OCKHAM'S RAZORS

A User's Manual

ELLIOTT SOBER

Ockham's Razors

Ockham's razor, the principle of parsimony, states that simpler theories are better than theories that are more complex. It has a history dating back to Aristotle, and it plays an important role in current science. The razor also gets used in everyday life and in philosophy. This book evaluates the principle and discusses its many applications. Fascinating examples from different domains provide a rich basis for contemplating the principle's promises and perils. It is obvious that simpler theories are beautiful and easy to understand; the hard problem is to figure out why the simplicity of a theory should be relevant to saying what the world is like. In this book, the ABCs of probability theory are succinctly developed and put to work to describe two "parsimony paradigms" within which this problem can be solved.

ELLIOTT SOBER is Hans Reichenbach Professor and William F. Vilas Research Professor in the Department of Philosophy, University of Wisconsin, Madison. In 2014 the Philosophy of Science Association awarded him the Carl Gustav Hempel Award for lifetime achievement in philosophy of science. His publications include *Evidence and Evolution: The Logic Behind the Science* (2008), *Did Darwin Write the Origin Backwards?: Philosophical Essays on Darwin's Theory* (2011), and *Unto Others: The Evolution and Psychology of Unselfish Behavior* (1998, coauthored with David Wilson).

Ockham's Razors

A User's Manual

ELLIOTT SOBER

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107692534

First published 2015

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Sober, Elliott.
Ockham's razors : a user's manual / Elliott Sober for Ezra William Didier-Sober.
pages cm
Includes bibliographical references and index.
ISBN 978-1-107-06849-0 (hardback paper) 1. Reasoning. 2. Problem solving.
3. Simplicity (Philosophy) 4. William, of Ockham, approximately 1285–approximately
1349. 5. Philosophy, Medieval. I. Title.
BC177.S68 2015
121′.6 – dc23 2015006512

ISBN 978-1-107-06849-0 Hardback
ISBN 978-1-107-69253-4 Paperback

for Ezra William Didier-Sober

Contents

Acknowledgments

I am very grateful to a very large number of people for the help they have given me on this project. My debts are to Marilyn McCord Adams, Kristin Andrews, Martin Barrett, John Basl, David Baum, Craig Baxter, David Blumenfeld, Hayley Clatterbuck, Jordan Ellenberg, Branden Fitelson, Simon Fitzpatrick, Malcolm Forster, Hank Frankel, Martha Gibson, Mike Goldsby, Alison Gopnik, Paula Gottlieb, Casey Hart, Dan Hausman, Casey Helgeson, Cecelia Heyes, Jack Justus, Sharon Kaye, Matt Kopec, Marc Lange, James Lennox, Brian McLoone, Jeffrey McDonough, James Messina, Wayne Myrvold, Steven Nadler, Greg Nirshberg, Shannon Nolen, John Norton, Ron Numbers, David O'Brien, Emi Okayasu, Sarah Paul, Trevor Pearce, Mike Roche, William Roche, Bas Rokers, Eric Sampson, Daniel Schneider, Michael Shank, Michael Schon, Russ Shafer-Landau, Michael Shank, Larry Shapiro, Alan Sidelle, George Smith, Dennis Stampe, Mike Steel, Eric Stencil, Reuben Stern, Michael Stoelzner, Michael Titelbaum, Michael Tomasello, Olav Vassend, Peter Vranas, Naftali Weinberger, Allen Wood, and Rega Wood. I also have benefitted from holding a University of Wisconsin William F. Vilas Professorship.

Introduction

Two of Barcelona’s architectural masterpieces are as different as different could be. The Church of the Holy Family, designed by Antoni Gaudí (1852–1926), is only a few miles from the German Pavilion, built by Mies van der Rohe (1886–1969). Gaudí’s church is flamboyant, complex, and irregular. Mies’s pavilion is tranquil, simple, and rectilinear. Mies, the apostle of minimalist architecture, used the slogan “less is more” to express what he was after.[1] Gaudí never said “more is more,” but his buildings suggest that this is what he had in mind.

One reaction to the contrast between Mies and Gaudí is to choose sides based on a conviction concerning what all art should be like. If all art should be simple or if all art should be complex, the choice is clear. I reject both of these monistic norms; I am a pluralist about artistic simplicity and complexity because I see value in both. True, there may be extremes that are beyond the pale. We are alienated by art that is far too complex and bored by art that is far too simple, but between those two extremes there is a vast space of possibilities.[2] Different artists at different times and places have had different goals. Artists are not in the business of trying to discover the uniquely correct degree of complexity that all artworks should have. There is no such timeless ideal.

Science is different, at least according to many scientists. Einstein (1933) spoke for many when he said that “it can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation

[1] Robert Browning uses “less is more” in his 1855 poem “Andrea del Sarto, called ‘The Faultless Painter’.”

[2] Here it is useful to contrast Arnheim’s (1954) idea that good art creates order with Peckham’s (1967) thesis that art is valuable because of the disorder it induces.

of a single datum of experience." This influential point of view holds that the search for simple theories is not optional; rather, it is a requirement of the scientific enterprise. When theories get too complex, scientists reach for Ockham's razor, the principle of parsimony, to do the trimming. This principle says that a theory that postulates fewer entities, processes, or causes is better than a theory that postulates more, so long as the simpler theory is compatible with what we observe. This formulation of the principle is preliminary; it will be fine-tuned in what follows. For example, what does "better" mean?

The long history of the principle of parsimony reveals that there were several such principles in play, not just one. That's why the title of this book is in the plural; the subject at hand is Ockham's *razors.* Different thinkers have meant different things by parsimony and different justifications for principles of parsimony have been constructed. My goal in this book is to describe this diversity and to determine when parsimony is relevant and when it is not. It is obvious that simple theories may be beautiful and easy to remember and understand. The hard problem is to explain why the fact that one theory is simpler than another tells you anything about the way the world is.

Ockham's razor was prominent in the history of science, but it remains important in contemporary science as well. It was used to defend Copernican astronomy, but it now plays a role in evolutionary biology and in cognitive psychology. Scientists, philosophers, and statisticians have had their separate insights into Ockham's razor, so this book will touch a number of bases. Aristotle will be shoulder to shoulder with Akaike, and Newton and Darwin will each have their say. However, there is more to Ockham's razor than the use that is made of it in science. The principle of parsimony is deployed when non-scientists reason about non-scientific questions. This is no surprise, if scientific modes of reasoning are continuous with forms of reasoning that are used in everyday life. Another non-scientific use of parsimony is more puzzling. Philosophers sometimes appeal to Ockham's razor when they evaluate philosophical theories. If the arts and the sciences diverge in the status they accord to simplicity, where does philosophy belong? Is philosophy more like the arts or more like the sciences in terms of the value it should assign to parsimony?

This book is not simple, but the sequencing of chapters is. Chapter 1 provides a brief history of the divergent ideas that were developed before 1900 concerning why the principle of parsimony makes sense. Chapter 2 introduces the ABCs of probability theory and puts them to work; after describing some failed attempts in the twentieth century to use probability to elucidate and justify Ockham's razor, I describe two "parsimony paradigms" that allow justifications of the principle of parsimony to be made clear. Chapter 3 is about phylogenetic inference in biology and Chapter 4 is about chimpanzee mind-reading in psychology. In both instances, scientists have invoked principles of parsimony and there has been scientific controversy about parsimony's relevance. In Chapter 5, I examine the use of Ockham's razor in philosophy.

Parsimony arguments that draw conclusions about the way the world is from the fact that one theory is more parsimonious than another differ from each other at two levels. First, some of them succeed while others fail. Second, the successful arguments succeed for different reasons, and the unsuccessful arguments go wrong in different ways. Fortunately, this second sort of heterogeneity is not endless; there are recurrent patterns that bring order to the diversity found among the successes and to the diversity found among the failures. I have tried to find a simple philosophical framework for understanding parsimony, but I have been guided by the idea that the framework should not be too simple; it must be complex enough to capture the phenomena. As a philosopher, I am on the side of Mies and Einstein.

1 A history of parsimony in thin slices (from Aristotle to Morgan)

In this chapter I discuss some interesting historical cases in which Ockham's razor has been put to use, but the main focus is on the history of attempts to justify the principle of parsimony. Why buy the idea that simpler theories are better than theories that are more complex? A variety of answers to this question have been offered. Some think that the principle of parsimony is justified by what we have learned from observing nature. Others think that the razor has a theological justification. Still others think that valuing parsimonious theories is part of what it means to be rational. And yet another faction regards the principle as rock bottom – they think the principle is correct but that it can't be justified at all.

The snapshots I present in this chapter are varied, but there is agreement on something important – that parsimony is not an optional, aesthetic frill. Was this historical consensus on the right track? In this chapter, I use this history to begin assembling epistemological tools for assessing the status of parsimony considerations; the search for tools will continue in subsequent chapters.

The naming ceremony

It may seem that the inevitable start of our story is William of Ockham (*c.* 1285–1348), since it is he for whom the principle was named. The naming ceremony apparently occurred long after Ockham. In his 1649 book *On Christian Philosophy of the Soul*, Libert Froidmont (1587–1653) claims to coin the phrase. He speaks of a "*novacula occami*" (a novacula is a small knife or razor) in describing one of Ockham's critics, Gregory of Rimini (d. 1358), who

> excellently drew Ockham's razor . . . [against] its author, since Ockham multiplied entities without necessity . . . However, I call this axiom

> Ockham's and the nominalists' razor because they used that [axiom] to trim and shave off all distinct entities, leaving a plurality only of names. Hence they are designated by the name "nominalists." (Hübener 1983, pp. 84–85)[1]

It is pleasing that this name for the principle comes to us from someone whose name means *cold mountain*. Quine (1953a, p. 4) says that the principle of parsimony expresses a "taste for desert landscapes," but he could just as easily have said that the principle evokes the austere beauty of a frozen summit.

The much-cited slogan "entities should not be multiplied beyond necessity" does not appear in Ockham's writings, but he does say that "it is futile to do with more what can be done with fewer"[2] and that "plurality should not be posited without necessity."[3] Ockham was not the first person to have endorsed these maxims (Thorburn 1918); similar formulations are to be found in the writings of Thomas Aquinas (1224–1274) and in those of Ockham's teacher Duns Scotus (1266–1308).[4] "Ockham's razor" is an example from philosophy of Stigler's (1980) Law of Eponymy, that no scientific idea is named after its original discoverer. Stigler named the law after himself, even though he says that it was the sociologist Robert Merton who discovered it. With tongue in cheek, Stigler chose the law's name so that it would be an instance of itself.

Unfortunately, Ockham didn't say much about why the principle of parsimony ought to be followed (Adams 1987, p. 158). He relied on the fact that he and other philosophers found it sensible; the maxim was common ground, so Ockham felt he could use it as an undefended premise in his arguments.

[1] I am grateful to Rega Wood for this translation from the Latin and for calling my attention to Hübener's essay, which corrects Thorburn's (1918, p. 349) claim that the *novaculum nominalium* metaphor was unknown in the seventeenth century and was invented in the eighteenth by Condillac in his 1746 *Origine des Connaissances Humaines*. Thorburn (pp. 349–350) also says that the English variant, "Occam's razor," first appeared in William Hamilton's (1852, p. 590) *Discussions*.

[2] Ockham, 1986b, Tractatus de Corpore Christi, cap. 28, pp. 157–158.

[3] Ockham, 1986a, Ordinatio I, d.27, q.2, p. 202.

[4] Here are two relevant passages: (1) "If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices" (Aquinas 1945, p. 129); (2) "we should always posit fewer things when the appearances can be saved thereby . . . therefore in positing more things we should always indicate the manifest necessity on account of which so many things are posited" (Duns Scotus 1998, p. 349).

Ockham did not claim that pluralities never exist nor that complex theories are never true. For example, he says that the road to salvation that God created is *un*parsimonious and notes that this does not mean that the arrangement is flawed or unfitting (Adams 1987, p. 159). At first glance, his maxim seems to say nothing about the way the world actually is; it tells you when you should decline to postulate the existence of something, not whether that something exists. In contrast, there are other versions of the principle of parsimony that unmistakably make claims about the world. Ockham was well aware of a version of this type. For Ockham and other medievals, Aristotle was a starting point for philosophical reflection. Aristotle, not Ockham, provides a better beginning for our story, the seventeenth-century naming ceremony notwithstanding. We'll return to Ockham in due course.

Aristotle's principle that nature does nothing in vain

In his book *Movement of Animals*, Aristotle (384–322 BCE) says that "nature does nothing in vain, but always does what is best, from among the possibilities, for the substantial being of each kind of animal" (2, 704b11–17).[5] Aristotle invokes this principle in many passages. For example, in *The Generation of Animals* (II 5), he asks why males exist. Aristotle raises the question because he thinks that uniparental reproduction is a real possibility. Citing his principle, he concludes that males must play *some* functional role in reproduction. This led him to ask what biparental reproduction contributes that uniparental reproduction cannot. Although this application of Aristotle's principle may seem sensible, it must sound naïve and overstated to the modern ear that is schooled in the lessons of avoiding uncritical adaptationism (Gould and Lewontin 1979). "Nature does nothing in vain" sounds wrong when you think about ear lobes, eye colors, philtrums (the grooves under our noses), and male nipples. Aristotle's maxim seems to conflict with Charles Darwin's (1807–1882) comment in *The Origin of Species* that nature is peppered with structures that "bear the plain stamp of inutility" (Darwin 1859, p. 480).

Aristotle has a reply. When he refers to *nature* in his principle, he doesn't mean what we mean by "nature" – the totality of everything that happens in space and time. Rather, he has in mind the individual *natures* that different organisms possess. It is because of the tiger's nature that tigers have

[5] I am grateful to Paula Gottlieb for helpful discussion on Aristotle.

sharp teeth. Aristotle's natures give rise to the natural tendencies that lead organisms to develop various traits. These natural tendencies are subject to interference; tigers born without sharp teeth are still tigers. This is Aristotle's *natural state model* (Sober 1980). We can combine the modern concept of *nature* with Aristotle's concept of *natures* to formulate an Aristotelian point: not everything that happens *in nature* happens *because of the natures* of individual organisms. Aristotle says that his principle applies only to traits that are universal within a kind (*Generation of Animals*, V.1). Ear lobes and eye colors vary; Aristotle has no problem with the idea that they lack functions (Lennox 2001). Aristotle also allows that some traits may be byproducts; in *On the Soul* (III.12.434a31–2) he says that " . . . everything in nature exists for the sake of something or will be an accident of those things which are for the sake of something." Male nipples are not a counterexample to "nature does nothing in vain."

Aristotle's principle is *teleological*; it says that the nature of a thing drives it to achieve a goal. Although Aristotle didn't think that everything has a *telos*, he did think that teleology (the idea of "final causes") should not be restricted to biology. For example, in *On the Heavens* (ii, clr, 296b, 310b, 2–5), he says that it is in the nature of heavy objects in the sublunar sphere to fall towards the center of the Earth in a straight line, though, of course, objects often fail to do this. Rocks are goal-directed, just as tigers are.

The maxim that often comes to mind when people now think of Ockham's razor is negative, but it is usually understood to have a positive complement. There is "do not accept a postulate if it is *not* needed to explain anything," but there is also "accept a postulate if it *is* needed to explain something." Aristotle's principle is not a restatement of either. It isn't purely negative, and its positive message is teleological. Even so, Aristotle's *nature does nothing in vain* is an important part of the history of Ockham's razor, as I'll explain.

