). This number is vast, and far exceeds $N = 10^{150}$.

expanding indefinitely). Technically, 10^{150} is the total number of state changes that all the elementary particles in the universe can undergo throughout the duration of the universe. But since any subject making a specification undergoes a state change, and since any such subject comprises at least one elementary particle, it follows that 10^{150} bounds the total number of specifications by subjects in the universe. 10^{150} is a supremely generous bound. Indeed, the only subjects we know that specify events are animals and computers, each of which comprise a vast ensemble of elementary particles, and generate specifications in time periods vastly slower than the Planck time. In setting N equal to 10^{150}, we therefore ensure that the preceding table includes all the specifications of events ever formulated by subjects throughout cosmic history.[16]

Because each row of the preceding table denotes a subject S_i specifying an event E_i with a pattern D_i, implicit in each row is a chance hypothesis $\mathbf{H}_i$, side information $\mathbf{I}_i$, and a bounded complexity measure $\Phi_i = (\varphi_i, \lambda_i)$. Each of these must be present for S_i to establish that D_i is a specification of E_i. Consider now the probability

$$\delta = 1/(2N) = 1/2 \times 1/10^{150}$$

and focus only on those rows for which the probability $\mathbf{P}(D_i^* \mid \mathbf{H}_i) = p_i$ is strictly less than δ, that is,

$$\mathbf{P}(D_i^* \mid \mathbf{H}_i) < \delta.$$

Without loss of generality we may assume the preceding table (see p. 208) has been so arranged that the first K rows are precisely the

[16] Nor does quantum computation offer to increase the number of specifications that can be concretely realized. Though quantum computation offers to dramatically boost computational power by allowing massively parallel computations, it does so by keeping computational states indeterminate until the very end of a computation. This indeterminateness of computational states takes the form of quantum superpositions, which are deliberately exploited in quantum computation to facilitate parallel computation. The problem with quantum superpositions, however, is that they are incapable of concretely realizing specifications. A quantum superposition is an indeterminate state. A specification is a determinate state. Measurement renders a quantum superposition determinate by producing a pure state, but once it does so we are no longer dealing with a quantum superposition. Because quantum computation thrives precisely where it exploits superpositions and avoids specificity, it offers no means for boosting the number of specifications that can be concretely realized in the universe (for an overview of quantum computation see DiVincenzo, 1995).

ones that satisfy this inequality. We then define the specificational resources Δ as these first K rows (i.e., the "δ-bounded rows").

Suppose now that a subject S specifies an event E with a pattern D, and that this specification is made in relation to a chance hypothesis **H**, side information **I**, and a bounded complexity measure $\Phi = (\varphi, \lambda)$. Suppose further that $\mathbf{P}(D^* \mid \mathbf{H}) < \delta$. The triplet (S, E, D) is therefore one of the first K rows of the preceding table (in particular, D is one of the D_is for $1 \leq i \leq K$). Define next the saturated event D^*_Δ as the disjunction of all the D^*_is, that is,

$$D^*_\Delta = D^*_1 \vee D^*_2 \vee \cdots \vee D^*_K.$$

D^*_Δ is the event that at least one E_i falls under D^*_i for $1 \leq i \leq K$. Moreover, the probability of D^*_Δ is the probability that at least one E_i falls under D^*_i by chance.

What, then, is the probability of D^*_Δ? Clearly, any chance hypothesis $\hat{\mathbf{H}}$ characterizing the chance occurrence of D^*_Δ has to be consistent with the chance hypotheses $\mathbf{H}_1$ through $\mathbf{H}_K$, and therefore satisfy $\mathbf{P}(D^*_i \mid \hat{\mathbf{H}}) = \mathbf{P}(D^*_i \mid \mathbf{H}_i)$ for $1 \leq i \leq K$ (if $\hat{\mathbf{H}}$ fails adequately to characterize the chance occurrence of the disjuncts of D^*_Δ, much less will it adequately characterize the chance occurrence of D^*_Δ). Any such $\hat{\mathbf{H}}$ therefore satisfies the following inequality:

$$\begin{aligned}
\mathbf{P}(D^*_\Delta \mid \hat{\mathbf{H}}) &= \mathbf{P}(D^*_1 \vee D^*_2 \vee \cdots \vee D^*_K \mid \hat{\mathbf{H}}) \\
&\leq \mathbf{P}(D^*_1 \mid \hat{\mathbf{H}}) + \mathbf{P}(D^*_2 \mid \hat{\mathbf{H}}) + \cdots + \mathbf{P}(D^*_K \mid \hat{\mathbf{H}}) \\
&= \mathbf{P}(D^*_1 \mid \mathbf{H}_1) + \mathbf{P}(D^*_2 \mid \mathbf{H}_2) + \cdots + \mathbf{P}(D^*_K \mid \mathbf{H}_K) \\
&< K\delta \\
&= K/(2N) \\
&\leq 1/2.
\end{aligned}$$

It follows that the probability of D^*_Δ cannot but be strictly less than 1/2, and therefore that D^* has small probability with respect to the probabilistic resources Δ.[17]

[17] Note that $\hat{\mathbf{H}}$ neither presupposes the Kolmogorov Consistency Theory (see Bauer, 1981, pp. 371–4) nor miscarries should the Bell's inequalities be violated (see Sudbery, 1988, pp. 198–201). $\hat{\mathbf{H}}$ simply enables us to establish an upper bound for the probability of D^*_Δ, not to calculate it precisely.

We now face a probabilistic inconsistency (cf. Section 6.3): (1) The complement of D^*_Δ has probability strictly greater than 1/2; (2) E fell under D^*_Δ even though D^*_Δ has probability strictly less than 1/2, and even though D^*_Δ was identified without knowledge of E (this last point holds because D specifies E and D is a disjunct of D^*_Δ); and (3) all the probabilistic resources relevant to E's occurrence have been factored into D^*_Δ. The only way to resolve this probabilistic inconsistency is by concluding that E did not occur by chance. True, this conclusion is relativized to the probabilistic resources Δ. Thus, for S to continue attributing E to chance, S will have to doubt Δ's adequacy for controlling false positives and thereby eliminating chance (much as Bob doubted the adequacy of Alice's probabilistic resources, and Cathy in turn doubted Bob's). But S cannot legitimately entertain this doubt, for S has factored in all the probabilistic resources that can conceivably specify E by means of a δ-bounded pattern; and once all these δ-bounded patterns are factored in, the probability that E by chance conforms to at least one of them still remains strictly less than 1/2.

The small probability δ at once resembles and diverges from the significance levels in the Alice-Bob-Cathy story. Thus, just as it's more likely than not that throughout its lifetime Alice's journal will be free of false positives if the significance level of Alice's journal is set at $\alpha_A = 5.77 \times 10^{-7}$, so too it's more likely than not that throughout cosmic history the universe will be free of false positives if the level at which chance is eliminated is set at $\delta = 1/2 \times 1/10^{150}$. On the other hand, unlike Alice's journal, whose α-level can be challenged because of the existence of other journals, given the totality of specifications throughout cosmic history there are no missing pieces to challenge δ. With Alice employing a significance level $\alpha_A = 5.77 \times 10^{-7}$, Bob and Cathy can point to other journals that haven't been factored into Alice's α-level. But with a subject S employing the probability bound $\delta = 1/2 \times 1/10^{150}$, there is nothing a fellow subject S′ can point to that hasn't been factored into δ already. When Bob confronts Alice over the inadequacy of her α-level, it's as though Alice just learned that a lottery she's been playing is less difficult to win than she initially thought. On the other hand, by employing the probability bound δ, a subject S ensures that of all the lotteries actually being played, S has chosen the most difficult.

The difference between local and universal small probabilities is the difference between Alice's α of 5.77×10^{-7} and S's δ of $1/(2N) = 1/2 \times 1/10^{150}$. Probabilities less than Alice's α are small probabilities relative to probabilistic resources that, depending on circumstances, may need to be augmented to ensure that false positives remain less likely than not. Probabilities less than S's δ, on the other hand, are small probabilities relative to a maximal set of probabilistic resources that, irrespective of circumstances, need never be augmented to ensure that false positives remain less likely than not.