What justification does Aristotle provide for his maxim? In *Physics* (II.8) and in *Parts of Animals* (I.1), he defends the need for teleology by considering the alternative, which he calls "chance." This is the view that Aristotle attributes to Empedocles, who held that the order we see in the universe is due to objects moving randomly in the void. Some objects stick together. Stable combinations persist while unstable ones fall apart.[6] Aristotle claims that chance leaves various facts unexplained, pre-eminently the regular

[6] Empedocles's idea resembles Darwin's idea of natural selection.

development of the organisms in a species as they move from embryo to juvenile to adult. He also thinks that statements about chance presuppose the truth of other statements that are teleological. For example, if two people meet on the street "by chance," this is because the two had goals that did not include their meeting. This is Aristotle's "philosophical justification" for *nature does nothing in vain* (Lennox 2001, p. 214).

This justification is full of holes. First, the claim that teleological concepts are needed to explain *some* facts about nature is not enough to show that nature does *nothing* in vain. Second, Aristotle is wrong in his claim that Empedocles can't explain the highly regular ontogenies of organisms. Even if chance explains the origination of various features of organisms way back when, that does not mean that the descendants of those ancient ancestors must obtain those features by the same chancy process; they can obtain them by inheritance from their parents. Third, many chance statements are true without any implication of teleology. The example of two people meeting "by chance" is atypical; as we will see in the next chapter, statements about probability can be true without there being a goal or a plan. Finally, statements about function and purpose can be true without Aristotle's natural state model being right. One alternative is to use Darwin's theory of natural selection. According to this approach, the reason hearts have the function of pumping blood, not of making noise, is that hearts evolved because they pump blood, not because they make noise (Wright 1976).[7]

There is another passage in which Aristotle suggests a different justification for his principle. In *Generation of Animals* (V.8.788b21), he says that "... we assume, basing our assumptions on what we see, that nature does nothing in vain in so far as is possible in each case." The verb "to see" (from *οράω*) is not metaphorical, and so the question arises of how our visual observations tell us that nature does nothing in vain. Perhaps Aristotle's idea is that we observe that the principle *works*; it is supported by its many successful applications. The principle tells you to seek out the function that a biological structure has. When you discover the function, the principle scores a success. Understood in this way, the principle isn't prior to what we discover when we investigate nature, but is a useful after-the-fact summary of what those investigations have yielded (Gottlieb and Sober forthcoming). This defense of

[7] Another alternative to Aristotle is the account of function-claims developed by Cummins (1975). There are others.

nature does nothing in vain appeals to observations. Aristotle does not think that his principle is *a priori*.

Aristotle endorses a second principle when he discusses his idea of the *unmoved mover*. Unlike *nature does nothing in vain*, this second principle has nothing to do with goals. Aristotle thinks that each thing that moves is caused to move by something outside itself. He also holds that there are no actual infinities. Aristotle thinks that when we trace a present motion back to an earlier motion, and that earlier motion back to one that is still earlier, the chain must reach its first member, an unmoved mover, after finitely many steps. This conclusion leaves open how many unmoved movers there are. In *Physics* (8.6.259a), Aristotle fills in this blank: "We ought . . . to suppose that there is one rather than many . . . Here it is sufficient to assume only one mover, the first of unmoved things, which being eternal will be the principle of motion to everything else." Aristotle's example concerns the specifics of motion, but the principle he invokes is more general. The point of relevance is that Aristotle is using a minimum principle to justify a conclusion about the way the world is, and his principle expresses no commitment to teleology.[8] There is one unmoved mover rather than several because postulating just one suffices to explain the motions we observe. As we will see, subsequent thinkers use Aristotle's phrase *nature does nothing in vain* but give it a meaning that approximates his minimum principle. They thereby leave Aristotle's teleology behind.

How Ockham wields his razor

Ockham was a nominalist – he denied the existence of "universals." What are they? Let us begin with what they are not. Universals contrast with particulars. Particulars are the individual things that populate the universe – you, the hive of bees in the park, the Eiffel Tower, Planet Earth. Universals are supposed to be the properties that multiple individual things have in common. For example, Socrates and Plato are particulars, and both philosophize. Does that

[8] This isn't what is going on in another passage in which Aristotle appeals to the idea that less is more. In the *Posterior Analytics* (1.25.2), he says "we may assume the superiority *ceteris paribus* of the demonstration which derives from fewer postulates or hypotheses – in short, from fewer premises; for given that all these are equally well known, where they are fewer knowledge will be more speedily acquired, and that is a desideratum." Here the advantage attributed to minimality has nothing to do with determining what is true.

mean that the statement "Socrates and Plato both philosophize" describes three things – the two men plus the universal to which the two individuals belong? Ockham's answer is *no.* According to Ockham, the two individuals exist and there is Socrates's philosophizing and Plato's as well. Each of these properties is unique to the individual who has it. There is no universal here – there exists no property of philosophizing that is shared among particulars. It is the human mind's invention of concepts (in this instance, the concept of philosophizing) that fosters the illusion the universals exist.

Was Ockham's nominalism motivated by his passion for parsimony? Did he deny the existence of universals because he thought they aren't needed to explain anything? This is what Froidmont says in the passage from him that I quoted, but the claim is not borne out by Ockham's writings. Ockham rejects universals because he thinks that the idea of a universal is *contradictory*. He didn't think that universals *might* exist though considerations of parsimony should lead us to deny that they do. His denial cuts deeper; he thought that universals *cannot* exist. They are like round squares and married bachelors. The universal of being human is supposed to be a single thing, and it's also supposed to be found in each individual human being. So the universal is both one and many, which is impossible. Ockham didn't need Ockham's razor to be a nominalist (Spade and Panaccio 2011).[9]

So where in his theorizing does Ockham actually use his razor? One place is in his discussion of what it takes for something to change. Medieval philosophers, inspired by Aristotle, wanted to have a theory that characterizes what happens when a particular changes. For example, when Socrates changes from healthy to sick, what's going on? Many philosophers held that all change, including this one, must involve the production or destruction of a thing. Since Socrates exists before and after he changes from healthy to sick, there must be some additional thing that is involved, or so they thought. The "things" these philosophers were thinking about are not physical objects – for example, disease-causing micro-organisms. Rather, they were thinking about "qualities." There is a thing called "Socrates's health"; it was annihilated and replaced by a thing called "Socrates's sickness."

[9] Ockham's criticism of universals may remind you of another puzzle – the puzzle of the trinity. How can God be three persons and one person at the same time? Ockham thought that this idea is illogical. However, he did not reject the trinity; rather, he rested his religious conviction on faith, not reason (Kaye 2007).

Ockham thought that this theory of change is sometimes correct, but that there are cases where you can explain change without postulating a thing that is created or annihilated. In his *Summa Logicae* (I, c. 56, OPh I, pp. 182–183), Ockham says that changes due to locomotion are like this. For example, when a coiled rope is stretched out, the change is due simply to the physical parts being rearranged; there is no need to claim that there is a thing that is present at one time and absent at another. Ockham's point here is not that this additional thing *can't* exist; rather, he is arguing that there is no reason to postulate its existence (Adams 1987, pp. 277–285).

This philosophical discussion of how change should be understood may sound like it is worlds away from modern science. Scientists now want to know what causes different changes; the metaphysics of change in general – what change *is* – is not their cup of tea. In Chapter 5 I'll discuss whether parsimony arguments in philosophy are ever similar to parsimony arguments in science. But, for now, let's leave Ockham's metaphysics behind and turn to a second context in which he wields his razor.

In her excellent book on Ockham, Marilyn McCord Adams (1987, pp. 160–161) draws our attention to a passage in which Ockham wonders whether matter in the heavens is the same kind of thing as matter here on Earth. Ockham notes that no conclusive proof is possible here, but he does think that a "persuasive argument" can be given. It is this:

> it appears to me . . . that the matter in the heavens is of the same kind as the matter here below. And this is because plurality should never be posited without necessity, as has often been said. Now, however, there appears no necessity to posit matter of a different kind here and there, since everything that can be saved by [positing] diversity in matter can just as well or better be saved by [positing matter] identical in kind.[10]

The "saving" that Ockham has in mind is a theory's accurately representing what we observe. Here Ockham is applying his razor to justify a conclusion that is recognizably scientific: celestial and terrestrial motion obey the same laws. Ockham could not have known how important this example would be three and a half centuries after his death.

[10] The passage from Ockham is from *Reportatio* II, q. 18 (OTh V, 404). The translation from the Latin is by Rega Wood.

Ockham's principle of parsimony is really two principles. There is the *razor of silence* and there is the *razor of denial* (Sober 2009b). The contrast between the two razors parallels the difference between agnosticism and atheism. The razor that Ockham uses in discussing locomotion leads him to remain silent about whether a thing is created or destroyed when a rope is uncoiled. He doesn't deny that something is destroyed when the rope is uncoiled; instead, he says that we don't need this postulate to explain what happens. Ockham also uses the razor of silence in his discussion of other categories from Aristotelian metaphysics, and this has led some commentators to suggest that the razor of silence is the only razor that Ockham deploys (Spade and Panaccio 2011). But Ockham's discussion of celestial and terrestrial matter shows that there is a second razor in his shaving kit. Here he considers two incompatible theories and uses the razor to decide which is better. The upshot is to deny the less parsimonious hypothesis, not to remain silent about whether it is true.

Geocentric and heliocentric astronomy

In his 1543 work *On the Revolutions of the Heavenly Spheres*, Nicholas Copernicus (1473–1543) proposed an astronomical model in which the Earth and the other planets revolve around the Sun.[11] Copernican heliocentrism eventually displaced the earth-centered model of Claudius Ptolemy (90–168), which had ruled the roost for some 1400 years. Few changes in the history of science have been as momentous.

Scientists who think of their subject as data-driven find the Copernican revolution difficult to comprehend. There was no observation that Copernicus could cite that refuted the Ptolemaic model. Such observations would come later – for example, from Galileo's using his telescope in 1610 to observe the phases of Venus.[12] Before then, each of the two models did a good job of accurately representing the angular motions of the Sun and the planets as seen from Earth. For Copernicus, the tie-breaker is that his heliocentric theory is simpler. In Book 1, Chapter 10, he says "we thus follow Nature, who

[11] I am indebted in this section to Malcolm Forster and Mike Shank for very useful discussion.

[12] Although Galileo's observations did establish that Venus goes around the Sun, they did not discriminate between the Copernican system and a third alternative – the system of Tycho Brahe (1546–1601), who held that the Sun goes around the Earth and that other planets go around the Sun.

producing nothing in vain or superfluous, often prefers to endow one cause with many effects." Copernicus's only pupil, Georg Joachim Rheticus, follows his teacher's lead, except that he adds God to Nature in explaining why the greater simplicity of the heliocentric model is a reason to think that the model is true. Here is what Rheticus says in his work of 1540, *A First Account of Copernicus's Book on Revolutions*:

> Mathematicians as well as physicians must agree with the statements emphasized by Galen here and there: "Nature does nothing without purpose" and "So wise is our Maker that each of his works has not one use, but two or three or often more." Since we see that this one motion of the Earth satisfies an almost infinite number of appearances, should we not attribute to God, the creator of nature, that skill which we observe in the common makers of clocks? For they carefully avoid inserting in the mechanism any superfluous wheel or any whose function could be served better by another with a slight change of position. (Rosen 1959, pp. 137–138)

Rheticus's mention of Galen points to the biological pedigree of a principle that here finds its application in astronomy.

Copernicus's strategy of justification-by-simplicity was passed on to his scientific descendants. In his 1632 *Dialogue on the Two Chief World Systems*, Galileo Galilei (1564–1642) has his character Salviati cite simplicity as a reason in favor of heliocentrism:

> if precisely the same effect follows whether the Earth is made to move and the rest of the universe stay still, or the Earth alone remains fixed while the whole universe shares one motion, who is going to believe that nature (which by general agreement does not act by means of many things when it can do so by means of few) has chosen to make an immense number of extremely large bodies move with inconceivable velocities, to achieve what could have been done by a moderate movement of one single body around its own center? (Galileo 1632, pp. 116–117)

A few pages later, Salviati invokes the authority of Aristotle to support something that Aristotle and his followers had never endorsed, heliocentrism:

> It is much more probable that the diurnal motion belongs to the Earth alone than to the rest of the universe excepting the Earth. This is supported by a very true maxim of Aristotle's, which teaches that *frustra fit per plura quod potest fieri per pauciora* [more causes are in vain when fewer suffice]. (Galileo 1632, p. 123)

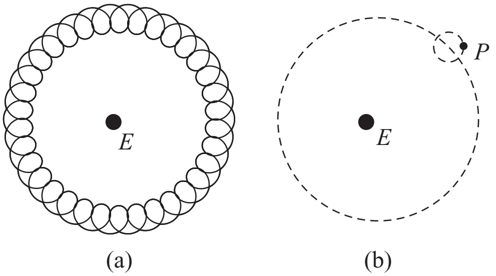


Figure 1.1

Using Aristotle against the Aristotelians is a clever rhetorical strategy, even though the slogan is not to be found in Aristotle (Myrvold 2003, p. 403).

Given that Copernicus and his followers defended heliocentrism by appeal to Ockham's razor, let's now turn to an examination of the geocentric theory that they opposed. Ptolemy worked out his geocentric model so that it would fit the detailed observations that astronomers had made of the heavens.[13] To get his theory to capture these observations, he postulated *epicycles*, an idea that Apollonius of Perga had developed in the third century BCE. The Greek term "epicycle" means *on the circle*. In Ptolemaic astronomy, planets revolve around the Earth in curly-cues (the technical term is "epitrochoid"), as shown in Figure 1.1(a). The curly-cue is the upshot of the two circular motions shown in Figure 1.1(b). The planet moves along a smaller circle that is centered on a larger circle that itself goes round the Earth. A planet's motion is like the motion of a point on the edge of a spinning top that sits on the edge of a moving merry-go-round. The big circle is the deferent; the small circle is the epicycle. Although Ptolemy thought that the Earth is at the center of the universe (i.e., at the exact center of the sphere that holds the fixed stars), it isn't exactly centered relative to the motions of planets. With respect to a

[13] Ptolemy develops a detailed physical model in his *Planetary Hypotheses*, not in his much more widely known work, *The Almagest*. *The Almagest* was mainly devoted to providing a device for describing and predicting what Earth-bound observers see, though there is discussion of a geocentric model in Book 9. What mattered to the descriptive and predictive task was "saving the phenomena" (i.e., capturing what people observe), with no serious commitment to the thesis that the Earth is "really" at the center of the universe. Copernicus and his followers took *The Almagest* to be offering a theory of what moves and what is stationary, though, in fact, this is not its main subject.

planet, the Earth is off-center a bit, sitting a short distance from the geometric center of the planet's deferent, in an "eccentric" position. If you draw a line from the Earth through the deferent's center, your line will then reach a third point, the equant. The distance from the Earth to the center of the deferent is the same as the distance from the deferent's center to the equant. This third point is important to the Ptolemaic system because the angular rate at which the center of a planet's epicycle moves along its deferent is constant when the rate is calculated from the equant.