We thus define a *local small probability* as a small probability relative to a set of probabilistic resources that, depending on circumstances, may need to be augmented. On the other hand, we define a *universal small probability* as a small probability relative to a privileged set of probabilistic resources that, irrespective of circumstances, need never be augmented (Δ is a case in point). Given these definitions, it is useful to introduce two further distinctions: The probabilistic resources used to define local (resp. universal) small probabilities will be known as *local* (resp. *universal*) *probabilistic resources*; what's more, the probability bounds α (resp. δ) below which probabilities become local (resp. universal) small probabilities will be called *local* (resp. *universal*) *probability bounds*.[18]

Although the logic of the Generic Chance Elimination Argument is identical for local and universal probability bounds, it simplifies considerably for the universal probability bound δ. Since the universal probabilistic resources Δ that define δ are fully adequate for all circumstances, we can combine (5), (6), and (7) of the GCEA, and rewrite it more compactly as follows:

(5–7) S finds that $p = \mathbf{P}(D^* \mid \mathbf{H})$ is strictly less than δ.[19]

Given the universal probability bound δ, the GCEA therefore reduces to a subject learning that an event E has occurred (1), identifying a

[18] Universal probability bounds, though not called by that name, appeared in Emile Borel's work on small probabilities. Borel (1962, p. 28) referred to probabilities falling below a universal probability bound as "probabilities which are negligible on the cosmic scale." According to him, "when the probability of an event is below this limit [= universal probability bound], the opposite event may be expected to occur with certainty, whatever the number of occasions presenting themselves in the entire universe." Borel held to a universal probability bound of 10^{-50}. Without a clear conception of specification and probabilistic resources, Borel's derivation of this number was less than fully rigorous. My own $\delta = 1/2 \times 1/10^{150}$ is a rigorous version of Borel's universal probability bound.

[19] Explicit reference to Δ can be omitted here since Δ is implicit in δ.

pattern D that delimits E (3), computing the probability of D^* ((2) and (4)), checking that this probability is strictly less than δ (5–7), and finally establishing that D is detachable from E ((8) and (9)). Once these conditions are satisfied, S is warranted inferring that E did not occur by chance.

6.6 THE INFLATIONARY FALLACY

In justifying the universal probability bound δ ($= 1/2 \times 1/10^{150}$), I assumed a noninflationary big-bang cosmology. In an inflationary universe, what we regard as the known physical universe (i.e., the sum total of energy that can potentially interact with us causally) is just one of a multitude of causally isolated subuniverses. According to Alan Guth, the totality of these causally isolated subuniverses contains many more than the 10^{80} elementary particles constituting the known physical universe, which for Guth is just the subuniverse we happen to inhabit (see Guth and Steinhardt, 1989). Whereas a noninflationary universe is a rather small place with room for few probabilistic resources, an inflationary universe is a much bigger place with room for, by some accounts, infinitely many probabilistic resources. The question therefore arises whether in hitching my universal probability bound δ to a noninflationary big-bang cosmology, I haven't undercut its ability to demarcate small probabilities. Cosmological theories come and go. Why, then, should we take δ seriously?

The only reason not to take a probability bound seriously is if we've omitted relevant probabilistic resources. The key word here is "relevant." Not every set of probabilistic resources is relevant for deciding whether an event happened by chance. Consider two state lotteries both of which have printed a million lottery tickets. One lottery sells all million tickets, the other sells only two tickets. Ostensibly both lotteries have the same number of probabilistic resources. Nevertheless, the relevant probabilistic resources for deciding whether the first lottery produced a winner by chance greatly exceed those of the second. Probabilistic resources are opportunities for an event to happen. To be relevant to an event, these opportunities need to be actual and not merely possible. Lottery tickets sitting on a shelf collecting dust might just as well never have been printed.

This much is uncontroversial. But suppose next we know nothing about the number of lottery tickets sold, and are informed simply that the lottery had a winner. Suppose furthermore the probability of any lottery ticket producing a winner is extremely low. Now what do we conclude? Does it follow that numerous lottery tickets were sold? Not at all. We are entitled to this conclusion only if we have independent evidence that numerous lottery tickets were sold. But absent such evidence, we have no grounds for asserting the existence of numerous lottery tickets, or even that the lottery was conducted fairly and that its outcome was due to chance. It is illegitimate to take an event, decide in advance it must be due to chance, and then propose numerous probabilistic resources because otherwise chance would be implausible. This is the inflationary fallacy, and it is utterly bogus.

We do not invent probabilistic resources simply to prop an otherwise failing chance hypothesis. Rather, we determine independently whether there really are enough probabilistic resources to render chance plausible. This reversal of commonsense logic where we fixate on chance, and then madly rush to invent the probabilistic resources necessary to preserve chance is the great pipedream of late twentieth-century cosmology and biology. It leads to such fanciful statements as "the universe is a free lunch" and "life is a cosmic imperative." The moment one posits infinitely many probabilistic resources, anything possible becomes certain (probabilistically this follows from the Strong Law of Large Numbers). The inflationary fallacy blurs the distinction between the actual and the possible. Moreover, it does so tendentiously in hopes of escaping the claims of the actual. The bubble universes of inflationary cosmology, the many worlds of quantum physics, and the possible worlds of metaphysics all serve to inflate our probabilistic resources so that what otherwise seems absurd to chance becomes not only plausible but inevitable.

Call them what you will – bubble universes, many worlds, possible worlds – there is something deeply unsatisfying about positing these entities simply because chance requires it. In each case the posited entities are causally isolated from our space–time manifold. Hence the only evidence in their favor is their ability to render chance plausible. This is pushing inference to the best explanation too far. It is legitimate to posit an entity to explain a phenomenon only if the entity is at least in principle capable of interacting with the phenomenon. But entities posited simply to inflate probabilistic resources are utterly inert and self-contained. They do not interact with our space–time manifold. They exist solely to make chance intelligible.

And yet, intelligibility is precisely what these posited entities undermine. Unlimited probabilistic resources are an epistemological nightmare. They transform Hume's problem of induction, which in the end amounts simply to an acknowledgment that our inductions might be wrong, into a positive reason for thinking that anything might happen anywhere at anytime. Despite a life of unremitting charity and self-sacrifice, did Mother Teresa in her last days experience a neurological accident that caused her to become an ax murderer? Though bizarre, this is a logical possibility. Moreover, as a logical possibility, it is sure to happen in some possible world. And what is to prevent that world from being ours? So long as our probabilistic resources are limited, a Generic Chance Elimination Argument quickly dispenses with this possibility – to be sure, not with absolute certainty (Hume's problem of induction is after all a problem), but to a moral certainty beyond reasonable doubt. But with unlimited probabilistic resources, we lose any rational basis for eliminating chance.

The problem here is that statistical decision theory breaks down for probabilistic resources that are causally disconnected from our space–time manifold. For instance, it is illegitimate to argue against Mother Teresa becoming an ax murderer by arguing that the subcollection of possible worlds where she becomes an ax murderer has small probability. Such higher-order probabilities are incoherent. We can assign probabilities within possible worlds, but not to collections of possible worlds. Possible worlds are fully self-contained. Consequently, there exists no chance process to pick out the possible world we inhabit. Indeed, no law, no reason, no chance process, no God – nothing – determines why we are in this possible world as opposed to another (laws, reasons, chance processes, and divinities belong strictly inside possible worlds, and cannot transcend them). Alternatively, there can be no epistemological road-map to guide us through the set of possible worlds to our own actual world. Perhaps Shakespeare was a genius. Perhaps Shakespeare was an imbecile who just by chance happened to string together a long sequence of apt phrases. Unlimited probabilistic resources ensure not only that we'll never know, but also that we have no rational basis for preferring one to the other.

We can now answer our earlier question, Why should we take δ seriously? We need to take δ seriously because it omits no relevant probabilistic resources. Once we reject the inflationary fallacy, we are back to our own little universe. As long as we haven't underestimated the probabilistic resources in our universe, δ will incorporate all the probabilistic resources we ever need. How, then, do we know we haven't underestimated the probabilistic resources in our universe? Certainly I don't mean to suggest that δ is written in stone. δ is based on our best current understanding of physics and cosmology. As this understanding develops, δ may well need to be revised (whether up or down is unclear). In making this admission, however, let's be clear what it would mean to underestimate the probabilistic resources in our universe. Three numbers determine δ: (1) the number of elementary particles in the universe, (2) the rate at which physical states can change, and (3) the length of time during which the universe can sustain specifying agents like ourselves. In Section 6.5 I provided upper bounds for these numbers, respectively, 10^{80}, 10^{45}, and 10^{25}. δ varies inversely with these numbers, so that as they increase, δ decreases. Hence if these numbers are off, then so is δ.

To underestimate the probabilistic resources in our universe is therefore to underestimate one of these numbers. Are there more than 10^{80} elementary particles in the universe? Is the rate at which physical states can change greater than 10^{45} per second? Does the length of time during which the universe can support specifying agents exceed 10^{25} seconds? Our best current science answers a firm No to each of these questions. This is not to say these numbers won't change. But it is to say they better not change simply to prop an otherwise failing chance hypothesis. If these numbers change, it will be because experimental scientists have done the hard work of uncovering new facts about the universe's extent, duration, structure, and dynamics. On the other hand, it will not be because armchair theorists have propounded yet another inflationary fallacy.

6.7 THE LAW OF SMALL PROBABILITY

All the pieces are now in place to explicate and justify the Law of Small Probability (LSP). In Section 2.2 we identified the Law of Small Probability with the claim that *specified events of small probability* *do not occur by chance*. We then rewrote this claim as the following formula in the first-order predicate logic:

$$\forall X[oc(X) \ \& \ sp(X) \ \& \ SP(X) \rightarrow \ \sim ch(X)].$$

Here ∀X denotes a universal quantifier that ranges over events, & denotes conjunction, → denotes implication, ∼ denotes negation, and *oc*, *sp*, *SP*, and *ch* denote the one-place predicates $oc(X) =$ X *has occurred*, $sp(X) =$ X *is specified*, $SP(X) =$ X *has small probability*, and $ch(X) =$ X *is due to chance*.

Because these predicates were formulated in the pretheoretic days before we had an adequate account of specification and small probability, we now need to reformulate them. We therefore fix a subject S who identifies both a requisite precondition $\Sigma = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Phi = (\varphi, \lambda))$ and probabilistic resources Ω. Together S, Σ, and Ω serve as parameters (i.e., background constants) for the reformulated predicates. The reformulated predicates are now defined as follows:

$oc(E) =_{def}$ S learns that E has occurred.
$sp(E) =_{def}$ S has identified a pattern D such that $D^* = E$ and D is detachable from E with respect to Σ.
$SP(E) =_{def}$ S finds that $\mathbf{P}(E \mid \mathbf{H})$ has small probability with respect to Ω.
$ch(E) =_{def}$ S is not warranted inferring that E did not occur according to the chance hypothesis **H**.

Here E is an arbitrary event. Note that because S, Σ, and Ω are parameters, they are actually embedded in these predicates. Technically, therefore, we could write, $oc(E) = oc(E; S, \Sigma, \Omega)$, $sp(E) = sp(E; S, \Sigma, \Omega$, $SP(E) = SP(E; S, \Sigma, \Omega)$, and $ch(E) = ch(E; S, \Sigma, \Omega)$, with variables to the left of the semicolon, parameters to the right.

Although these reformulated predicates appear to take some liberties with the original predicates, in fact they capture precisely what the original predicates were trying to assert. Moreover, they are precisely what we need to justify the Law of Small Probability. Indeed, I am going to justify the Law of Small Probability by deriving it from the Generic Chance Elimination Argument. Since the Generic Chance Elimination Argument was itself justified in Sections 6.1 through 6.3, deriving the Law of Small Probability from it will suffice to justify the

Law of Small Probability. It's clear from inspection that the Law of Small Probability and the Generic Chance Elimination Argument are engaged in substantially the same task, namely, eliminating chance through small probabilities by means of specifications. By reformulating these predicates as we have, the Law of Small Probability logically encapsulates the Generic Chance Elimination Argument, so that the antecedent *oc*(E) & *sp*(E) & *SP*(E) entails (1)–(9) of the GCEA and the consequent $\sim ch$(E) is logically equivalent to (10) (cf. Section 6.2).

Let us now consider the rationale behind these reformulations.

oc(E). *As originally formulated, this predicate asserts that E has occurred. So formulated, this predicate says nothing about the rational agent who learns that E has occurred. Since this predicate and the others need to be formulated specifically for their role in eliminating chance, the rational agent who attempts to eliminate chance needs to be explicitly cited in these predicates. As the chief item of interest in a chance-elimination argument, E does not occur in isolation from rational agents. Chance-elimination arguments are after all arguments that rational agents produce in response to events about which they have knowledge – unidentified events have no place here. To be of use in a chance-elimination argument, it therefore isn't enough for an event simply to have occurred. Rather a rational agent – what we are calling a subject S – must know that the event has occurred. This is the rationale for reformulating oc(E).*

sp(E). *As originally formulated, this predicate asserts that E is specified. For E to be specified, however, presupposes a pattern D that matches E (i.e., $D^* = E$), as well as a requisite precondition $\Sigma = (\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{I}, \Phi = (\varphi, \lambda))$ that detaches D from E – see Sections 5.4 and 5.2 respectively (note that because D matches E, it occurs in this predicate as a bound variable, and thus can be suppressed). Moreover, as part of a chance-elimination argument, D, E, and Σ cannot be taken in isolation from a subject S who employs Σ to show that D specifies E. This is the rationale for reformulating sp(E).*

SP(E). *As originally formulated, this predicate asserts that E has small probability. The very idea of an event having small probability, however, presupposes a set of probabilistic resources Ω with respect* *to which small probabilities can be evaluated (cf. Sections 6.1–6.3). Moreover, since this predicate is going to appear in a chance-elimination argument, there has to be somebody – a subject S – to establish that E has small probability with respect to Ω. This is the rationale for reformulating SP(E).*

ch(E)*. As originally formulated, this predicate asserts that E occurred by chance. Nonetheless, because our concern with this predicate is how it enters a chance-elimination argument, to deny or affirm that E occurred by chance is not in the first instance to make a metaphysical claim about the causal process underlying E's occurrence. Rather, it is to make an epistemic claim about a subject S's warrant for attributing E to a chance hypothesis* ***H****. Given a chance hypothesis* ***H*** *that could conceivably explain E, S is obliged to retain* ***H*** *as a live possibility until a positive warrant is found for rejecting it. Chance-elimination arguments are after all elimination arguments. Because chance is always the default option, only active refutation eliminates it. To eliminate chance, a subject S must have positive warrant for inferring that E did not occur according to the chance hypothesis* ***H****. On the other hand, to retain chance a subject S must simply lack warrant for inferring that E did not occur according to the chance hypothesis* ***H****.*

Prima facie, our reformulation of ch(E) seems unnecessarily roundabout. We want to know whether an event E has occurred by chance, and not whether we lack warrant for attributing E to something other than chance. As an epistemic matter, however, we can only know the latter, and not the former. Indeed, it is only by means of the latter that we can assert the former (i.e., it is only by lacking/possessing warrant for attributing E to something other than chance that we can assert that E did/didn't occur by chance). Although the two negations that appear in our reformulation of ch(E) seem to complicate matters unduly, one of these negations drops out when we negate ch(E):

$$\sim ch(E) = \text{S is warranted inferring that E did not occur according to the chance hypothesis } \boldsymbol{H}.$$

ch is therefore more naturally formulated in terms of ~ch rather than the other way round. Here, then, is the rationale for reformulating ch(E).

Suppose now we are given an arbitrary event E, a subject S who identifies both a requisite precondition $\Sigma = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Phi = (\varphi, \lambda))$ and probabilistic resources Ω, and the predicates *oc, sp, SP*, and *ch* as we have just reformulated them. Next, consider the conjunction

$$oc(\mathrm{E}) \,\&\, sp(\mathrm{E}) \,\&\, SP(\mathrm{E}).$$

To assert this conjunction against this background is then logically equivalent to asserting (1) through (9) of the GCEA for the special case where the pattern D matches E (because D matches E, it occurs in this conjunction as a bound variable, and thus can be suppressed). To prove this equivalence is a simple matter of bookkeeping.