Why did Ptolemy and his followers place the Earth at the center with planets doing curly-cues around it? The question divides in two: why did they think that the Earth is stationary while other objects revolve around it? And given an answer to the first question, why did they introduce epicycles to describe what they saw? The first question is answered by the fact that when a heavy object is dropped from the top of a tall building or from a tree, it falls straight down. And objects on the surface of the Earth don't normally fly off the surface in the way that objects on the edge of a moving merry-go-round often fly off the merry-go-round. If the Earth moves, shouldn't these everyday observations be otherwise? Physics would later show that the answer to this question is *no*, but that is not what most ancient astronomers thought. Aristotle embraced the idea that the Earth is stationary and is at the center of the universe, but the grip of geocentrism cannot be completely explained by a simplistic appeal to his influence. Our experience is that the Earth is solid beneath our feet; when we look into the sky, we see objects moving. The stars have uniform motions, and they move in unison across the sky. In contrast, the planets exhibit *retrograde motion*, and they don't do this in unison. If you watch a planet over the course of a year, you will see it move smoothly across the sky, then back up, and then move forward again. The word "planete" means *wanderer* in Greek. A deferent-plus-epicycle model captures retrograde motion. Viewed from above, Mars moves around Earth in a curly-cue; viewed from Earth, Mars moves forward, then back, and then forward again. Retrograde motion occurs when a planet moves on its epicycle in a direction opposite to the direction in which the epicycle is moving on its deferent.

By putting the Sun near the center, Copernicus was able to explain retrograde motion without postulating epicycles. As long as the Earth and the other planets revolve around the Sun with planets closer to the Sun completing their orbits faster than ones that are farther out, observers on

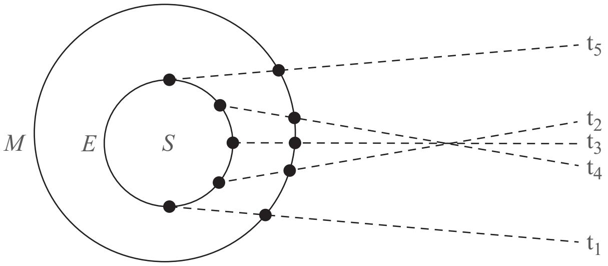


Figure 1.2

Earth will see the wanderers doing retrograde motion. This is illustrated in Figure 1.2, which traces hypothetical locations of Earth and Mars as each revolves around the Sun from time t_1 to time t_5. In this figure, the Earth gets about two-thirds of the way round the Sun during a stretch of time in which Mars gets only about a fifth of the way round. At each time, Mars is seen from Earth as having a certain position in the sky. Mars seems to move forward from time t_1 to time t_2, then back up in the passage from t_2 to t_3 to t_4, and then move forward again from t_4 to t_5.[14] Viewed from above, Mars isn't doing curly-cues, but when we Earthlings look at Mars, we see retrograde motion.

This brief account may lead you to think that Copernicus didn't use epicycles and that this made his theory more parsimonious than Ptolemy's. The trouble with this suggestion is that Copernicus *did* use epicycles. He did so because his observations did not permit him to say that planetary paths are circles. Rather than simply saying that planets move in ovals and leaving it at that, Copernicus felt compelled to talk about deferents and epicycles. In this respect, Copernicus was old-school; he shared with Ptolemy a commitment to the idea that circles are perfect and that the motions of heavenly objects must therefore be constructed just from circles. Epicycles allowed Copernicus to show how a system of circles can result in an oval planetary path.[15]

[14] Craig McConnell provides good animations that show how Ptolemy's system and Copernicus's account for retrograde motion: http://faculty.fullerton.edu/cmcconnell/Planets.html.

[15] Goldstein and Hon (2005, p. 88) say that "the path of a planet is not a significant concept" for Copernicus, though on the same page they say that "Copernicus admits [in *On Revolutions* V.4] that the path of a planet, due to the compounding factors of the eccentric and the little epicycle, is not a circle." Goldstein and Hon's main point is that Kepler was the first astronomer to hold that planets have trajectories through space without being embedded in physical "orbs." It is debatable whether Kepler

So what becomes of the boast that the Copernican system is simpler than Ptolemy's? Some have argued that Copernicus postulated fewer epicycles than Ptolemy, but this has been contested (Kuhn 1957, p. 171; Neugebauer 1957, pp. 204–205; Dijksterhuis 1961, pp. 293–294). The contrast I want to draw between Ptolemy's and Copernicus's systems does not involve counting epicycles; rather, the idea is that the Copernican model *predicts* regularities that the Ptolemaic model can only *accommodate.* This difference between the two models impressed Copernicus and his followers. In what follows I'll describe what impressed them, leaving until the next chapter the question of whether they ought to have been impressed.

Let's begin with the following regularity:

(B-O) Each superior planet is at its brightest when it is in opposition to the Sun.

The pre-Copernican term "superior" means that the planet is "above" the Sun. This is geocentric terminology, but Copernicus used it anyway (as I will) as a label for a certain set of planets (Mars, Jupiter, and Saturn), each of which takes more than one Earth-year to complete its journey across the sky. The term "opposition" means that we Earthbound observers see the planet rise at sunset and set at sunrise – the planet does the opposite of what the Sun does. The B-O generalization is an observation statement and is stated in a language that is neutral between the two theories.[16]

How does the Copernican system explain the B-O regularity? Let's begin with a planet's being superior. The theory says that a planet is superior exactly when it is farther from the Sun than the Earth is. The theory also entails that a planet is in opposition to the Sun if and only if the Earth is precisely between the two. To understand this point, consider Figure 1.3, which shows Earth and a superior planet both going counterclockwise around the Sun; the Earth is rotating on its axis, also in a counterclockwise direction. The Earth is at location E_1 when the planet is at P_1, and the Earth is at E_2 when the planet is at P_2. Notice that the Earth completes a quarter of its path round the Sun

was the first to think this (Shank 2003), but even if he was, the innovation does not undercut the idea that Copernicus's system involves planetary paths. Here I am using the term "path" so that it is compatible with planets sitting on orbs, but does not require this.

[16] The idea is that observations are neutral relative to the theories under test, not that they are absolutely theory-neutral; see Sober (2008a) for discussion.

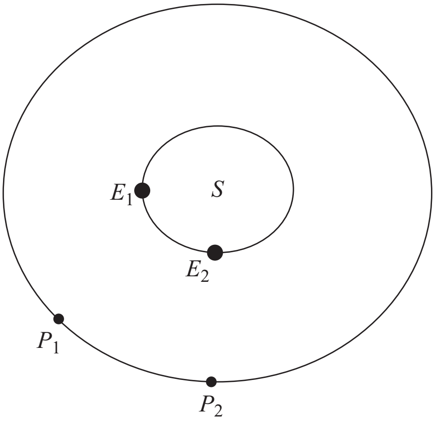


Figure 1.3

in the same amount of time that the planet completes only one eighth of its circuit. When the Earth and the planet are at E_2 and P_2 respectively, observers on Earth start to see the planet when they see the Sun set, and they cease to see it when they see the Sun rise. This isn't true when the Earth is at E_1 and the planet is at P_1; in that case, the planet becomes visible during the night and ceases to be visible well after sunrise. So a superior planet is in opposition to the Sun precisely when the Earth is between it and the Sun. A second consequence of the Copernican system is that the planet appears brightest to earthbound observers when the planet is in opposition to the Sun. This is because the planet is closer to the Earth when the locations are E_2 and P_2 than when the locations are E_1 and P_1. The B-O generalization falls handily into place in the Copernican model.

Ptolemy's geocentric model, in contrast, does not predict the B-O regularity. The heliocentric and the geocentric systems agree that a planet is in opposition to the Sun when the earth is exactly in between them. Figure 1.4 shows three possible positions that planet P might occupy, each consistent with P's being in opposition to the Sun. Notice that opposition leaves open whether P is brightest, dimmest, or somewhere in between.[17] True, the geocentric model can accommodate the regularity by choosing one of these epicycles

[17] Here and in later figures, I draw the Sun without an epicycle. This is Ptolemy's preference in *The Almagest* (3.4). He says that putting the Sun on an eccentric rather than on an epicycle is "simpler and is performed by means of one motion instead of two" (Ptolemy 150, p. 153).

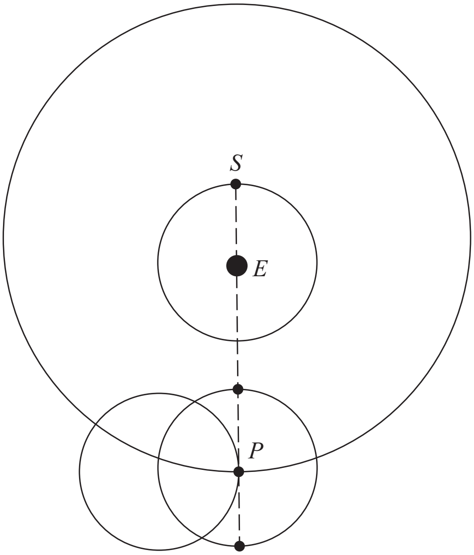


Figure 1.4

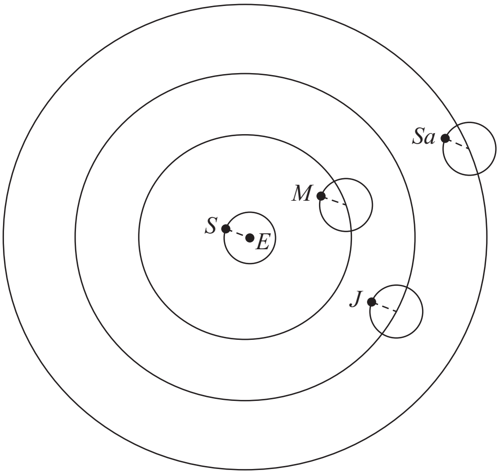


Figure 1.5

rather than the others. But this just means that the geocentric model says that the B-O regularity is *possible*; there is no entailment that it must be so. Matters get worse for the Ptolemaic system when we consider all the superior planets, not just one of them. Figure 1.5 depicts a geocentric model with the three superior planets known to Ptolemy and Copernicus – Mars, Jupiter, and Saturn. For the Ptolemaic model to capture the B-O regularity, the arrow that points from the center of a planet's epicycle to the planet must remain parallel with the arrow that points from Earth to Sun. This parallelism of the four arrows must be maintained as the three superior planets move along their epicycles while those epicycles and the Sun all circle the Earth. According to the Ptolemaic model, this coordination is a mere coincidence.

The astronomical objects depicted in Figure 1.5 resemble four clocks lying face-up on a table with one of them at the center and the other three revolving around it. Each clock has a single hand and the four hands tick through the hours in perfect synchrony. That the clock hands move in parallel goes beyond the fact that one clock is at the center with the other three revolving around it. It is in this sense that the Ptolemaic model regards the B-O regularity as a coincidence that the model can accommodate but does not predict. For Copernicus's heliocentric model, the B-O regularity is no coincidence at all.

The distinction between prediction and accommodation also applies when each model is asked to explain a second regularity:

(I-N) The inferior planets (Mercury and Venus) are always observed to be near each other and near the Sun.

The Ptolemaic hypothesis – that the Sun, Mercury, and Venus each orbit a stationary Earth – treats these orbits as independent patterns. To explain I-N, Ptolemy is forced to add a postulate, that the line from Earth (or more precisely, the equant) to the (mean) Sun happens to coincide with the line from Earth to the center of Mercury's epicycle and the line from Earth to the center of Venus's epicycle. This lining-up arrangement is depicted in Figure 1.6(a). Figure 1.6(b) depicts the heliocentric model; it automatically ensures that the I-N regularity holds. According to the Copernican model, Mercury and Venus stay closer to the Sun than Jupiter does simply because the Sun is at the center and the order of the orbits is Mercury, Venus, Earth, and Jupiter. In each diagram in Figure 1.6, draw an angle whose vertex is at the Earth and that is just wide enough to cover the widest separation that Mercury and Venus can have. This angle automatically includes the Sun in the heliocentric model, but does so in the geocentric model only because of the line-up stipulation depicted in Figure 1.6(a). The basic geocentric model

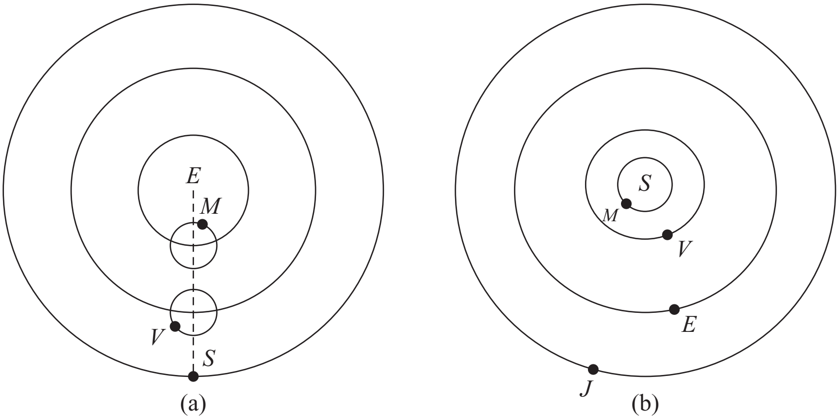


Figure 1.6

is forced to accommodate I-N by adding a postulate, whereas the heliocentric model explains I-N without needing any such supplementation.

In his *Cosmographic Mystery* of 1596, Johannes Kepler (1571–1630) looks back at Copernicus's theory and says that his confidence in it was first established by its "magnificent agreement with everything that is observed in the heavens." Kepler then adds:

> However, what is far more important . . . for the things at which from others we learn to wonder, [is that] only Copernicus magnificently gives the explanation, and removes the cause of wonder . . . Reasons are supplied for a great many . . . matters for which Ptolemy for all his many motions could give no reason. (Kepler 1596, pp. 75f.; quoted in Lange 1995, pp. 506–507)

Kepler's contrast between explanation and its absence resembles the contrast I have drawn between prediction and accommodation. But two questions remain: why does the fact that one model predicts what another can only accommodate count as evidence in favor of the former? And what does this difference have to do with Ockham's razor? These questions will be addressed in Chapter 2. For the moment, the point is that Copernicus and his followers thought that these dots are all connected.

To better grasp the distinction I am drawing between prediction and accommodation, consider a different inference problem that is much more mundane. In a single week last summer in Madison, my friend Susan told me that she went to Lake Mendota each day and each day she saw a red sailboat. Consider two possible explanations of these reports:

(ONE) There was a single sailboat that Susan saw seven times.
(SEVEN) Susan saw seven different sailboats, one each day.

ONE predicts that Susan's reports will agree with each other about the color of what she saw. In contrast, the SEVEN hypothesis can accommodate the agreement among the reports, but it predicts no such thing. Notice that this difference between the two hypotheses disappears if we flesh them out as follows:

(ONE+) There was a single red sailboat that Susan saw seven times.
(SEVEN+) Susan saw seven different red sailboats, one each day.

Both of these beefed-up hypotheses predict that Susan's seven reports will agree with each other about the color of what she saw. The non-plused hypotheses are barebones; the plus hypotheses are obtained by fleshing out the barebones. When I say that Copernicus predicts what Ptolemy only accommodates, I am talking about their barebones models, not what happens when they are fleshed out. This simple example also will be analyzed in the next chapter.

A final word on epicycles: it was Kepler, not Copernicus, who proposed a heliocentric model that did without epicycles. Kepler's decisive step was to abandon his predecessor's fixation on circles. Kepler's orbits were ellipses straight up, not ellipses constructed by piling circles upon circles. A consequence of this innovation was that a piece of astronomical vocabulary became a metaphor for a larger idea. When people now say that a postulate is an epicycle, they are making a criticism. They usually mean that the postulate is unsupported by evidence and that it was introduced for the sole purpose of keeping a defective theory afloat. Epicycles in this modern sense are what Ockham's razor is supposed to slice away.