As soon as (1) through (9) obtain, the GCEA automatically approves (10), namely, that S is warranted inferring E did not occur according to the chance hypothesis **H**. But this last claim is just $\sim ch(\mathrm{E})$. It follows that $\sim ch(\mathrm{E})$ is a consequence of $oc(\mathrm{E}) \,\&\, sp(\mathrm{E}) \,\&\, SP(\mathrm{E})$, and hence that the relation between these formulas takes the form of a conditional:

$$oc(\mathrm{E}) \,\&\, sp(\mathrm{E}) \,\&\, SP(\mathrm{E}) \rightarrow \sim ch(\mathrm{E}).$$

What sort of conditional is this? Certainly not an entailment (it's logically possible for the consequent to be mistaken despite the antecedent being true). What's at stake in this conditional is epistemic justification, or what philosophers who study conditional logic call assertibility. Thus, given the antecedent $oc(\mathrm{E}) \,\&\, sp(\mathrm{E}) \,\&\, SP(\mathrm{E})$, we are epistemically justified asserting the consequent $\sim ch(\mathrm{E})$.[20]

We can generalize the preceding conditional slightly. The predicate $sp(\mathrm{E})$ asserts that some pattern D matches E (i.e., $\mathrm{D}^* = \mathrm{E}$). But in the GCEA, all that's required is for S to identify a pattern D that delimits E (i.e., E entails D^*). Given the parameters S, Σ, and Ω, we therefore define the following predicate:

$sp(\mathrm{D}, \mathrm{E}) =_{\mathrm{def}}$ S has identified the pattern D such that D delimits E and D is detachable from E with respect to Σ.

Here D is an arbitrary pattern and E an arbitrary event. Given this predicate we now form the following conditional:

$$oc(\mathrm{E}) \,\&\, \exists \mathrm{D}[sp(\mathrm{D}, \mathrm{E}) \,\&\, SP(\mathrm{D}^*)] \rightarrow \sim ch(\mathrm{E}).$$

[20] Cf. Adams (1975), Lewis (1976), Ellis (1978), Appiah (1985), and Jackson (1987; 1991). Conditional logic is a vast field of study, and one best left for another occasion.

Here the existential quantifier ranges over patterns. For this conditional, the antecedent is logically equivalent to (1) through (9) of the GCEA, and the consequent is logically equivalent to (10). As before, proving these equivalences is a matter of bookkeeping.

This, then, is the final form of the Law of Small Probability. Note that it logically encapsulates the Generic Chance Elimination Argument of Section 6.2. Because E is arbitrary in this conditional, we can place it under the scope of a universal quantifier. If we do this, we obtain the following definitive statement of the Law of Small Probability:

The Law of Small Probability. *Suppose a subject S has identified both a requisite precondition* $\Sigma = (\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{I}, \Phi = (\varphi, \lambda))$ *and probabilistic resources* Ω. *Then the following formula defines the Law of Small Probability:*

$$\forall X\{oc(X) \;\&\; \exists D[sp(D, X) \;\&\; SP(D^*)] \rightarrow \sim ch(X)\}.$$

Here X ranges over events and D over patterns.

In Section 2.2 we formulated a statement of the design inference that employed the earlier, weaker version of the Law of Small Probability. If we now substitute this last version of the Law of Small Probability, we obtain the following definitive statement of the design inference (cf. Section 2.2):

The Design Inference. *Suppose a subject S has identified all the relevant chance hypotheses* $\mathcal{H}$ *that could be responsible for some event E. Suppose further that S has identified (1) a probability measure* $\boldsymbol{P}$ *that estimates likelihoods with respect to the chance hypotheses in* $\mathcal{H}$, *(2) side information* $\boldsymbol{I}$, *(3) a bounded complexity measure* $\Phi = (\varphi, \lambda)$ *that characterizes S's problem-solving ability, and (4) probabilistic resources* Ω *that characterize S's needs and interests in controlling false positives. Then the following argument defines the design inference:*

P1: $oc(\mathrm{E})$
P2: $(\forall \mathbf{H} \in \mathcal{H})(\exists \mathrm{D})[sp(\mathrm{D}, \mathrm{E}; \mathbf{H}) \;\&\; SP(\mathrm{D}^*; \mathbf{H})]$
P3: $(\forall \mathrm{X})(\forall \mathbf{H} \in \mathcal{H})\{[oc(\mathrm{X}) \;\&\; (\exists \mathrm{D})[sp(\mathrm{D}, \mathrm{X}; \mathbf{H}) \;\&\; SP(\mathrm{D}^*; \mathbf{H})]] \rightarrow \sim ch(\mathrm{X}; \mathbf{H})\}$
P4: $\sim reg(\mathrm{E})$
P5: $reg(\mathrm{E}) \vee (\exists \mathbf{H} \in \mathcal{H}) ch(\mathrm{E}; \mathbf{H}) \vee des(\mathrm{E})$
C: $des(\mathrm{E})$.

A few clarifications are in order. First, the validity of this argument is straightforward (cf. Section 2.2). Second, once **H** in $\mathcal{H}$ is fixed, the predicates *oc, sp, SP*, and *ch* operate with the usual parameters S, $\Sigma = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Phi = (\varphi, \lambda))$, and Ω. Hence quantifying over the parameter **H** makes perfect sense in P2, P3, and P5. Third, P2 properly combines and strengthens what were Premises 2 and 3 in Section 2.2. Fourth, since the Law of Small Probability holds for all parameters S, $\Sigma = (\mathbf{H}, \mathbf{P}, \mathbf{I}, \Phi = (\varphi, \lambda))$, and Ω, placing the Law of Small Probability under the quantifier $\forall \mathbf{H} \in \mathcal{H}$ in P3 is perfectly legitimate so long as S, **P**, **I**, $\Phi = (\varphi, \lambda)$, and Ω remain fixed (which in this case they do). Fifth, we define *reg*(E) and *des*(E) as in Section 2.2 (note the discussion there): *reg*(E) $=_{\text{def}}$ there is some regularity **R** that accounts for E; *des*(E) $=_{\text{def}}$ ~*reg*(E) & $(\forall \mathbf{H} \in \mathcal{H})$ ~*ch*(E; **H**). Our discussion of the Law of Small Probability and the design inference is now complete.

7

Epilogue

I want in these closing pages to tie together some loose ends. Let's begin with a criticism of the design inference. The question can be raised whether the design inference isn't merely a technique for detecting coincidences, and if so, whether design isn't in the end a trivial mode of explanation. What underlies this criticism is the identification of coincidence with brute concurrence of events for which explanation is not only superfluous, but downright unwelcome. Galen Pletcher (1990, pp. 205–6) describes this sense of coincidence as follows:

> To call something a "coincidence" explains nothing. It does exactly the opposite: it asserts that the fact that two events are closely related – in time, and in other ways – does not need to be explained. It says more than that the relation between them cannot (at present) be explained.... To call something a coincidence is to express (even if only implicitly or perhaps even unwittingly) the opinion that it is misguided to search for an explanation (in the proper sense) of the coinciding of the phenomena at issue.

This criticism fails to recognize that not all coincidences are best left unexplained. Yes, the fact that the Shoemaker-Levy comet crashed into Jupiter exactly 25 years to the day after the Apollo 11 moon landing is a coincidence probably best left unexplained. But the fact that Mary Baker Eddy's writings on Christian Science bear a remarkable resemblance to Phineas Parkhurst Quimby's writings on mental healing is a coincidence that deserves to be explained, and is best explained by positing Quimby as a source for Eddy (cf. Martin, 1985, pp. 127–30).

The term "coincidence" properly speaking has two senses. The most basic sense is simply that of a concurrence (i.e., two or more things coming together), with no prejudice against how the concurrence is to be explained or whether the concurrence even has an explanation. The other sense is that of a concurrence for which inquiring after an underlying explanation is misguided and destined to fail. How do we know when it is fruitless to explain a coincidence?

Presumably, as we go through life trying to understand the world, we don't have an a priori list of items deemed unworthy of explanation. Rather, we conclude that something is unworthy of explanation only after we have given it the old college try, and found that in attempting to explain it we're in worse shape than before we started. Coincidences always begin as open problems and only later, after repeated frustration, become closed books.