Descartes and Leibniz on God and the laws of nature

René Descartes (1596–1650) and Gottfried Wilhelm von Leibniz (1646–1716) thought that there is a theological reason why the laws of nature must be simple, but they came to that conclusion by different routes.

These two philosophers shared a second conviction – that all material events have mechanical explanations. Material objects are like machines, and what happens to them can be explained by the to-ing and fro-ing of tiny particles – both those that comprise the machines and those that impinge on them from the external environment.[18] In Descartes's book *The World*, which he withdrew from publication when he learned that the Holy Office (the Inquisition) had just condemned Galileo's heliocentrism, Descartes expresses this reductionist commitment as follows:

> If you find it strange that . . . I do not use the qualities called "heat," "cold," "moistness," and "dryness," as do the [scholastic] philosophers, I shall say to you that these qualities appear to me to be themselves in need of explanation. Indeed, unless I am mistaken, not only these four qualities, but also all the others (indeed all the forms of inanimate bodies) can be explained without the need of supposing for that purpose any other thing in their matter than the motion, size, shape, and arrangement of its parts. (Descartes 1633, Chapter 5)

In the same spirit, Leibniz maintains that "all natural phenomena could be explained mechanically if we understood them well enough" (Leibniz 1969, p. 478).

Although both thinkers hold that material events have mechanistic explanations, they also hold that the physical laws that do the explaining cannot themselves be explained mechanistically. It is here that theology comes to the rescue. However, Descartes and Leibniz brought God into the picture by different routes. Descartes thought that the laws of physics can be derived from God's *nature*, whereas Leibniz did the derivation from premises concerning God's *goals*. This may sound like a small difference. Why does it matter whether one starts with the assumption that God is unchangeable rather than with the assumption that God wants to make the world a perfect place? Regardless of the answer to this question, Descartes had nothing but disdain for the project that Leibniz would later pursue. In the fourth of his *Meditations on First Philosophy*, Descartes says that God "is capable of countless things

[18] Descartes was not a mechanist about everything; he embraced the dualist doctrine that the mind is immaterial. Conscious experience has no spatial location and so it does not occur "in nature" if nature is limited to the totality of events that have spatio-temporal location. It is interesting that some of what we now think of as mental phenomena were, for Descartes, part of the world of matter (Descartes 1662, p. 169).

whose causes are beyond my knowledge. And for this reason alone I consider the customary search for final causes to be totally useless in physics; there is considerable rashness in thinking myself capable of investigating the purposes of God" (Descartes 1641). Three years later he repeats the prohibition in the *Principles of Philosophy* (1644, I: 28): "When dealing with natural things we will . . . never derive any explanations from the purposes which God or nature may have had in view when creating them. For we should not be so arrogant as to suppose that we can share in God's plans."

Descartes's derivations

In *The World*, Descartes (1633) asks his readers to imagine a world that consists solely of mechanical goings-on – there is matter in motion, but no creatures with conscious experience. God brings this purely material world into existence and then imposes upon it a set of physical laws that will govern the motions of objects. The first two of the three laws that Descartes describes are these:[19]

> Law A: Each part of matter, taken by itself, always continues to be in the same state until collision with others forces it to change . . . And so once it has begun to move, it will continue always with the same force, until others stop it or slow it down.
>
> Law B: When a body pushes another, it cannot give it any motion without at the same time losing as much of its own, nor can it take any of the other's away except if its motion is increased by just as much.

Descartes says that "these two rules follow in an obvious way from this alone, that God is immutable, and acting always in the same way, he always produces the same effect." From this fact about God, Descartes also derives a conservation law:

> Thus, assuming that he placed a certain quantity of motions in the totality of matter from the first instant that he had created it, we must admit that he always conserves in it just as much, or we would not believe that he always acts in the same way.

[19] In this section, I use the translations of Descartes from Garber (1992, pp. 198–199, p. 281, and p. 286) and I am indebted to Garber's analysis of Descartes's arguments.

In the *Principles of Philosophy* (1644, II 36), Descartes argues along the same lines for his conservation principle, but he introduces a new element, an "exception":

> We also understand that there is perfection in God not only because he is in himself immutable, but also because he works in the most constant and immutable way. Therefore, with the exception of those changes which evident experience or divine revelation render certain, and which we perceive or believe happen without any change in the creator, we should suppose no other changes in his works, so as not to argue for an inconstancy in him.

In these passages, Descartes does not use the word "simplicity," but one consequence of what he says is that the universe has a kind of simplicity. It is simpler to have laws that are fixed for all time than to have the laws of nature perpetually changing. We know that our universe is governed by unchanging laws because we know that God, an immutable being, made it so.

Scholars have disagreed about the exact relationship that Descartes sees between *a priori* propositions about God and propositions that characterize the laws of nature (Nadler 1990). One context in which this question can be raised concerns two of the passages just quoted. In *The World* Descartes says that the laws "follow" from God's immutability, whereas in *Principles* he says that it is God's immutability, plus additional premises about the changes we observe and the changes that the Bible describes, that constrain what we should take the laws to be. Perhaps Descartes uses the concept of "following" in the first passage to indicate a relation that is less exacting than the relation of logical consequence. Or perhaps his usage is close to the modern meaning, in which case we should conclude that Descartes changed his mind – he backed off from the stronger relation postulated in *The World* to the weaker relation described in *Principles*. In any event, the second formulation is interesting from the point of view of thinking about Ockham's razor. God's immutability apparently tells us that we should postulate no more changes than we know about from observation and from sacred texts. This is a principle of parsimony.

In *Principles of Philosophy* (II: 39), Descartes states his second law of nature and asks why it is true. The law says that "all motion is in itself rectilinear," where "motion in itself" means the motion that will occur if no outside causes impinge. Descartes says that this law of straight-line motion owes its truth to

> the immutability and simplicity of the operation by which God preserves motion in matter. He always preserves the motion in the precise form in which it is occurring at the very moment when he preserves it, without taking account of how it was moving [a moment before].

A fact about the simplicity of God's interaction with matter explains why motion exhibits what is now called the Markov property (a topic we will revisit in the next two chapters). What happens next depends just on the system's present state; its prior state is "forgotten." This is parsimonious, since what happens next depends on less rather than on more.

Leibniz on the best of all possible worlds[20]

People who aren't professional philosophers, if they have heard of Leibniz at all, are apt to have heard of him because he was the butt of a joke. Leibniz held that God creates the world we see around us by choosing from a set of possible worlds; he chose to actualize this world, rather than any of the alternatives, because this world of ours is the best of all possible worlds. Voltaire (1694–1778) satirized Leibniz in his 1759 novella *Candide*, not by name, but in the form of Dr. Pangloss, who breezily insisted that even the most egregious of evils has a silver lining that more than compensates for the harm the evil does.[21] Voltaire was ridiculing Leibniz's solution to the problem of evil: how can an all-knowing, all-powerful, and entirely benevolent deity permit there to be so much evil in the world? Why do bad things happen to good people, not just occasionally, but massively?

When Leibniz maintains that our world is "best," he has moral goodness in mind, but he means something more. He says in Section 6 of his 1686 *Discourse on Metaphysics* that "God has chosen the most perfect world, that is to say, the one that is at the same time the simplest in hypotheses and the richest in phenomena." This statement entails that there is no conflict between simple laws and diverse phenomena – both can be maximized and that is exactly

[20] I am grateful to David Blumenfeld for helping me improve my understanding of Leibniz.

[21] Palmer (2002) argues that Voltaire, though well aware of Leibniz's optimism, found his main model for Pangloss in the popular writer Noel Antoine Pluche (1688–1761). My thanks to Trevor Pearce for pointing this out to me.

what God does.[22] Indeed, Leibniz thought that maximizing both is not just possible, but that having simple laws is both necessary and sufficient for God to maximize diversity of phenomena (Leibniz 1969, p. 648; Blumenfeld 1995).[23]

Why does Leibniz think that the simplest set of laws will produce more diverse phenomena than laws that are more complex? Sometimes he defends this idea by means of analogies:

> God makes the most things he can and what obliges him to seek simple laws is the need to find a place for as many things as can be put together; if he made use of other laws, it would be like trying to make a building with round stones, which makes us lose more space than they occupy. (Leibniz, 1969, p. 211; quoted in Blumenfeld 1995, p. 389)

In another passage Leibniz compares God's creating the universe to

> certain games in which all the spaces on a board are to be filled according to definite rules, but unless we use a certain device, we find ourselves at the end blocked from the difficult spaces and compelled to leave more spaces vacant than we need to or wished to. Yet there is a definite rule by which a maximum number of spaces can be filled in the easiest way. (Leibniz 1969, p. 487; quoted in Blumenfeld 1995, p. 391)

In *Theodicy* (1710, Section 208), Leibniz states the general principle that underlies these two examples: God chooses simple laws because "the more intricate processes take up too much ground, too much space, too much place, too much time that might have been better employed." Leibniz's explanation is puzzling even if we accept the assumption that simpler processes consume less of some resource. It is simpler to have all water boil at 100 degrees Celsius

[22] Rescher (1982, p. 11) disagrees with this interpretation; he says that Leibniz sees a conflict between simplicity of laws and diversity of phenomena, which God resolves by finding the optimal balance.

[23] Whereas Leibniz saw no conflict between simple laws and moral perfection, Leibniz's interlocutor Nicholas Malebranche (1638–1694) saw things differently. In his 1680 *Treatise on Nature and Grace*, Malebranche says that the universe is morally imperfect. Malebranche's God had to choose and did so by giving absolute priority to ensuring that the universe has simple laws (Nadler 2008).

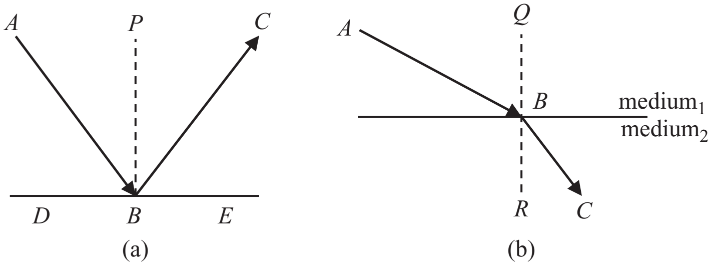


Figure 1.7

than to have water in different times and places boil at different temperatures. In this example and in many others, simpler laws deliver monotony, not diversity.

Leibniz's idea is not just that we can explain why the laws of nature are simple by appealing to God's goals. He thinks, in addition, that we can use ideas about simplicity and God's goals to help us discover what those laws are. Leibniz develops this theme in his discussion of optics in his 1696 *Tentamem Anagogicum*. Consider the law of reflection (Figure 1.7[a]). A light beam starting at point *A* bounces off a flat mirror and ends up at point *C*. The light gets from point *A* to point *C* by hitting the mirror at point *B*; in doing so, the angle of incidence (the angle *ABP*) is equal to the angle of reflection (the angle *PBC*). By going from *A* to *B* to *C*, the light travels a shorter distance than it would do if it went from *A* to *D* to *C* or from *A* to *E* to *C*; the shortest path from *A* to the mirror and then to *C* passes through *B*, a result obtained by Hero of Alexandria (10–70 CE). In the law of refraction (Figure 1.7 [b]), the light ray moves from point *A*, which is in one medium, to point *C*, which is in another. For example, the two media might be air and water, or air and glass. In this case, the light bends; it does not follow the shortest path from *A* to *C*.[24]

In the *Tentamem Anagogicum*, Leibniz (1969, pp. 477–486) makes his case for teleology by describing how things go wrong when one views the problem

[24] The law of refraction furnishes another instance of Stigler's Law of Eponymy. The optical law is now often called Snell's Law. Snell died in 1626 without publishing his discovery; Thomas Harriot had the law by 1602, though he too did not publish. Descartes was the first European to publish it, in his 1637 *Discourse on Method*. Some scholars think that all these figures were Johnny-come-latelies, since the real discoverer was the tenth century Ibn Sahl in Baghdad, who wrote "On Burning Mirrors and Lenses." See Young (2012).

purely mechanistically. The wrong turn he considers is Descartes's explanation of the laws of reflection and refraction in his 1637 *Optics*. Descartes does not talk about the goal of getting the light from *A* to *C*. Rather, Descartes describes what happens to the light when it reaches *B*, having departed from *A*, and proposes an explanation, based solely on that history, for why the light then assumes a trajectory that takes it to *C*. In explaining the law of reflection, Descartes asks the reader to consider a tennis ball bouncing off a hard surface. It approaches that surface at an angle and its motion decomposes into two components, which Descartes calls "determinations" – one that goes straight down and the other that goes horizontally. Descartes asserts that the collision with the ground affects the vertical component while the horizontal component remains the same. Assuming finally that the ball's total speed is the same before and after the bounce, he deduces the law of reflection. The mechanistic derivation of the law of refraction is similar. Instead of a ball's bouncing off the ground, the ball hits a thin linen sheet at an angle and tears through it. Again there is a vertical and a horizontal determination; the impact with the sheet affects the former, but the horizontal component "must always remain the same as it was, because the sheet offers no opposition at all to the determination in this direction." Descartes concludes that the ball loses some of its downward motion but none of its horizontal motion, so the trajectory bends. His analysis does not allow him to say how much bending will occur; he fails to recover the precise formulae of the sine law (Garber 1992, pp. 188–193; McDonough forthcoming).

Leibniz complains that Descartes's explanations are "extremely forced and not intelligible enough." Though Leibniz doubts the specifics of Descartes's proposals, he agrees that mechanistic explanations must exist for these optical phenomena. Leibniz nonetheless insists that the mechanistic approach is deficient, in two respects: it fails to explain why the laws of mechanics are as they are and it is a poor guide to discovering new laws. Both these deficiencies are remedied by appealing to God's goal that the universe be governed by laws of maximal perfection.

When Leibniz says that the perspective of final causes facilitates the discovery of laws, he is thinking of Fermat's use of the principle of least time to explain both reflection and refraction. In both cases, the path followed is the "easiest," where a path from point *A* to point *C* is easiest if it minimizes the product of distance and resistance. The light gets from *A* to *C* in the least time by following the path of least resistance. In the case of reflection off a mirror,

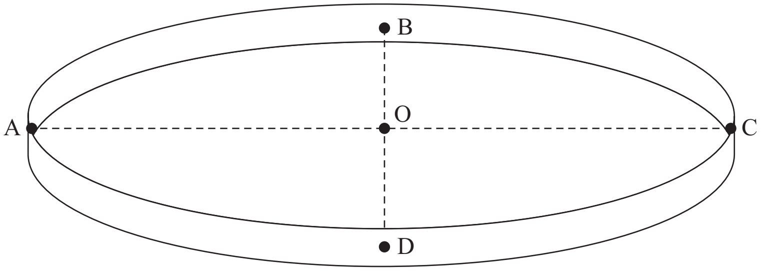


Figure 1.8

there is just one medium that the light moves through, so the easiest path is just the shortest distance from *A* to a point on the mirror and then to *C*. In the case of refraction, there are two media and the light does not take the shortest path if the media offer different degrees of resistance. Rather, the easiest path will have the light spending more time in the medium of lesser resistance. Think of the path a life guard on a beach will take to reach a swimmer in distress as quickly as possible. Whereas Descartes seeks to explain why the light will go to *C* (rather than to some other point), given that it already travelled from *A* to *B*, Leibniz seeks to explain why the light passes through *B* (rather than through some other point), given that it goes from *A* to *C*.