How then do we distinguish the fecund coincidences, those worth explaining, from the sterile coincidences, those better left unexplained? The design inference leaves this decision to the Explanatory Filter (cf. Section 2.1), letting it decide whether a coincidence is best explained in terms of a regularity, chance, or design. As we have seen repeatedly throughout this monograph, there are plenty of coincidences that cry out for explanation, that are readily submitted to the Explanatory Filter, and that upon being submitted to the Explanatory Filter yield one of these three modes of explanation.

On the other hand, there are coincidences for which we haven't a clue how to submit them to the Explanatory Filter. Consider the following coincidence recounted by Carl Jung (1973, p. 10):

> I noted the following on April 1, 1949: Today is Friday. We have fish for lunch. Somebody happens to mention the custom of making an "April fish" of someone. That same morning I made a note of an inscription which read: "Est homo totus medius *piscis* ab imo." In the afternoon a former patient of mine, whom I had not seen for months, showed me some extremely impressive pictures of fish which she had painted in the meantime. In the evening I was shown a piece of embroidery with fish-like sea-monsters in it. On the morning of April 2 another patient, whom I had not seen for many years, told me a dream in which she stood on the shore of a lake and saw a large fish that swam straight towards her and landed at her feet. I was at this time engaged on a study of the fish symbol in history. Only one of the persons mentioned here knew anything about it.

What explains the overwhelming amount of fish imagery that bombarded Jung around the first of April 1949? All attempts to explain this coincidence, much less submit it to the Explanatory Filter, have struck me as fruitless. This coincidence seems best left unexplained and regarded as a sterile coincidence.

The criticism that the design inference attempts the futile task of explaining sterile coincidences is therefore easily dispensed with. Not all coincidences are sterile, and for those that are not the design inference provides a useful explanatory tool. Even so, there is a related worry that the design inference might inadvertently get applied to sterile coincidences and end up assigning them a significance they don't deserve. Specifically, the worry is that absent a suitable naturalistic story, attributing a coincidence to design is tantamount to invoking supernatural powers.

To appreciate what's at stake here, recall the case of Nicholas Caputo (cf. Section 1.2). Caputo, as the Democratic clerk from Essex County, New Jersey, selected the Democrats to head the ballot line forty out of forty-one times in his county. Caputo was supposed to have obtained his ballot line selections – which unduly favored his own political party – by chance. Nevertheless, if chance was responsible, Caputo succeeded in bringing about a specified event of probability one in fifty billion. An application of the design inference thus concluded that Caputo's ballot line selections were not due to chance but to design. Now it's clear that there is a naturalistic story to be told here. The simplest one is that Caputo cheated, pretending the ballot line selections were obtained by chance when in fact he deliberately rigged them to favor the Democrats.

But consider now the following scenario which from a probabilistic point of view is isomorphic to the Caputo case: A parapsychology experiment is under way. Alice is the subject and Bob is the experimenter. Bob flips a coin forty-one times, each time asking Alice to predict the face of the coin that is about to appear. In forty out of the forty-one times Alice predicts correctly. If chance is operating, Alice has therefore succeeded in specifying an event of probability one in fifty billion. As in the Caputo case, an application of the design inference concludes that the coincidence between Alice's predictions and Bob's coin tossing was not due to chance but to design. No doubt, there are naturalistic stories to be told here as well. For example, Bob may have been cheating so that instead of flipping the coin, he secretly compelled it to land as predicted by Alice (this is essentially what Bill Murray did playing a parapsychologist in the movie *Ghostbusters*). Suppose, however, that experimental controls were tight, and that cheating was precluded. If the experiment happened as described, what's wrong with attributing the outcome to design?

In Chapter 2 we defined design as the set-theoretic complement of the disjunction regularity-or-chance. Nothing in this definition entails a causal story, much less an intelligent agent, much less still a supernatural or occult power. Taken in its most fundamental sense, the word *design* signifies a *pattern* or *blueprint*. The key step in any design inference is showing that an event conforms to a pattern. Frequently the reason an event conforms to a pattern is because an intelligent agent arranged it so (cf. Section 2.4). There is no reason, however, to turn this common occurrence into a metaphysical first principle.

We can determine whether an event conforms to a pattern without having to explain why the conformity exists. Thus, even though in practice inferring design is the first step in identifying an intelligent agent, taken by itself design does not require that such an agent be posited. The notion of design that emerges from the design inference must not be confused with intelligent agency. Though they operate in tandem, the two are separate notions. Whether an event conforms to a pattern is a separate question from what caused the event to conform to the pattern. The reference to "design" in the design inference arises in the first instance because in the structure of the inference, patterns and events coincide.

When the design inference infers design, its primary effect is to limit our explanatory options. Only secondarily does it help identify a cause. To identify a cause we need to investigate the particulars of the situation where design was inferred. Simply put, we need more details. In the Caputo case, for instance, it seems clear enough what the causal story is, namely, that Caputo cheated. In the probabilistically isomorphic case of Alice and Bob, however, we may have to live without a causal explanation. To be sure, we can follow the parapsychology community in positing psi, that factor or faculty supposedly responsible for parapsychological coincidences. Nevertheless, the explanatory benefits of invoking psi remain far from clear.

If we now leave aside the worry that inferring design may require us to tell unacceptable causal stories, still another worry is likely to remain. Consider the Shoemaker-Levy coincidence mentioned in passing a few pages back. The Shoemaker-Levy comet crashed into Jupiter exactly 25 years to the day after the Apollo 11 moon landing. What are we to make of this coincidence? Do we really want to explain this coincidence in terms of design? What if we submitted this coincidence to the Explanatory Filter and out popped design? Granted, as a noncausal notion design doesn't require that we invoke supernatural powers. Our intuitions strongly suggest, however, that the comet's trajectory and NASA's space program were operating independently, and that at best the coincidence should be referred to chance – certainly not to design.

This worry is readily dispatched. The fact is that the design inference does not yield design all that easily, especially if probabilistic resources are sufficiently generous. It is simply not the case that unusual and striking coincidences automatically generate design as the conclusion of a design inference. There is a calculation to be performed. *Do the calculation. Take the numbers seriously. See if the underlying probabilities really are small enough to yield design.* Martin Gardner (1972) is quite correct when he notes, "The number of events in which you participate for a month, or even a week, is so huge that the probability of noticing a startling correlation is quite high, especially if you keep a sharp outlook." The implication he means to draw, however, is not correct, namely, that therefore startling correlations/coincidences may uniformly be relegated to chance.

I stress again, *Do the probability calculation!* The design inference is robust and easily resists counterexamples of the Shoemaker-Levy variety. Assuming, for instance, that the Apollo 11 moon landing specifies the crash of Shoemaker-Levy into Jupiter (a generous concession at that), and that the comet could have crashed at any time within a period of a year, and that the comet crashed to the very second precisely 25 years after the moon landing, a straightforward probability calculation indicates that the probability of this coincidence is no smaller than 10^{-8}. This simply isn't all that small a probability, especially when considered against the backdrop of the entire solar system. Certainly this probability is nowhere near the universal probability bound of 10^{-150} calculated in Section 6.5. I have yet to see a convincing application of the design inference that infers design for coincidences that our ordinary inclination attributes to chance.

Having shown how the design inference withstands certain criticisms and worries, I want to conclude this epilogue by stating what I take to be the main significance of the design inference for science. It is this: The design inference detects and measures *information*. Increasingly, scientists are recognizing the importance of information. Manfred Eigen, for instance, regards information as the central problem of biology, and one he hopes to unravel through algorithms and natural laws (1992, p. 12). David Chalmers, in attempting to explain human consciousness, proposes that "just as physics assumes the existence of properties of nature such as space, time, energy, charge and mass, so must a theory of consciousness posit the existence of a new fundamental property: information" (Horgan, 1994, p. 94; see also Chalmers, 1996, ch. 8). Keith Devlin (1991, p. 2) ponders whether "*information* should be regarded as . . . a basic property of the universe, alongside matter and energy (and being ultimately interconvertible with them)."

Like design, information is a noncausal notion. The transmission of information, though frequently mediated by transfers of energy across a communication channel, is properly understood in terms of the correlations between what happens at the two ends of a communication channel – and thus without reference to any intervening causal process. As Fred Dretske (1981, p. 26) remarks,

> It may seem as though the transmission of information . . . is a process that depends on the causal inter-relatedness of source and receiver. The way one gets a message from s to r is by initiating a sequence of events at s that culminates in a corresponding sequence at r. In abstract terms, the message is borne from s [to] r by a causal process which determines what happens at r in terms of what happens at s.
>
> The flow of information may, and in most familiar instances obviously does, depend on underlying causal processes. Nevertheless, the information relationships between s and r must be distinguished from the system of causal relationships existing between these points.

How, then, does the design inference detect and measure information? For information to pass from a source s to a receiver r, a message M′ emitted at s must suitably constrain a message M″ received at r. Moreover, the amount of information passing from s to r is by definition the negative logarithm to the base 2 of the probability of M′ (see Dretske, 1981, chs. 1 and 2). For the design inference to detect information, the message M′ emitted at s has to be identified with a pattern D, and the message M″ received at r has to be identified with an event E. Moreover, for M′ to suitably constrain M″ must then mean that D delimits E. Under the provisional assumption that E is due to the chance hypothesis **H**, one detects the transmission of information from s to r provided that D is detachable from E and $\mathbf{P}(D^* \mid \mathbf{H})$ has small probability. Moreover, one measures the amount of information transmitted from s to r as $-\log_2 \mathbf{P}(D^* \mid \mathbf{H})$. This account of information in terms of design is entirely consistent with the classical account in terms of symbol strings (cf. Shannon and Weaver, 1949).

References

Adams, E. W.: *The Logic of Conditionals*. Dordrecht: Reidel, 1975.

Appel, K., Haken, W., and Koch, J.: Every planar map is four colorable. *Illinois Journal of Mathematics*, 21:429–567, 1977.

Appiah, A.: *Assertion and Conditionals*. Cambridge, UK: Cambridge University Press, 1985.

Balcazar, J. L., Diaz, J., and Gabarro, J.: *Structural Complexity I*. Berlin: Springer, 1988.

Balcazar, J. L., Diaz, J., and Gabarro, J.: *Structural Complexity II*. Berlin: Springer, 1990.

Barrow, J. D.: *Theories of Everything: The Quest for Ultimate Explanation*. New York: Fawcett, 1991.

Barrow, J. D., and Tipler, F. J.: *The Anthropic Cosmological Principle*. Oxford, UK: Oxford University Press, 1986.

Bauer, H.: *Probability Theory and Elements of Measure Theory*. London: Academic Press, 1981.

Bertola, F., and Curi, U.: *The Anthropic Principle*. Cambridge, UK: Cambridge University Press, 1993.

Blum, M.: A machine-independent theory of the complexity of the recursive functions. *Journal of the Association for Computing Machinery*, 14:322–36, 1967.

Bogdan, R. J. (ed.): *Henry E. Kyburg, Jr. & Isaac Levi*. Dordrecht: Reidel, 1982.

Borel, E.: *Le Hasard*, 2nd ed. Paris: Presses Universitaires de France, 1948.

Borel, E.: *Probabilities and Life*, trans. M. Baudin. New York: Dover, 1962.

Borel, E.: *Probability and Certainty*, trans. D. Scott. New York: Walker, 1963.

Bourne, L. E., Jr., Dominowski, R. L., and Loftus, E. F.: *Cognitive Processes*. Englewood Cliffs, N.J.: Prentice-Hall, 1979.

Bradley, R., and Swartz, N.: *Possible Worlds: An Introduction to Logic and Its Philosophy*. Indianapolis, Ind.: Hackett, 1979.

Brooks, D. R., and Wiley, E. O.: *Evolution as Entropy*, 2nd ed. Chicago: University of Chicago Press, 1988.

Carnap, R.: *The Continuum of Inductive Methods*. Chicago: University of Chicago Press, 1952.

Carnap, R.: Statistical and inductive probability. 1955. In *Readings in the Philosophy of Science*, ed. B. A. Brody. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Carnap, R.: *Logical Foundations of Probability*, 2nd ed. Chicago: University of Chicago Press, 1962.

Carnap, R., and Jeffrey, R. C. (eds.): *Studies in Inductive Logic and Probability*, vol. 1. Berkeley, Calif.: University of California Press, 1971.

Chaitin, G. J.: On the length of programs for computing finite binary sequences. *Journal of the Association for Computing Machinery*, 13:547–69, 1966.

Chalmers, D. J.: *The Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford University Press, 1996.

Cicero, M. T.: *De Natura Deorum*, trans. H. Rackham. 45 BC. Cambridge, Mass.: Harvard University Press, 1933.

Coombs, C. H., Dawes, R. M., and Tversky, A.: *Mathematical Psychology: An Elementary Introduction.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Cramér, H.: *Mathematical Methods of Statistics.* Princeton: Princeton University Press, 1946.

Davidson, D.: *Essays on Actions and Events.* Oxford, UK: Clarendon Press, 1980.

Davies, P. (ed.): *The New Physics.* Cambridge, UK: Cambridge University Press, 1989.

Davis, M. D., and Weyuker, E. J.: *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science.* Orlando, Florida: Academic Press, 1983.

Dawkins, R.: *The Blind Watchmaker.* New York: W. W. Norton, 1987.

Day, J. P.: *Inductive Probability.* New York: Humanities Press, 1961.

de Duve, C.: *Vital Dust: Life as a Cosmic Imperative.* New York: Basic Books, 1995.

de Finetti, B.: *Theory of Probability*, 2 vols. New York: Wiley, 1974.

DeGroot, M. H., Fienberg, S. E., and Kadane, J. B.: *Statistics and the Law.* New York: Wiley, 1986.

Dembski, W. A.: Randomness by design. *Nous*, 25(1):75–106, 1991.

Dembski, W. A.: Randomness. In *The Routledge Encyclopedia of Philosophy*, ed. Edward Craig. London: Routledge (in press).

de Moivre, A.: *The Doctrine of Chances.* 1718. Reprint. New York: Chelsea, 1967.

Detlefsen, M., and Luker, M.: The four-color theorem and mathematical proof. *Journal of Philosophy*, 77(12): 803–20, 1980.

Devlin, K.: *Logic and Information.* Cambridge, UK: Cambridge University Press, 1991.

Diaconis, P.: *Group Representations in Probability and Statistics.* Hayward, Calif.: Institute of Mathematical Statistics, 1988.

DiVincenzo, D. P.: Quantum computation. *Science*, 270:255–61, 1995.

Dornstein, K.: *Accidentally, On Purpose: The Making of a Personal Injury Underworld in America.* New York: St. Martin's Press, 1996.

Drake, F., and Sobel, D.: *Is Anyone Out There?: The Scientific Search for Extraterrestrial Intelligence.* New York: Delacorte Press, 1992.

Dretske, F. I.: *Knowledge and the Flow of Information.* Cambridge, Mass.: MIT Press, 1981.

Dubins, L. E., and Savage, L. J.: *Inequalities of Stochastic Processes: How to Gamble if you Must.* New York: Dover, 1976.

Earman, J.: *A Primer on Determinism.* Dordrecht: Reidel, 1986.

Earman, J.: *Bayes or Bust.* Cambridge, Mass.: MIT Press, 1992.

Edwards, A. W. F.: Are Mendel's results really too close? *Biological Review,* 61:295–312, 1986.

Edwards, P. (ed.): *The Encyclopedia of Philosophy,* 8 vols. New York: Macmillan, 1967.

Eigen, M.: *Steps Towards Life: A Perspective on Evolution,* trans. P. Woolley. Oxford, UK: Oxford University Press, 1992.

Ellis, B. D.: A unified theory of conditionals. *Journal of Philosophical Logic,* 7:107–24, 1978.

Evans, C.: *The Casebook of Forensic Detection: How Science Solved 100 of the World's Most Baffling Crimes.* New York: Wiley, 1996.

Feller, W.: *An Introduction to Probability Theory and Its Applications,* vol. 1, 3rd ed. New York: Wiley, 1968.

Fine, K.: Vagueness, truth and logic. *Synthese,* 30:265–300, 1975.

Fine, T. L.: *Theories of Probability.* New York: Academic Press, 1973.

Finkelstein, M. O.: *Quantitative Methods in Law: Studies in the Application of Mathematical Probability and Statistics to Legal Problems.* New York: Macmillan, 1978.