Although "the principle of the easiest path" (also known as the principle of least time) seems to work well for the simple examples depicted in Figure 1.7, Leibniz denies that it is fully satisfactory. Light sometimes takes the *longest* time. Consider the mirrored wall shown in Figure 1.8; the wall forms an oval enclosure around a light source that is at the central point *O*. The figure shows four paths that light takes from *O* to the wall and back again; two of them minimize travel time, but two maximize it. The light takes all four paths, not just the two that are easiest. This leads Leibniz to propose a principle that "supersedes" the principle of the easiest path. It says that "in the absence of a minimum it is necessary to hold to the *most determined*, which can be the *simplest* even when it is a *maximum*." The simple case of reflection off a flat mirror shown in Figure 1.7(a) illustrates this "principle of the most determined path" (McDonough 2009). The thing to notice about point *B* in Figure 1.8 is that it is unique; it has no "twin." The same cannot be said of point *D*; the path from *A* to *D* to *C* has the same length as the path from *A* to *E* to *C*. A similar point applies to refraction (Figure 1.7[b]). There is a unique path from *A* to *C* that minimizes travel time. Any path that takes longer has a twin.

Returning to the oval wall in Figure 1.8, we can see that Leibniz's idea of uniqueness needs to be understood *locally*. There are four paths from *O* to the mirror and back again. Each is *locally* unique, but not globally. Consider point *C*. Each point around *C* has a twin, but *C* does not. *A* is not a twin of *C* because they aren't in the same locality. Indeed, the need to construe uniqueness locally is visible even in the simplest case – reflection off a flat mirror. In Figure 1.7(a), the path from *A* to *B* to *C* is the shortest of the paths that hit the mirror, but there is, in addition, a direct path from *A* to *C* that bypasses the mirror entirely. In reality, there are *two* paths from *A* to *C* that are *locally* unique.

Leibniz summarizes the approach he favors, and its results, in Section 22 of his 1686 *Discourse on Metaphysics*:

> The way of final causes . . . is often useful for understanding important and useful truths, which one would be a long time seeking by the other more physical route; of this fact anatomy can provide significant examples. I believe, too, that Snell, who first discovered the rules of refraction, would have waited a long time to find them if he had sought first to discover how light is formed. But apparently he followed the method which the ancients used in catoptrics, which is in fact that of final causes. For seeking the easiest way in which to direct a ray from one given point to another through reflection by a given plane (assuming that the easiest way is the plan of nature), they discovered the equality of the angles of incidence and reflection . . . This method, I believe, Snell, and later independently of him Fermat, applied most ingeniously to refraction. For when rays in the same media observe a ratio between the sines which is equal to the ratio of the resistances of the media, this happens to be the easiest, or at least the most determined way to pass from a given point in one medium to a given point in another.

Notice Leibniz's mention of anatomy.

How do Leibniz's views about the most determined path connect with his idea that God wants simple laws? In the passage just quoted, Leibniz seems to allow that the easiest path might not be the one that is most determined. Has simplicity (easiness) fallen away as a crucial consideration? This is not plausible, given Leibniz's repeated insistence that God chooses the simplest laws and does so by finding laws that are the most determined. If simplicity means parsimony and parsimony means minimizing some quantity, how can the principle of the most determined path be an instance of Ockham's razor? Leibniz is clear that he does not care whether a physical quantity is minimized or maximized as long as it is unique.[25] As far as I know, Leibniz does not explicitly address the question of why the most determined path is simplest, but an answer he might have endorsed can be obtained by counting *solutions* to a problem. When a candidate law has a twin, God has a choice. He can actualize both the candidate in question and its twin, so there will be two paths from *A* to *C* that reflect off the mirror in Figure 1.7(a), not just one. But now the flood gates open. If God embraces both the *ADC* and the *AEC* paths, what about all the others? Surely this multiplicity of pathways is anything but simple. The alternative is for God to be arbitrary; God could choose *ADC* and reject *AEC* even though he has no reason to do so. This is ruled out by the *principle of sufficient reason*, which is central to Leibniz's philosophy; God does *nothing* without having a reason. A law that is most determined is maximally simple because, without arbitrariness, it embodies a *unique* solution.

Leibniz's strategy of solving problems by seeking solutions that uniquely maximize or minimize bore tremendous fruit. In 1746, Maupertuis introduced his *law of least action*, an idea that Euler was adumbrating around the same time. It began as a result about the motion of point masses, but then it grew:

> The laws of movement and of rest deduced from this principle being precisely the same as those observed in nature, we can admire the application of it to all phenomena. The movement of animals, the vegetative growth of plants . . . are only its consequences; and the spectacle of the universe becomes so much the grander, so much more beautiful, the worthier of its Author, when one knows that a small number of laws, most wisely established, suffice for all movements. (Maupertuis 1746, p. 286)

For Maupertuis, "nature is thrifty in all its actions."[26] The idea was elaborated and refined by Lagrange, Hamilton, Jacobi, and Weierstrass. The resulting "variational principles" found many applications in classical physics and similar techniques are used today in relativity theory and quantum mechanics.

[25] Leibniz may have been moved by the fact that minimizing a magnitude M is the same as maximizing $-M$.

[26] Voltaire took aim at Maupertuis in his 1753 satire *Histoire du Docteur Akakia et du Natif de Saint Malo*.

The teleological interpretation remains controversial even when the mathematical derivation is sound. In a simple example like the law of refraction, it is true that the path followed is the one of least time; it is another matter to say that the light follows this path *because* it minimizes travel time. But even without the teleological add-on, the usefulness of minimal principles in science has suggested to many that nature is following the dictates of Ockham's razor.

Newton on avoiding the luxury of superfluous causes

In the second edition (1713) of his *Mathematical Principles of Natural Philosophy* (the first edition appeared in 1687), Isaac Newton (1643–1727) states four *Rules of Reasoning in Philosophy*. By "natural philosophy," Newton meant something close to what we now mean by "science." He wasn't talking about philosophy in the modern sense of a subject that (for better or worse) is often separate from science. Here are the first three rules; I'll discuss the fourth in Chapter 4:

> *Rule I. No more causes of natural things should be admitted than are both true and sufficient to explain their phenomena.* As the philosophers say: Nature does nothing in vain, and more causes are in vain when fewer suffice. For nature is simple and does not indulge in the luxury of superfluous causes.
>
> *Rule II. Therefore, the causes assigned to natural effects of the same kind must be, so far as possible, the same*. Examples are the cause of respiration in man and beast, or of the falling of stones in Europe and America, or of the light of a kitchen fire and the sun, or of the reflection of light on our Earth and the planets.
>
> *Rule III. Those qualities of bodies that cannot be intended and remitted [i.e., qualities that cannot be increased and diminished] and that belong to all bodies on which experiments can be made should be taken as qualities of all bodies universally.* For the qualities of bodies can be known only through experiments; and therefore qualities that square with experiments universally are to be regarded as universal qualities; and qualities that cannot be diminished cannot be taken away from bodies. Certainly idle fancies ought not to be fabricated recklessly against the evidence of experiments, nor should we depart from the analogy of nature, since nature is always simple and ever consonant with itself. (Newton 1687, pp. 794–795, brackets in the original)

The voice of Aristotle can be heard in Rule I, but only in echo. Newton's rule involves no commitment to teleology. Whereas Aristotle was mainly thinking about biological examples when he said that nature does nothing in vain, Newton was casting his net more widely. He was stating a fully general principle about the postulation of causes. And whereas Aristotle sees his principle as a summary of what empirical inquiry has yielded, Newton sees his rule as guiding that inquiry.[27]

Rule II is about unification. Similar effects have similar causes. Newton's example of the descent of a stone in Europe and America is close to an example that made him famous; Newton explains the motion of the moon by appeal to the same laws that he uses to explain the motion of a falling apple here on Earth. This contrasts with Aristotelian physics, which treats terrestrial and celestial motion differently. For Aristotle, the natural motion of an apple is to move in a straight line towards the center of the Earth, whereas the natural motion of the moon is to move in a circle. For Newton, both move with constant velocity unless acted upon by an external force. Did Newton see the irony of favorably alluding to Aristotle in Rule I even though Aristotle violated the principle espoused in Rule II?

Newton's Rule III is about induction. It says that we are entitled to generalize from our finite and very limited experience. Newton proposed a *universal* law of gravitation, one that he thought applies to *all* places and *all* times, even though the evidence for this law comes from a very small sample of places and times. Newton returns to the theme of parsimony and simplicity in his discussion of this rule when he says "nor are we to recede from the analogy of Nature, which uses to be simple, and always consonant to itself." This "analogy" allows us to transfer what we find to be true of the objects we observe to those that we fail to observe, where this failure is due to their being too distant from us in space or time, or to their being too small. Here again, Newton may be thinking critically of Aristotelian physics.

Newton is saying in all three of these rules that a principle of parsimony or simplicity (I'll use these terms more or less interchangeably) is a reliable guide to discovering true theories. It isn't just that simple theories are beautiful and easy to understand. When one theory is simpler than another, that is supposed to be a reason to think that the simpler theory is true and the more complex theory is false. Newton is not saying that simplicity is the *only* consideration.

[27] The phrase in Rule 1 "more causes are in vain when fewer suffice" translates Newton's Latin ("frustra fit per plura quod fieri potest per pauciora"), which is right out of Duns Scotus and Ockham.

Theories also must conform to what we observe. A simple theory that runs afoul of the observations is *too* simple.

In these passages from the *Rules of Reasoning*, Newton offers a brief defense of his taste for simplicity, but it doesn't go very deep. He says that "nature does nothing in vain," but he does not explain why one should agree that this is so. He cites "the philosophers," but Newton did not believe that a proposition is true just because some philosophers say that it is. Newton's statement and discussion of his rules conform to the dictates of methodological naturalism (in the sense that God goes unmentioned in them), but that is not because he thought that God is irrelevant to the question of why the methodological advice is sound. In an unpublished *Treatise on Revelation* (*c.* 1670–1680), Newton enunciates some "Rules for methodizing/construing the Apocalypse."[28] Rule 9 concerns the use of simplicity in Biblical interpretation:

> *Rule 9. To choose those constructions which without straining reduce things to the greatest simplicity.* The reason of this is manifest by the precedent Rule. Truth is ever to be found in simplicity, & not in the multiplicity & confusion of things. As the world, which to the naked eye exhibits the greatest variety of objects, appears very simple in its internal constitution when surveyed by a philosophic understanding, & so much the simpler by how much the better it is understood, so it is in these visions. It is the perfection of God's works that they are all done with the greatest simplicity. He is the God of order & not of confusion. And therefore as they that would understand the frame of the world must endeavor to reduce their knowledge to all possible simplicity, so it must be in seeking to understand these visions.

Newton's "constructions" are the interpretations we should place on Biblical passages. The preceding rule to which he refers is this:

> *Rule 8. To choose those constructions which without straining reduce contemporary visions to the greatest harmony of their parts.*

This reference to "harmony of parts" resembles what Newton says about "the analogy of nature" and nature's "always [being] consonant to itself" in his third rule of reasoning. But even more striking is an idea that does not surface in the *Rules of Reasoning* that Newton added to his *Mathematical Principles of*

[28] The Yahuda Ms. 1.1 in the National Library of Israel, Jerusalem, Israel: http://www.newtonproject.sussex.ac.uk/view/texts/normalized/THEM00135.

Natural Philosophy. The principle of parsimony is correct because the universe is made by God. The perfection of God's law-making is reflected in the simplicity of his laws. God is said to be "the God of order and not of confusion," a phrase reminiscent of Descartes's insistence in his 1641 *Meditations on First Philosophy* that God is no deceiver. God makes the world simple so that we humans can grasp its laws and see from those laws that the world was made by a benevolent deity.

It would be a mistake to think that Newton took this discussion of God to pertain only to the interpretation of religious texts. For Newton, rules of reasoning are *general*; they apply to reasoning about any subject matter at all. This idea has an ancient philosophical pedigree. From Aristotle to the present, logicians have held that the correctness of a rule of reasoning does not depend on the subject matter to which it is applied. For example, the following two arguments are deductively valid, meaning that if the premises are true, then the conclusion must be true:

Socrates is a human being.	The Parthenon is made of stone.
All human beings are mortal.	All things that are made of stone are hard.
————————	————————
Socrates is mortal.	The Parthenon is hard.

Not only are both arguments valid; they are valid for the same reason. The arguments have the same *logical form*. Logical validity has nothing to do with the subject matters of arguments. So it was with Newton's view of parsimony. For him, the principle of parsimony applies to Biblical revelations just as it applies to planetary revolutions, and it has a unitary justification.

Aristotle's "nature does nothing in vain" was an expression of his teleology. The natures of individual tigers and rocks cause them to behave as they do and their goal-directedness stems from those individual natures. Aristotle was not invoking a divine intellect when he proposed this teleological picture (Lennox 2001). Newton and Leibniz retained the Aristotelian phrase, but they changed the subject – from particular rocks and tigers to laws and principles. And the source of the teleology was shifted as well – from the natures of individual tigers and rocks to God himself. For Leibniz, the simple laws that light obeys do not stem from light's having a goal, but from God's having the goal of creating a world of maximal perfection (McDonough 2009). Newton was on the same page; the laws of nature are simple because God is perfect and so is his handiwork. He is the God of order, not of confusion. Aristotle's resonant phrase was not discarded; rather, it was assigned a new meaning.

Hume on the principle of the uniformity of nature

What would it be like to evaluate the principle of parsimony without bringing in God? One option is to argue that parsimony's justification comes from *a priori* mathematics and logic. Another is to argue that its justification comes from empirical facts about nature. A third is to contend that the principle has no justification even though it is something that we need to assume in constructing our picture of the world. This last option is what one finds in the work of David Hume (1711–1776).

In his third rule of reasoning, Newton talks about the "analogy of nature." Hume talks about something similar – *the principle of the uniformity of nature*. This is the idea that the future will resemble the past. This principle says that our universe possesses a kind of simplicity; regularities that exist in the past will persist into the future. Just as Newton says that celestial and terrestrial motion obey the same laws, Hume's principle says that past and future do the same. What Newton said about space, Hume's principle asserts about time. Hume maintains that the principle of the uniformity of nature is a presupposition of all inductive inferences. Whereas Newton tried to justify a principle of simplicity by invoking God, Hume says that the principle of the uniformity of nature cannot be justified, either theologically or in any other way.