Finkelstein, M. O., and Levenbach, H.: Regression estimates of damages in price-fixing cases. In *Statistics and the Law,* eds. M. H. DeGroot, S. E. Fienberg, and J. B. Kadane, pp. 79–106. New York: Wiley, 1986.

Finkelstein, M. O., and Levin, B.: *Statistics for Lawyers.* New York: Springer-Verlag, 1990.

Fisher, R. A.: Statistical tests. *Nature,* 136:474, 1935.

Fisher, R. A.: *The Design of Experiments,* 4th ed. Edinburgh: Oliver and Boyd, 1947.

Fisher, R. A.: *Statistical Methods and Statistical Inference.* Edinburgh: Oliver and Boyd, 1956.

Fisher, R. A.: *Experiments in Plant Hybridisation.* Edinburgh: Oliver and Boyd, 1965.

Freedman, D., Pisani, R., and Purves R.: *Statistics.* New York: Norton, 1978.

Frege, G.: *The Basic Laws of Arithmetic,* trans. and ed. M. Furth. Berkeley, Calif.: University of California Press, 1964.

Frege, G.: *Translations from the Philosophical Writings of Gottlob Frege,* 3rd ed., eds. P. Geach and M. Black. Oxford, UK: Basil Blackwell, 1985.

Gardner, M.: Arthur Koestler: Neoplatonism rides again. *World,* 87–9, 1 Aug. 1972.

Garey, M., and Johnson, D.: *Computers and Intractability.* New York: W. H. Freeman, 1979.

Giere, R. N.: Objective single case probabilities and the foundations of statistics. In *Logic, Methodology and Philosophy of Science IV,* eds. P. Suppes et al., pp. 467–83. Amsterdam: North Holland, 1973.

Gleick, J.: *Genius: The Life and Science of Richard Feynman.* New York: Random House, 1992.

Goguen, J. A.: The logic of inexact concepts. *Synthese,* 19:325–73, 1968–69.

Goldreich, O.: Randomness, interactive proof, and zero-knowledge, a survey. In *The Universal Turing Machine: A Half-Century Survey*, ed. R. Herken, pp. 377–405. Berlin: Kammerer und Unverzagt, 1988.

Goldreich, O., Goldwasser, S., and Micali, S.: How to construct random functions. *Journal of the Association for Computing Machinery*, 33(4):792–807, 1986.

Goldwasser, S., Micali, S., and Rackoff, C.: The knowledge complexity of interactive proofs. *Proceedings of the 17th Association for Computing Machinery Symposium on the Theory of Computing*, 291–305, 1985.

Goodman, N.: *Fact, Fiction, and Forecast*, 4th ed. Cambridge, Mass.: Harvard University Press, 1983.

Gorenstein, D.: *The Classification of Finite Simple Groups*, vol. 1. New York: Plenum, 1983.

Gorenstein, D.: Classifying the finite simple groups. *Bulletin of the American Mathematical Society*, 14:1–98, 1986.

Grim, P. (ed.): *Philosophy of Science and the Occult*, 2nd ed. Albany, N.Y.: State University of New York Press, 1990.

Grimmett, G. R., and Stirzaker, D. R.: *Probability and Random Processes*. Oxford, UK: Oxford University Press, 1982.

Gudder, S. P.: *Quantum Probability*. Boston: Academic Press, 1988.

Guth, A., and Steinhardt, P.: The inflationary universe. In *The New Physics*, ed. P. Davies, pp. 34–60. Cambridge, UK: Cambridge University Press, 1989.

Hacking, I.: *Logic of Statistical Inference*. Cambridge, UK: Cambridge University Press, 1965.

Hacking, I.: *The Emergence of Probability*. Cambridge, UK: Cambridge University Press, 1975.

Halliday, D., and Resnick, R.: *Fundamentals of Physics*, 3rd ed. extended. New York: Wiley, 1988.

Hamming, R. W.: *Coding and Information Theory*, 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall, 1986.

Hawking, S. W.: *A Brief History of Time: From the Big Bang to Black Holes*. New York: Bantam, 1988.

Heyting, A.: *Intuitionism: An Introduction*, 3rd rev. ed. Amsterdam: North Holland, 1971.

Horgan, J.: Can science explain consciousness? *Scientific American*, 271(1):88–94, 1994.

Howson, C., and Urbach, P.: *Scientific Reasoning: The Bayesian Approach*, 2nd ed. LaSalle, Ill.: Open Court, 1993.

Jackson, F.: *Conditionals*. Oxford, UK: Blackwell, 1987.

Jackson, F. (ed.): *Conditionals*. Oxford, UK: Oxford University Press, 1991.

Jung, C. G.: *Synchronicity*, trans. R. F. C. Hull. Princeton: Princeton University Press, 1973.

Kac, M., Rota, G.-C., and Schwartz, J. T.: *Discrete Thoughts*. Boston: Birkhäuser, 1986.

Kahneman, D., and Tversky, A.: Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3:430–54, 1972.

Kahneman, D., and Tversky, A.: On the psychology of prediction. *Psychological Review*, 80:237–51, 1973.
Kauffman, S. A.: *The Origins of Order: Self-Organization and Selection in Evolution*. New York: Oxford, 1993.
Kaufmann, A.: *Introduction to the Theory of Fuzzy Subsets*, vol. 1. Orlando, Fla.: Academic Press, 1975.
Keynes, J. M.: *A Treatise on Probability*. 1921. Reprint. Cambridge, UK: Cambridge University Press, 1988.
Klir, G. J., and Folger, T. A.: *Fuzzy Sets, Uncertainty, and Information*. Englewood Cliffs, N.J.: Prentice-Hall, 1988.
Knobloch, E.: Emile Borel as a probabilist. In *The Probabilistic Revolution*, vol. 1, eds. L. Krüger, L. J. Daston, and M. Heidelberger, pp. 215–33. Cambridge, Mass.: MIT Press, 1990.
Knuth, D. E.: *Seminumerical Algorithms*, 2nd ed. In *The Art of Computer Programming*, vol. 2. Reading, Mass.: Addison-Wesley, 1981.
Koestler, A.: *The Roots of Coincidence: An Excursion into Parapsychology*. New York: Random House, 1972.
Kolmogorov, A.: Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii* (in translation), 1(1): 3–11, 1965.
Krajicek, J., and Pudlak, P.: Propositional proof systems, the consistency of first order theories and the complexity of computations. *Journal of Symbolic Logic*, 54(3):1063–79, 1989.
Krüger, L., Daston, L. J., and Heidelberger, M.: *The Probabilistic Revolution*, 2 vols. Cambridge, Mass.: MIT Press, 1990.
Küppers, B.-O.: *Information and the Origin of Life*. Cambridge, Mass.: MIT Press, 1990.
Lakatos, I.: Falsification and the methodology of scientific research programmes. In *Criticism and the Growth of Knowledge*, eds. I. Lakatos and A. Musgrave, pp. 91–196. Cambridge, UK: Cambridge University Press, 1970.
Laplace, P. S. de: *Essai Philosophique sur les Probabilités*. 1814. Portion translated and reprinted into English. In *The World of Mathematics*, 4 vols., ed. J. R. Newman, pp. 1301–09. Redmond, Wash.: Tempus, 1988.
Lasota, A., and Mackey, M. C.: *Probabilistic Properties of Deterministic Systems*. Cambridge, UK: Cambridge University Press, 1985.
Laudan, L.: *Progress and Its Problems: Towards a Theory of Scientific Growth*. Berkeley, Calif.: University of California Press, 1977.
Le Cam, L.: *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.
Lehman, D.: *The Perfect Murder: A Study in Detection*. New York: The Free Press, 1989.
Lenoir, T.: *The Strategy of Life: Teleology and Mechanics in Nineteenth Century German Biology*. Dordrecht: Reidel, 1982.
Leslie, J.: *Universes*. London: Routledge, 1989.
Lewis, D.: Probabilities of conditionals and conditional probabilities. In *Conditionals*, ed. F. Jackson, pp. 76–101. Oxford, UK: Oxford University Press, 1991 (1976).