Why was Hume unwilling to invoke the goals of a benevolent creator to justify the principle of the uniformity of nature? Although his 1779 *Dialogues Concerning Natural Religion* might suggest that this unwillingness stemmed from a deep suspicion of the design argument for the existence of God, the situation is more complicated. In this book, Cleanthes is the defender of the design argument and Philo is usually its skeptical critic. I say "usually" because, in Part 12, Hume has Philo provide an unambiguous endorsement of the argument:

> A purpose, an intention, a design, strikes everywhere the most careless, the most stupid thinker; and no man can be so hardened in absurd systems, as at all times to reject it. That Nature does nothing in vain, is a maxim established

> in all the schools, merely from the contemplation of the works of Nature, without any religious purpose; and, from a firm conviction of its truth, an anatomist, who had observed a new organ or canal, would never be satisfied till he had also discovered its use and intention. One great foundation of the Copernican system is the maxim, That Nature acts by the simplest methods, and chooses the most proper means to any end; and astronomers often, without thinking of it, lay this strong foundation of piety and religion. The same thing is observable in other parts of philosophy: And thus all the sciences almost lead us insensibly to acknowledge a first intelligent Author; and their authority is often so much the greater, as they do not directly profess that intention.[29]

At the end of the *Dialogues*, Philo changes his tune; he says that the only thing that can be concluded about the existence of an intelligent designer is that "the cause or causes of order in the universe probably bear some remote analogy to human intelligence." Hume puts these words in italics. A more tepid endorsement of the design argument for the existence of God is hard to imagine, but, still, it *is* an endorsement. At that point in the *Dialogues*, the narrator draws the book to a close by saying that Cleanthes, the defender of natural theology, is closer to the truth than Philo. So where did Hume stand? Was the narrator speaking for Hume in this instance, or was the narrator a rhetorical device behind whom Hume was hiding?

Apart from whatever degree of skepticism Hume actually felt concerning explanations that appeal to God's goals, we do know that Hume had a more fundamental reason for denying that the principle of the uniformity of nature can be justified by appeal to an intelligent designer. It is here that Hume's distinction between *relations of ideas* and *matters of fact* comes into play. The former are comprised of statements like "triangles have three sides"; reason alone tells us that such propositions are true (they are *a priori*). The latter category is comprised of propositions that can be known only via observation (they are *a posteriori*). The principle of the uniformity of nature is not like the statement about triangles; you'd be contradicting yourself if you said that some triangles fail to have three sides, but there is no contradiction in saying that the future will fail to resemble the past. And since, for Hume,

[29] It is interesting that Hume here links the design argument to both biology and astronomy.

every proposition must either state a relation of ideas or a matter of fact, the principle of the uniformity of nature must state a matter of fact. If a statement that describes a matter of fact has a justification, that justification must come from our past observations. In this light, let's consider the following argument:

> Nature has been uniform in the past.
>
> ————————
>
> The future will resemble the past.

The premise means that what happened in 2001 resembled what happened earlier, that the same is true of what happened in 2002, in 2003, and so on. Now we stand in the present year. Should we expect that what happens next year will resemble what happened earlier? Does our prior experience justify this expectation? In the above argument, you obviously can't deduce the conclusion from the premise, but you might think that the premise renders the conclusion highly probable. Hume's view is that this is an inductive inference, and so the rule of inference you are using to connect premise to conclusion involves the assumption that nature is uniform. For Hume, any attempt to justify the principle of the uniformity of nature by an inductive inference from past experience would be question-begging. He puts the point like this:

> If there be any suspicion that the course of nature may change and that the past may be no rule for the future, all experience becomes useless and can give rise to no inference or conclusion. It is impossible, therefore, that any arguments from experience can prove this resemblance of the past to the future, since all these arguments are founded on the supposition of that resemblance. (Hume 1748, p. 37)

Hume's conclusion is that there is no source from which the principle of the uniformity of nature can receive its justification. It is part of human "habit and custom" to expect the future to resemble the past, but this expectation cannot be rationally justified. Even if the design argument established the existence of God, that would not help. For Hume, the principle of the uniformity of nature is rock bottom – it can't be justified at all.

Kant's demotion of God

Immanuel Kant (1724–1804) is often described as the philosophical giant who effected a grand synthesis of empiricism and rationalism.[30] This broad-brush statement might lead you to expect that his treatment of the principle of parsimony somehow combines the theistic rationalism of Leibniz and the skeptical empiricism of Hume. The truth is a bit messier; Kant retains elements from each, though the rationalist's theism and the empiricist's skepticism both fall by the way.

Kant agreed with Hume that the principle of the uniformity of nature is presupposed in all inductions. But Kant thought that Hume's principle is part of a deeper and more encompassing principle, *that nature is a unity*. By this he meant that everything that happens in nature is governed by a small number of simple, basic laws. Hume's unity of past and future is just a part of this bigger picture. Although Kant agreed with Hume that the proposition that nature is a unity cannot be established by observation or by an *a priori* proof that it is true, he thought that something can be said in its favor – believing that nature is a unity is forced on us by *reason.*

In his *The Critique of Pure Reason*, Kant (1787) cites "the well-known scholastic maxim, that . . . principles must not be unnecessarily multiplied" (A652 B680) and says that it presupposes that nature is a unity. He also says that "reason presupposes the systematic unity of the various powers [that different objects have] . . . and that parsimony in principles is not only an economical requirement of reason, but is one of nature's own laws" (A650 B678). The quest for parsimonious theories and the quest for unifying theories are joint enterprises. Reason obliges us to seek both, and therefore to believe that nature is simple and unified. However, we have no independent assurance that nature has these features. The fact that we cannot know that nature is simple and unified traces back to Hume's skepticism, but the claim that reason requires us to think of nature in this way is not Humean. Kant has a "thicker" notion of reason than Hume did. For Hume, reason is merely the faculty that allows us to recognize the truth of analytic propositions (like "triangles have three sides"). For Kant, reason is something more; it can show us that there are synthetic *a priori* truths (like "every event has a cause"), a possibility excluded from Hume's philosophy.

[30] I am immensely grateful to James Messina for his help on this section.

Kant thinks that the principle of parsimony is different in kind from other principles that also play a central role in organizing our quest for scientific knowledge. For example, he holds that the postulate that every event has a cause is "necessary for the possibility of experience," but he does not accord that status to the postulate that nature is simple. Kant thinks of "experience" as the formation of true, justified, and objective judgments, where by "objective," he has in mind the distinction between the subjective temporal order of experiences and the objective temporal order of events in the world. These can differ, as when you see the door of a house and then see its roof even though the door and the roof exist simultaneously. Kant gives a complicated argument in the Second Analogy of *The Critique of Pure Reason* to show that the causal principle is crucial to drawing this distinction. He concludes that "every event has a cause" is a justified, objective judgment, and that it is *a priori* true. Why isn't "nature is simple" in the same boat? Perhaps Kant would say that its truth isn't needed to ground the distinction between objective and subjective. Whatever the reason, Kant claims that we have no *a priori* guarantee that the world is actually simple. Yet, reason commands us to assume that the laws of nature are simple and unifying.[31]

The logical maxim that instructs us to seek unifying and simple theories has an additional presupposition. The maxim obliges us to think that the kind of unity that nature exhibits is just what would be true if nature were the product of a divine intellect; we must think that nature is *as if* it were made by God (A672 B700). This "as if" marks a fundamental departure from Leibniz and Descartes, who both think that there are good arguments for the existence of God. Kant is convinced that no such argument is to be had. He thinks that the faculty of reason has a tendency to trick us into believing that there is a God. Reason's quest for unification can lead us down the garden path; the

[31] Friedman (1992) and others have maintained that Kant's argument for "every event has a cause" is simultaneously an argument for the stronger thesis that the future will resemble the past, in the sense that there are at least some persistent causal regularities. Whether or not this is right, Kant views the principle of parsimony and the systematic unity of nature in a different light; that's why he discusses them in the Dialectic, not in the Analytic, where the Analogies are located. Kant's faculty of reason (in contrast to the understanding) enjoins us to think, not just that there are persistent regularities, but that those regularities fit into a system that is simple and unifying. Kant does not think that a transcendental argument can show that the laws of nature must have these properties.

result is that we postulate the existence of a Being who unifies *everything*. Here reason over-reaches itself by introducing a postulate that is inherently unknowable. Reason, if not carefully monitored and constrained, leads us to fall victim to a "transcendental illusion" (B350–352). Kant thinks that this is what happened to Leibniz. For Kant, the "as if" statement is something that an atheist can consistently accept. Could atheists consistently maintain, not just that there is no God, but that thinking about God is entirely irrelevant to doing science? For Kant, this is going too far. Thinking of nature as if it were created by God isn't *idle*; on the contrary, it is *indispensable*. We can't develop our science without using the idea of God. Here we see a link to Leibniz.

Leibniz thinks that the existence of God can be established by a rational argument and that the existence of God justifies the principle of parsimony. Kant removes this theistic foundation and replaces it with the faculty of reason; it is reason, not God (or nature either), that is the source of the imperative that directs us to value unification and parsimony. As Kant says, reason does not beg; it commands (A653 B681). But Kant doesn't completely jettison the idea of God. The *idea* of God is indispensable, even though the *existence* of God is unknowable, at least as far as Kant's theoretical philosophy is concerned.

Kant places great weight on the distinction between the theoretical and the practical. Science and morality are different; the point of scientific theories is to provide knowledge of the way the world is, whereas the point of moral principles is to guide behavior. Thought is the theoretical sphere; action is the sphere of the practical. Kant thinks that embracing the logical maxim that bids you search for unifying and parsimonious scientific theories does not oblige you to be a theist. Not so for the oughts of morality! Kant says, in *The Critique of Practical Reason*, that embracing these requires you to believe that there is a God. Theism is a commitment of Kant's practical philosophy, not of his theoretical philosophy (Buchdahl 1969, pp. 523–530; Guyer 2006, p. 234).

Just as Kant draws a distinction between theoretical and practical, he also distinguishes between the theoretical enterprises of biology and physics. It is in biological theories, not in physics, that teleological concepts play an indispensable role. In Section 75 of his 1790 *Critique of the Power of Judgment*, Kant says that there can be no Newton of a blade of grass. His point was not that there can be no general biological theories. Rather, Kant's idea is that Newton showed that a successful physical theory can do without teleological concepts, but that this could never happen in biology. Purely mechanistic theories are fine for planets, but not for plants and animals. In Section 66, Kant says that "the dissectors of plants and animals . . . assume as indisputably necessary the maxim that nothing in such a creature is in vain, and they put the maxim on the same footing of validity as the fundamental principle of all natural science, that nothing happens by chance." Notice that the Aristotelian slogan here resurfaces to express a commitment to biological teleology, not to endorse the broader and non-teleological thesis that we saw at work in Newton's *Rules of Reasoning*.

Returning finally to Kant's thesis that reason obliges us to seek theories that are parsimonious and unifying, I want to consider Kant's claim that it would be irrational to decline to embrace this so-called logical maxim. Kant thinks that the faculty of reason, by definition, aims to find theories of this sort. This is what Kant is getting at when he writes that "what is peculiarly distinctive of reason . . . is that it prescribes and seeks to achieve its systematization, that is, to exhibit the connection of its parts in conformity with a single principle" (A645 B673). This justification of parsimony and unification strikes me as thin. If reason, by definition, obliges us to have these goals, why shouldn't we say "no" to reason and sign up under the banner of *schmeason*, which requires no such commitment? Kant thinks that all hell breaks loose if we take this step. He says that

> the law of reason which requires us to seek for this unity, is a necessary law, since without it we should have no reason at all, and without reason no coherent employment of the understanding, and in the absence of this no sufficient criterion of empirical truth. In order therefore to secure an empirical criterion we have no option save to presuppose the systematic unity of nature as objectively valid and necessary. (A651 B679)

This is a very strong statement – that there is *no* alternative set of rules that provides "a sufficient criterion for empirical truth." My reply is that there do exist criteria for empirical truth that ignore or contravene parsimony and unification. For example, consider the discussion of "counterinduction" in twentieth century explorations of Hume's problem of induction. Counterinduction tells you to infer that the future will go counter to what you observed in the past (Burks 1953, Black 1954, Salmon 1967). Suppose that today you remove half the balls from an urn (without replacement), see what proportion in this sample is green, and then use that information to predict what you'll observe tomorrow when you look at the balls that remain in the urn.

If the frequency of green balls in today's sample turns out to be 25 percent, counterinduction tells you to infer that the frequency of green balls in tomorrow's will be 75 percent. Counterinduction sounds crazy. The philosophical challenge is to show why induction is better than counterinduction. Whether or not this challenge can be met, the fact remains that induction is not the only possible rule for forming judgments about empirical truth.[32]

Whewell's consilience of inductions

William Whewell (1794–1866) was an influential historian and philosopher of science who took Newton's physics as his premier case study for understanding scientific inference. In the third edition of his *Philosophy of the Inductive Sciences*, he says the following about Newton's first rule of reasoning:

> the Rule . . . expresses one of the most important tests which can be given of a sound physical theory. It is true, the explanation of one set of facts may be of the same nature as the explanation of the other class: but then, that the cause explains *both* classes, gives it a very different claim upon our attention and assent from that which it would have if it explained one class only. The very circumstance that the two explanations coincide, is a most weighty presumption in their favour. It is the testimony of two witnesses in behalf of the hypothesis; and in proportion as these two witnesses are separate and independent, the conviction produced by their agreement is more and more complete. When the explanation of two kinds of phenomena, distinct, and not apparently connected, leads us to the same cause, such a coincidence does give a reality to the cause, which it has not while it merely accounts for those appearances which suggested the supposition. This coincidence of propositions inferred from separate classes of facts, is exactly what we noticed in the *Novum Organon Renovatum* (b. ii. c. 5, sect. 3), as one of the most decisive characteristics of a true theory under the name of *Consilience of Inductions.* (Whewell 1968, p. 330)

This passage makes the fairly modest claim that two pieces of evidence are better than one if they come from different "classes of facts" or "kinds of

[32] Instead of arguing that nothing other than reason could furnish a criterion for empirical truth, could Kant maintain that we human beings are psychologically incapable of embracing any alternative to reason? This brings him closer than he wants to be to Hume's position. And the psychological claim seems false.

phenomena."[33] But Whewell also says something stronger; he says that the confirmation provided by two different kinds of phenomena is "one of the most decisive" justifications a theory can have. The section of Whewell's *Novum Organon Renovatum* to which he refers reads as follows:[34]

> We have here spoken of the prediction of facts *of the same kind* as those from which our rule was collected. But the evidence in favour of our induction is of a much higher and more forcible character when it enables us to explain and determine cases of a *kind different* from those which were contemplated in the formation of our hypothesis. The instances in which this has occurred, indeed, impress us with a conviction that the truth of our hypothesis is certain. No accident could give rise to such an extraordinary coincidence. No false supposition could after being adjusted to one class of phenomena, exactly represent a different class, where the agreement was unforeseen and uncontemplated. That rules springing from remote and unconnected quarters should thus leap to the same point, can only arise from that being the point where truth resides.
>
> Accordingly the cases in which inductions from classes of facts altogether different have thus *jumped together*, belong only to the best established theories which the history of science contains. And as I shall have occasion to refer to this peculiar feature in their evidence, I will take the liberty of describing it by a particular phrase; and will term it the *Consilience of Inductions*.
>
> It is exemplified principally in some of the greatest discoveries. Thus it was found by Newton that the doctrine of the Attraction of the Sun varying according to the Inverse Square of this distance, which explained Kepler's *Third Law* . . . explained also his *First* and *Second Laws* . . . although no connexion of these laws had been visible before . . . Here was a most striking and surprising coincidence, which gave to the theory a stamp of truth beyond the power of ingenuity to counterfeit. (Whewell 1968, p. 153)

This second passage advances a bolder thesis than the first one does; it says that consilience of induction *cannot err* – that a false theory cannot pass the

[33] What does it mean for two phenomena explained by a given theory to be of different "kinds?" Whewell (1968, p. 332) realized that this is a good question. Does it just mean that the theory's predecessor failed to treat them as the same? Or should the notion of kinds of phenomena be given an ahistorical interpretation? Another good question concerns Whewell's claim that the testimony of two witnesses provides stronger evidence for a theory that unifies them to the degree that the two are separate and independent; see Sober (1989) for discussion and a counterexample.