Lipton, P.: *Inference to the Best Explanation.* London: Routledge, 1991.
Luce, D., Bush, R. R., and Galanter, E. (eds.): *Handbook of Mathematical Psychology,* vol. 1. New York: Wiley, 1963.
Martin, W.: *The Kingdom of the Cults,* rev. ed. Minneapolis, Minn.: Bethany House, 1985.
Mazur, J. E.: *Learning and Behavior,* 2nd ed. Englewood Cliffs, N.J.: Prentice Hall, 1990.
McGee, V.: *Truth, Vagueness, and Paradox.* Indianapolis, Ind.: Hackett, 1991.
McKeon, R. (ed.): *The Basic Works of Aristotle.* New York: Random House, 1941.
Monod, J.: *Chance and Necessity.* New York: Vintage, 1972.
Mood, A. M., Graybill, F. A., and Boes, D. C.: *Introduction to the Theory of Statistics,* 3rd ed. New York: McGraw-Hill, 1974.
Newman, J. R. (ed.): *The World of Mathematics,* 4 vols. Redmond, Wash.: Tempus, 1988.
Paley, W.: *Natural Theology: Or, Evidences of the Existence and Attributes of the Deity, Collected from the Appearances of Nature.* 1802. Reprint. Boston: Gould and Lincoln, 1852.
Parthasarathy, K. R.: *Probability Measures on Metric Spaces.* New York: Academic Press, 1967.
Patterson, W.: *Mathematical Cryptology for Computer Scientists and Mathematicians.* Totowa, N.J.: Rowman and Littlefield, 1987.
Peirce, C. S.: The red and the black. 1878. In *The World of Mathematics,* 4 vols., ed. J. R. Newman, pp. 1311–17. Redmond, Wash.: Tempus, 1988.
Peirce, C. S.: The general theory of probable inference. 1883. In *Philosophical Writings of Peirce,* ed. J. Buchler, pp. 190–215. New York: Dover, 1955.
Peirce, C. S.: *Collected Papers,* 6 vols., eds. C. Hartshorne and P. Weiss. Cambridge, Mass.: Harvard University Press, 1931.
Peterson, C. R., and Beach, L. R.: Man as an intuitive statistician. *Psychological Bulletin,* 68:29–46, 1967.
Plantinga, A.: *Warrant and Proper Function.* New York: Oxford University Press, 1993.
Pletcher, G. K.: Coincidence and explanation. In *Philosophy of Science and the Occult,* ed. P. Grim, 2nd ed., pp. 205–13. Albany, N.Y.: State University of New York Press, 1990.
Polanyi, M.: *Personal Knowledge: Towards a Post-Critical Philosophy.* Chicago: University of Chicago Press, 1962.
Polanyi, M.: Life transcending physics and chemistry. *Chemical and Engineering News,* 21 Aug. 1967.
Polanyi, M.: Life's irreducible structure. *Science,* 113:1308–12, 1968.
Popper, K. R.: The propensity interpretation of probability. *British Journal for the Philosophy of Science,* 10:25–42, 1959.
Posner, M. I. (ed.): *Foundations of Cognitive Science.* Cambridge, Mass.: MIT Press, 1989.
Prigogine, I., and Stengers, I.: *Order out of Chaos: Man's New Dialogue with Nature.* New York: Bantam, 1984.

Ramsey, F. P.: *Philosophical Papers*, ed. D. H. Mellor. Cambridge, UK: Cambridge University Press, 1990.

Reichenbach, H.: *The Theory of Probability*. Berkeley, Calif.: University of California Press, 1949.

Reid, C.: *Hilbert–Courant*. New York: Springer, 1986.

Reid, T.: *Lectures on Natural Theology*, eds. E. Duncan and W. R. Eakin. 1780. Washington, D.C.: University Press of America, 1981.

Reinsel, R.: Parapsychology: An empirical science. In *Philosophy of Science and the Occult*, ed. P. Grim, 2nd ed., pp. 187–204. Albany, N.Y.: State University of New York Press, 1990.

Russell, B.: *The Problems of Philosophy*. Oxford, UK: Oxford University Press, 1959.

Salmon, W. C.: *The Foundations of Scientific Inference*. Pittsburgh, Pa.: University of Pittsburgh Press, 1967.

Savage, L.: *The Foundations of Statistics*, 2nd rev. ed. 1954. Reprint. New York: Dover, 1972.

Schwartz, B.: *Psychology of Learning and Behavior*, 2nd ed. New York: Norton, 1984.

Shannon, C. E., and Weaver, W.: *The Mathematical Theory of Communication*. Urbana, Ill.: University of Illinois Press, 1949.

Stalnaker, R. C.: *Inquiry*. Cambridge, Mass.: MIT Press, 1984.

Stigler, S. M.: *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, Mass.: Harvard University Press, 1986.

Sudbery, A.: *Quantum Mechanics and the Particles of Nature*. Cambridge, UK: Cambridge University Press, 1988.

Suppes, P., and Zinnes, J. L.: Basic measurement theory. In *Handbook of Mathematical Psychology*, vol. I, eds. D. Luce, R. R. Bush, and E. Galanter, pp. 1–76. New York: Wiley, 1963.

Swinburne, R.: *The Existence of God*. Oxford, UK: Oxford University Press, 1979.

Taylor, R.: *Action and Purpose*. Englewood Cliffs, N.J.: Prentice-Hall, 1966.

Thaxton, C. B., Bradley, W. L., and Olsen, R. L.: *The Mystery of Life's Origin*. New York: Philosophical Library, 1984.

Torgerson, W. S.: *Theory and Methods of Scaling*. New York: Wiley, 1958.

Tribe, L. B.: Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, 84:1329–93, 1971.

Turing, A. M.: Computing machinery and intelligence. *Mind*, 59:434–60, 1950.

Tversky, A., and Kahneman, D.: Belief in the law of small numbers. *Psychological Bulletin*, 76:105–10, 1971.

Tversky, A., and Kahneman, D.: Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5:207–32, 1973.

Tymoczko, T.: The four color problem and its philosophical significance. *Journal of Philosophy*, 76(2):57–83, 1979.

van Fraassen, B. C.: *The Scientific Image*. Oxford, UK: Oxford University Press, 1980.

van Inwagen, P.: *Metaphysics*. Boulder, Colo.: Westview Press, 1993.

van Lambalgen, M.: Algorithmic information theory. *Journal of Symbolic Logic*, 53(4):1389–400, 1989.
VanLehn, K.: Problem solving and cognitive skill acquisition. In *Foundations of Cognitive Science*, ed. M. I. Posner, pp. 527–79. Cambridge, Mass.: MIT Press, 1989.
von Mises, R.: *Probability, Statistics, and Truth*, 2nd ed. New York: Dover, 1957.
Waldrop, M. M.: *Complexity: The Emerging Science at the Edge of Order and Chaos*. New York: Simon and Schuster, 1992.
Walley, P.: *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall, 1991.
Wegener, I.: *The Complexity of Boolean Functions*. Stuttgart: Wiley-Teubner, 1987.
Wicken, J. S.: *Evolution, Thermodynamics and Information*. New York: Oxford University Press, 1987.
Wilansky, A.: *Topology for Analysis*. Malabar, Fla.: Krieger, 1983.
Wilson, C., and Weldon, J.: *Close Encounters: A Better Explanation*. San Diego: Master Books, 1978.
Wimsatt, W.: Randomness and perceived-randomness in evolutionary biology. *Synthese*, 43:287–329, 1980.
Wittgenstein, L.: *Culture and Value*, ed. G. H. von Wright, trans. P. Winch. Chicago: University of Chicago Press, 1980.
Wittgenstein, L.: *Remarks on the Foundations of Mathematics*, eds. G. H. von Wright, R. Rhees, and G. E. M. Anscombe, trans. G. E. M. Anscombe, rev. ed. Cambridge, Mass.: MIT Press, 1983.
Yockey, H. P.: *Information Theory and Molecular Biology*. Cambridge, UK: Cambridge University Press, 1992.
Zabell, S. L.: Buffon, Price, Laplace: Scientific attribution in the 18th century. *Archive for History of Exact Sciences*, 39(2):173–81, 1988.
Zurek, W. H. (ed.): *Complexity, Entropy and the Physics of Information*. Redwood, Calif.: Addison-Wesley, 1990.

Index