[34] This book is Whewell's "renovation" of Francis Bacon's 1620 *Novum Organon*.

consilience test. This was no slip of the pen; the claim of infallibility occurs in other passages:

> No false supposition could after being adjusted to one class of phenomena, exactly represent a different class, where the agreement was unforeseen and uncontemplated. (Whewell 1968, p. 153)

Whewell thinks that this thesis is borne out by the history of science, which "offers no example in which a theory supported by such consiliences, had been afterwards proved to be false" (Whewell 1968, p. 295; see also p. 331). Whatever may be said of the science that Whewell knew about, subsequent scientific developments have not been kind to this bold pronouncement (Van Fraassen 1985, p. 267; Janssen 2002, p. 39). Newton's theory of gravitation was replaced by Einstein's, and the wave theory of light, which Whewell also praises for its consilience, gave way to quantum mechanics.

Perhaps Whewell's argument can be purged of its overstatement. After all, a rule of inference can be highly reliable even if it sometimes leads from true premises to a false conclusion. Maybe the history of science shows that consilience leads to truth often enough that we should give it great weight even though it fails to provide conclusive proof that a theory is true. We saw earlier that Hume maintained that an inductive justification of induction would be circular, but let us nevertheless consider what an inductive justification of the consilience of inductions would look like.

To assess the track record of consilience, we must avoid the mistake of selective attention; we need to count the failures as well as the successes. What would a failure of consilience involve? Consider a biologist who discovers that grass is green because it contains chlorophyll. This finding may prompt the biologist to formulate a general theory, that *all green plants* are green for the same reason. But there is a theory that is even more general; it states that *all green organisms* are green because they contain chlorophyll. This theory makes a prediction about the common iguana (*Iguana iguana*) and it is false. These iguanas are green but they don't contain chlorophyll. For every unifying theory that we think is correct, there are other unifying theories that go astray.[35] The quest for unification has yielded some beautiful successes,

[35] It isn't necessary for my argument that some flesh-and-blood scientist actually used the method of consilience to reach an erroneous solution; the point is that the method leads to a mistake, whether or not anyone ever made the mistake by using

but we should not ignore its many failures if we want to assess its track record.[36]

There is another problem with the track record defense of consilience. As noted, Whewell counted Newton’s theory of gravitation and the wave theory of light as important successes for his consilience principle. He could not have known that this was a mistake. If theories once refuted stay refuted, while theories that are strongly supported by the evidence at hand sometimes turn out to be wrong when new evidence rolls in, we need to watch out for a selection bias; the apparent track record to date over-estimates the success rate of unification and it is hard to say by how much.[37]

My skepticism about track record defenses of consilience carries over to track record defenses of simplicity and Ockham’s razor. These skeptical worries should not be set aside just because a scientific genius claims that Ockham’s razor is a good idea on the grounds that it has worked in practice. The genius I have in mind is Einstein, who said the following in his 1933 Herbert Spencer lecture: “Our experience hitherto justifies us in believing that nature is the realisation of the simplest conceivable mathematical ideas.” Einstein’s eye is drawn to scientific successes, notably his own. However, we must not neglect the failures.

Inspired by Whewell, whom he knew personally, Darwin repeatedly invokes the unifying power of his theory of evolution as one of its chief merits. Here is a passage from his book *The Variation of Animals and Plants under Domestication*, which he published after *the Origin* appeared:

the method. Dividing by zero is a bad idea whether or not anyone ever divided by zero.

[36] Forster (1988) is careful to distinguish the consilience of inductions from the idea that the unifying power of a theory is evidence that the theory is true. The former idea involves a theory’s predicting an observation that was not known when the theory was formulated; the latter can hold even when the data were known before the theory was invented. For discussion of the relevance of such chronological considerations, see Hitchcock and Sober (2004).

[37] Norton (2000, p. 167) warns against an additional selection effect in connection with track-record defenses of simplicity: “Imagine that some true theories can be expressed simply in a mathematical formalism ready to hand and that others cannot. In this scenario we are most likely to find the first type of theory and less likely to find the second. So the fact that our physics texts are bursting with theories in simple mathematical clothing would just reflect our inability to discover the ones that require complicated expression.”

> In scientific investigations it is permitted to invent any hypothesis, and if it explains various large and independent classes of facts it rises to the rank of a well-grounded theory. The undulations of the ether and even its existence are hypothetical, yet every one now admits the undulatory theory of light. The principle of natural selection may be looked at as a mere hypothesis, but rendered in some degree probable by what we already know of the variability of organic beings in a state of nature, – by what we positively know of the struggle for existence, and the consequent almost inevitable preservation of favourable variations, – and from the analogical formation of domestic races. Now this hypothesis may be tested, – and this seems to me the only fair and legitimate manner of considering the whole question, – by trying whether it explains several large and independent classes of facts; such as the geological succession in organic beings, their distribution in past and present times, and their mutual affinities and homologies. If the principle of natural selection does explain these and other large bodies of facts, it ought to be received. (Darwin 1868, vol. 1, p. 9; quoted in Gayon 1998, pp. 31–32)

This passage avoids Whewell's overstatement, but sometimes Darwin was less cautious, as when he added the following remark to the sixth and final edition of *the Origin*, published in 1872:

> It can hardly be supposed that a false theory would explain, in so satisfactory a manner as does the theory of natural selection the several large classes of facts above specified. It has recently been objected that this is an unsafe method of arguing. But it is a method used in judging of the common events of life, and has often been used by the greatest natural philosophers. The undulatory theory of light has thus been arrived at; and the belief in the revolution of the Earth on its own axis was until lately supported by hardly any direct evidence. (Darwin 1959, p. 748)

Notice Darwin's comment that the rules of reasoning at work in his theorizing are not limited to biology; they are used in everyday life and in physics. Once again we see an old idea reaffirmed: rules of reasoning are not subject-matter specific.

Darwin (1859, p. 459) starts the last chapter of *the Origin* by saying that the book is "one long argument" for his theory of evolution. He structured this argument under Whewell's influence, so it must have disappointed Darwin that Whewell never accepted the theory. Neither did another important philosopher of the time. This philosopher was tepid about Darwin's theory and tangled with Whewell over the nature of scientific inference; it is to this opponent of Whewell that I now turn.

Mill tries to cut the razor down to size

In Chapter 24 of his 1865 book *An Examination of Sir William Hamilton's Philosophy*, John Stuart Mill (1806–1873) takes issue with Hamilton's views on the principle of parsimony. Mill says that Hamilton's big mistake is thinking that the principle of parsimony needs to be grounded on an "ontological theory" – that is, on a theory about the way the world is. Mill quotes Hamilton as saying that "nature never works by more and more complex instruments than are necessary." He says that Hamilton approves of Aristotle's statement that "God and Nature never operate without effect . . . they never operate superfluously . . . but always through one rather than through a plurality of means." Mill raises an eyebrow at Hamilton's "never." How could any human being ever know that there isn't even one exception in the whole history of the universe to Hamilton's sweeping pronouncements?

Mill's positive thesis is that the principle of parsimony is "a case of the broad practical principle, not to believe anything of which there is no evidence . . . The assumption of a superfluous cause, is a belief without evidence; as if we were to suppose that a man who was killed by falling over a precipice, must have taken poison as well." For Mill, the principle of parsimony is purely methodological; it "implies no theory concerning the propensities or proceedings of Nature. If Nature's ways and inclinations were the reverse of what they are supposed to be, it would have been as illegitimate as it is now, to assume a fact of Nature without any evidence for it."

Mill's justification of the principle of parsimony rests on an idea I'll call *evidentialism*; this is the thesis that every belief we have must be justified by its being supported by evidence.[38] Mill's evidentialism conflicts with Carnap's (1950) claim that we are entitled to adopt various fundamental assumptions – for example, that physical objects exist – even though we can offer no evidence that these assumptions are true. The justification for adopting such

[38] Formulated more quantitatively, evidentialism asserts one's confidence should be "proportional" to strength of evidence.

assumptions is that they are useful; for Carnap, believing that there are physical objects is *permissible*, not *obligatory*. I will discuss Carnap's suggestion in Chapter 5; I mention it here because it shows that Mill's evidentialism is not the only game in town. But even granting evidentialism, there is a problem with Mill's justification of Ockham's razor. In the example about the man's death, Mill says that we should assert a single cause (the falling over the precipice) and should not add to this the unjustified hypothesis that the man was poisoned. Notice that "falling" and "falling and poison" are compatible with each other. Matters change if we wish to compare hypotheses that are incompatible – "falling and no poison" versus "falling and poison." Suppose we have no evidence as to whether the man was poisoned. Ockham's razor is often taken to say that we should prefer the first hypothesis over the second in this pair of incompatible alternatives. Mill's reading of the maxim licenses no such conclusion. Mill is focusing exclusively on the *razor of silence*, but the principle of parsimony is often thought to embody a *razor of denial*. Whenever the principle of parsimony is offered as a reason for choosing between incompatible theories, it can't be the razor of silence that is doing the work. Perhaps evidentialism has something to say in favor of the razor of denial, but Mill does not make the connection clear.

James Clerk Maxwell, the evangelical physicist

Newton, as we have seen, was famous for providing a physical theory that unifies terrestrial and celestial motion. Almost two centuries later, James Clerk Maxwell (1831–1879) achieved a second grand unification – the unification of electricity and magnetism. Maxwell, like Newton, was a devout Christian. And like Newton again, he saw an important connection between science's search for unifying theories and the actions of a benevolent deity.

The connection that Maxwell draws between God and the principle of simplicity needs to be traced out carefully. Sometimes Maxwell sounds like William Paley (1743–1805), who argued in his 1802 book *Natural Theology* that complex adaptive structures such as the vertebrate eye are evidence for the existence of God. Maxwell follows Paley, though for Maxwell it is simple physical laws, not complex biological adaptations, that point to a theistic conclusion. When Maxwell argues in this way, he is not saying that the existence of God provides a reason to expect natural laws to be simple. Rather, he is saying the converse – that the fact that the laws of nature are simple is evidence for the existence of God.[39] This line of thinking does not have God guaranteeing the correctness of Ockham's razor.

Maxwell sometimes deviates from Paley's pattern. In *Natural Theology*, Paley wants to build his case for God's existence solely by constructing inferences that are based on what we observe in nature; that's what *natural* theology is all about. Maxwell, an evangelical, thought this too narrow a foundation. Scripture must play a central role in the discussion. The Bible assures us that God exists, and God's existence assures us that nature is governed by simple unifying laws. Without this theological guarantee, it is entirely possible that the human mind's passion for unification is a psychological quirk that we should attempt to resist. Maxwell's description of the human mind has a Kantian ring: "the human mind cannot rest satisfied with the mere phenomena which it contemplates, but is constrained to seek for the principles embodied in the phenomena" (Jones 1973). Given this psychological fact, Maxwell (1856, Vol. I, p. 377) asks whether this yearning for simplicity and unification is to be trusted: "Are we to conclude that these various departments of nature in which analogous laws exist, have a real interdependence; or that their relation is only apparent and owing to the necessary conditions of human thought?" For Maxwell, God lays this worry to rest (Stanley 2012).

Morgan's canon

C. Lloyd Morgan (1852–1936) was a Darwinian who grew dissatisfied with the Darwinian penchant for anthropomorphizing the mental lives of animals. Darwin and his follower George Romanes (1848–1894) emphasized the continuity of human beings with other animals, not just with respect to morphology, but also in connection with mind and behavior. For example, in Chapter 2 of his 1871 book *The Descent of Man*, Darwin tells stories about animal behavior to support the claim that language, self-consciousness, an aesthetic sense, and belief in God are qualitatively similar to (though not identical with) mental faculties found in non-human organisms. Here is a characteristic passage:

[39] Here Maxwell follows the lead of his predecessor in Cambridge, William Whewell, who makes this kind of argument in his 1833 Bridgewater Treatise.

> The tendency in savages to imagine that natural objects and agencies are animated by spiritual or living essences, is perhaps illustrated by a little fact which I once noticed: my dog, a full-grown and very sensible animal, was lying on the lawn during a hot and still day; but at a little distance a slight breeze occasionally moved an open parasol, which would have been wholly disregarded by the dog, had any one stood near it. As it was, every time that the parasol slightly moved, the dog growled fiercely and barked. He must, I think, have reasoned to himself in a rapid and unconscious manner, that movement without any apparent cause indicated the presence of some strange living agent, and no stranger had a right to be on his territory. (p. 67)

A year later, in his *Expression of the Emotions in Man and Animals*, Darwin (1872a) applies the terms "joy" (p. 76), "affection" (p. 117), "terror" (p. 77), "insulted" (p. 133), and "discontented, somewhat angry, or sulky" (p. 232) to non-human animals. Morgan regarded such statements as naïve. To counteract the tendency to read human mental states into the behaviors of other animals, he espoused a "canon." In his 1894 *Introduction to Comparative Psychology* he puts it like this:

> In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale. (Morgan 1894, p. 53)

In the book's 1903 second edition, Morgan adds the following clarification:

> To this, however, it should be added, lest the range of the principle be misunderstood, that the canon by no means excludes the interpretation of a particular activity in terms of the higher processes, if we already have independent evidence of the occurrence of these higher processes in the animal under observation. (Morgan 1894, p. 59)

Does this addition render the principle vacuous? An even-handed evidentialism of the sort that Mill espoused will say that *any* attribution of a faculty (whether it is higher or lower) to an organism requires evidence. This, apparently, is not what Morgan intends. Rather, Morgan thinks that our "default" assumption should be that the organism we are studying has lower faculties but not higher; if the evidence forces us to revise that assumption, so be it. This is the idea that hypotheses that attribute lower faculties to organisms are *innocent until proven guilty*, while hypotheses that attribute higher faculties are *guilty until proven innocent*.

Morgan’s canon exerted a powerful influence on the development of psychology in the twentieth century. Many, perhaps most, of its proponents interpreted it as a version of the principle of parsimony. For example, the influential behaviorist B. F. Skinner (1904–1990) took the canon to mean that one should explain an organism’s behavior without attributing any mental states at all to it if that is possible (and he *did* think that this is possible). Since Morgan said that postulating lower mental faculties is preferable to postulating higher faculties, a natural consequence of the canon is that postulating zero faculties is best of all. Skinner (1938, p. 4) refers to Morgan’s canon as a principle of parsimony. Non-behaviorists also found a lot to like in the canon. When the cognitive revolution began in the 1960s and cognitive ethology started to develop, students of animal cognition espoused a “principle of conservatism,” the idea being that it is a good idea to be conservative in the mental abilities one attributes to organisms to explain their behaviors (Cheney and Seyfarth 1990).

Although most friends of Morgan’s canon have thought that it is an instance of the principle of parsimony, this is not how Morgan saw things. He held that the simplest hypothesis about other organisms is that they have *higher* cognitive faculties. He writes:

> Is it not simpler to explain the higher activities of animals as the direct outcome of reason or intellectual thought, than to explain them as the complex results of mere intelligence or practical sense experience? (Morgan 1894, p. 54)

Why is this the simpler explanation? Morgan answers by proposing two analogies:

> The explanation of the genesis of the organic world by direct creative fiat is far simpler than the explanation of its genesis through the indirect method of evolution.
>
> The formation of the canyon of the Colorado by a sudden rift in the earth’s crust, similar to those which opened during the Calabrian earthquakes, is simpler than its formation by the fretting of the stream during long ages under various meteorological conditions.

In these and other examples, the explanation that Morgan says is simpler has a single powerful cause that achieves its effect in one fell swoop, while the one judged to be more complex postulates the cumulative action of many less potent causes acting together (Clatterbuck 2015). Morgan observes that "in these cases and in many others, the simplest explanation is not the one accepted by science (p. 55)." Applying this point to comparative psychology, Morgan concludes that the principle of simplicity *encourages* anthropomorphism[40]; it is not a prophylactic for keeping anthropomorphism at bay.[41] The point of the canon, according to Morgan, is to prevent us from falling prey to *over*simplification.

Morgan's idea that his canon is an anti-simplicity principle is interesting, but even more interesting is Morgan's conviction that the canon is justified by Darwin's theory of evolution. This claim will be striking for those who see Morgan's canon as an instance of the principle of parsimony; it contrasts sharply with the other approaches to Ockham's razor that I have surveyed in this chapter. The other authors I have described think that Ockham's razor is a very general principle that applies across all scientific subject matters, and so does not depend for its justification on the findings of any particular science. Morgan thinks his canon has a more limited applicability, so he sees room for the possibility that it might be justified by the findings of a single science.

Morgan formulates the problem of justifying his canon by asking the reader to consider three "divergently ascending grades of organisms." Species *a* represents human beings and *b* represents dogs; he says that *a* has ascended to a higher level than *b*, and *b* has risen higher than *c*. Each of these organisms may exhibit to some degree each of three "ascending faculties ... in mental development," which he numbers 1, 2, and 3. How might the degree of development of these three mental faculties differ among the three taxa? Morgan says that the most plausible view, according to Darwin's theory, is given by the idea he terms the *method of variation*, "according to which any one of the faculties 1, 2, or 3, may in *b* and *c* be either increased or reduced relative to its development in *a* (p. 57)." The method of variation places no constraints

[40] Morgan (1894, pp. 369–372) changes his tune when he discusses a deceitful Maltese terrier.

[41] Morgan was not the first to connect simplicity and anthropomorphism. In his *Treatise of Human Nature*, Hume (1739, 1.3.16.3) says that "it is from the resemblance of the external actions of animals to those we ourselves perform, that we judge their internal likewise to resemble ours ... When any hypothesis, therefore, is advanc'd to explain a mental operation, which is common to men and beasts, we must apply the same hypothesis to both." Hume sounds like he is channeling Newton's *Rules of Reasoning* here. See Buckner (2013) for discussion.

on the distribution of characteristics among the taxa that the evolutionary process produces; the method of variation is the method of anything goes. Morgan takes the method of variation to have the consequence that

> any animal may be at a stage where certain higher faculties have not yet been evolved from their lower precursors; and hence we are logically bound not to assume the existence of these higher faculties until good reasons shall have been shown for such existence. (Morgan 1894, p. 59)

Morgan is right that the method of variation says that it is possible for an organism to have lower but not higher faculties, but the method equally allows that an organism may have higher but not lower faculties. Yet, Morgan's canon is selective; it warns against the possibility of mistakenly attributing a higher faculty, but does not mention that one also can err in denying that a higher faculty is present. Since the method of variation describes no asymmetry between "higher" and "lower" in the evolutionary process, Morgan's attempt to derive an epistemological asymmetry between "higher" and "lower" from that method was doomed from the start.

The distinction between the razor of silence and the razor of denial cries out for application to Morgan's discussion. He says that we are "bound not to assume the existence of . . . higher faculties" unless we have evidence for their existence. Does this mean that we should assume the non-existence of higher faculties in that circumstance, or is Morgan saying merely that we should withhold assent? The distinction between *assuming notX* and *not assuming X* is all-important. If we sometimes ought to *deny* that dogs have higher mental faculties on the grounds that attributing higher faculties to them isn't needed to explain their behavior, Morgan's justification of his canon does not explain why. In addition, Morgan's argument does not identify anything special about *higher* mental faculties. It is equally true that one should not assume the existence of *lower* faculties in an organism unless one has evidence that they are present. Despite its evolutionary trappings, Morgan's justification of his canon comes to little more than a plea for evidentialism. We will see in Chapters 3 and 4 that the bearing of parsimony on psychological hypotheses is richer than this. Yet, the fact remains that Morgan's justification of his canon is thin.

Maybe the evolutionary process exhibits an asymmetry that Morgan neglected to put to work in his discussion of the canon. Let us think about this possibility by considering his terminology of "higher" and "lower." If Morgan's use of "higher" and "lower" is Darwinian, what do these terms mean? There now is debate concerning where Darwin stood on the question of "evolutionary progress" (Shanahan 2004, pp. 285–294). Each side has its favorite citations. On the one hand, Darwin (1872b) says in a letter to the American paleontologist Alpheus Hyatt that "after long reflection, I cannot avoid the conviction that no innate tendency to progressive development exists." On the other, Darwin writes at the very end of *the Origin* that "the production of the higher animals" is "the most exalted object which we are capable of conceiving." Perhaps these two statements do not conflict – the *process* of evolution has no tendency to make progress, but this does not prevent us from admiring the *products* of that process (Sober 1994, 2011a). Rather than enter into this debate, I want to focus on a distinction that Darwin drew and that has remained standard in evolutionary biology ever since. Consider a lineage that extends from an ancestral population to a descendant population. If there was evolution, some of the traits found in descendants were not present in their ancestors, but this allows that there may be other traits that are present in both. Let us call traits "old" when they are present in both ancestors and descendants and "new" if they are found just in the descendants. These terms describe when traits first appeared in the lineage. No trait can be both old and new. Recognizing the difference between old and new traits does not require one to say that descendants are better adapted or more complex or "higher" on some ladder-like scale of being. Can Morgan's canon be vindicated by interpreting it as saying that it is better to explain an organism's behavior by attributing old faculties to it rather than new ones? For Morgan, abstract reasoning is a higher faculty and sense perception is lower. If higher = new and lower = old, this means, since human beings now have both of these faculties, that there should exist an ancestor of present day human beings that has the lower but not the higher faculty. How does this idea bear on the question of which mix of higher and lower faculties one should find in our contemporaries – in dogs, for instance? Does it mean that we should expect dogs to have only the lower faculty, rather than both the lower and the higher faculties?

The answer is *no*. To see why, consider Figure 1.9, which represents a phylogenetic tree. At the top are the "leaves" that represent different groups of contemporary organisms with their characteristics described. As you move down the page, you are going back in time with contemporary groups finding

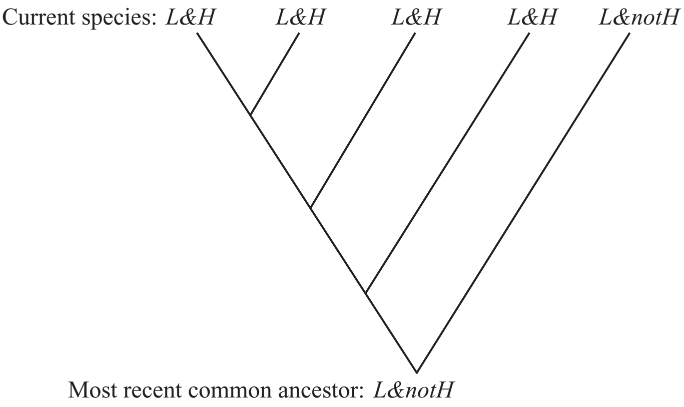


Figure 1.9

their common ancestors as lineages coalesce. At the bottom of the tree is the most recent common ancestor of the leaves. The left-most leaf represents human beings. We have the lower and the higher faculties (*L&H*). Trace the lineage that leads to us backwards in time and you eventually reach an ancestor that has *L&notH*. The higher faculty makes its appearance in this tree and is then retained in such a fashion that almost all of the leaves have *L&H*. Only one of the leaves is a "living fossil," exhibiting the mix of faculties that is found in the most recent common ancestor.

Now consider this puzzle: I choose one of the non-human leaves in this tree at random and tell you that it exhibits a behavior that would occur if the leaf had *L&H*, but also would occur if it had *L&notH*. What conclusion should you draw about the faculty or faculties that this leaf probably possesses? Given that almost all of the leaves have *L&H*, your best bet is that the leaf in question has both the lower and the higher faculties. It isn't impossible that the leaf has *L&notH*, but that is a long shot.[42] The point of this figure is not that evolution must always result in contemporary organisms having higher faculties almost as frequently as lower ones. The point is that defining "higher" and "lower" to mean *new* and *old* does not give Morgan the result he wants. If it were somehow a principle of evolution that new characteristics should now be

[42] It is a feature of the tree in this figure that the leaves have L more often than they have H. If faculties can never be lost, then L cannot occur less frequently than H. But, as Morgan realized, loss and gain are both evolutionary possibilities.

rare and old ones should now be common, that would vindicate Morgan's canon. But, alas, there is no such principle.

Suppose we entirely abandon an evolutionary perspective and define "higher" and "lower" as follows:

> Psychological faculty *H* is higher than faculty *L* if and only if an organism that has *H* must also have *L*, but not conversely.

Morgan often cites the examples of abstract reasoning, trial-and-error learning via sensory association, and reflex, saying that the first is higher than the second while the second is higher than the third. Although the method of variation says that any of these can be present in an organism without the others, in practice Morgan's picture seems to be that abstract learning requires trial-and-error learning, which in turn requires reflex (but not conversely). The idea is that higher faculties will be present in an organism only if lower faculties are also in place. This underwrites the expectation that organisms will have the higher faculty *H* less often than they'll have the lower faculty *L*. What is left hanging is whether *L&notH* must occur more often than *L&H* (Sober 1998, pp. 233–235). Here again, the leaves on the tree in Figure 1.9 are instructive.

Concluding comments

This chapter, despite its title, has not focused on the *uses* to which Ockham's razor has been put, though I did discuss a few examples under that heading. Rather, my main focus has been on attempts pre-1900 to *justify* the principle of parsimony.[43] In the Introduction I gave a preliminary formulation of the razor that says that simpler theories are "better" than theories that are more complex. This principle isn't difficult to justify if we are modest in what we mean by "better." Simpler theories are easier to understand and remember, and they often strike us as beautiful. Isn't that a sufficient justification for thinking that simpler theories are "better"? It is not, if we want to know whether and why parsimony is *epistemically* relevant.

There are various ways to spell out the concept of epistemic relevance. Suppose that a simpler theory (*S*) and a more complex theory (*C*) are logically

[43] This is why I have not discussed several important philosophers who emphasized the role of simplicity in science. Mach (1898) and Poincaré (1914) are examples.

incompatible with each other, but both are compatible with the observations you now have. Here are three questions: does the fact that *S* is simpler than *C* help justify the claim that

- *S* has a higher probability of being true than *C*?
- *S* is better supported by the observations than *C* is?
- *S* will make more accurate predictions in the future than *C* will?

These questions are not answered by pointing out that *S* is beautiful and easy to understand and remember. To put the point crudely, the issue is whether parsimony is relevant to figuring out what the *world* is like, rather than just reflecting some psychological fact about which theories we happen to like. In the next chapter, I'll refine these three questions and explain why I bother to separate them.

We can further clarify what epistemic relevance means by contrasting it with prudential relevance. The famous "wager" of Blaise Pascal (1623–1662) is useful here. Pascal argued that even if there is no evidence that God exists, it is prudent to believe in God.[44] Notice that Pascal's thesis refers both to a *proposition* and a *person*; he says that even if there is no evidence for the proposition that God exists, a person still can benefit by believing it. Epistemic relevance has to do with propositions, prudential relevance with people. In the three bulleted questions displayed above, I am comparing two theories and there is no mention of theorists.

The distinction drawn earlier between the razor of silence and the razor of denial is worth bearing in mind as we investigate these questions. As Mill made clear, it isn't hard to explain why the hypothesis *X* is superior to the hypothesis *X&Y* when you have evidence for *X* and no evidence at all that bears on *Y*. The razor of silence tells you to slice away *Y*, not in the sense of denying that *Y* is true, but in the sense of remaining silent about it. In contrast, when the razor of denial tells you to prefer the hypothesis *X&notY* over the hypothesis *X&Y*, a different razor is being used, and a different justification is needed.

God plays a prominent role in the history I've reviewed in this chapter. Descartes, Leibniz, Newton, and Maxwell proposed theistic justifications for Ockham's razor while Aristotle, Hume, Kant, Mill, and Morgan paint a very different picture. We will see in the next chapter that the waning of theistic

[44] For discussion of the wager, see Mougin and Sober (1994).

discussion and its replacement by secular ideas is a dominant theme in the twentieth century. It pretty much goes without saying in that century that God is not the answer. Atheists and agnostics will regard this as progress, but theists also have reason to be wary of the suggestion that God is the solution to the puzzle about parsimony. I illustrate this reason by way of a cautionary tale.

Newton believed that his laws of motion do not guarantee the stability of the solar system; he concluded that God periodically intervenes to keep things going smoothly. Newton did not see how a naturalistic explanation could be developed, so he invoked the supernatural instead. In his 1796 *Exposition of the System of the World*, Pierre-Simon Laplace (1749–1827) argued that Newton's appeal to divine intervention was unnecessary. Using Newton's own laws as a platform on which to build, Laplace developed an analysis that Newton did not anticipate, one that entails that planetary orbits are stable.[45] Atheists and agnostics have no time for explanations that appeal to divine intervention, but theists also have reason to be wary. Theists should realize that theistic explanation may be *premature*. As it is in physics, so it is in philosophy of science. In the next chapter we will examine several non-theistic justifications of Ockham's razor. Philosophers, scientists, and statisticians in the twentieth century have had a lot to say about this.

Many of the philosophers and scientists I have surveyed in this chapter link the epistemology of Ockham's razor with a metaphysical thesis. They think that simplicity and unification are epistemically relevant only if nature is simple. The next two chapters chip away at that link.

[45] Scientists now think that Laplace was wrong. The standard contemporary view is that the planetary system is not stable, though the planets will continue to revolve around the Sun without colliding for the next few billion years (Hayes 2007; Laskar 2008).

2 The probabilistic turn

Discussion of parsimony took a probabilistic turn in the twentieth century.[1] The project was to use probability theory to analyze and justify Ockham's razor. Not all of these efforts succeeded, but two of them did. I think there are two "parsimony paradigms" in which probability ideas show that parsimony is epistemically relevant. The two paradigms were developed within two different philosophical frameworks for understanding probability; one paradigm finds its home in Bayesianism, the other in frequentism. To set the stage for investigating probabilistic approaches to Ockham's razor, I'll start this chapter by providing a brief (and I hope accessible) primer on probability. But first I want to say a little about Bayesianism and frequentism.

Two philosophies of probability

Bayesianism is a philosophy of inference that traces back to a mathematical result (a theorem) obtained by Thomas Bayes (1701–1761). Bayes's (1764) theorem describes how the probability you assign to a hypothesis should be influenced by the new evidence you acquire. Bayesianism is now a general philosophy of scientific reasoning that has grown richer and more detailed than its eighteenth-century beginnings. This philosophy says that scientific reasoning has the attainable goal of figuring out how probable different scientific hypotheses are, given the evidence at hand. Or more modestly, it maintains that science is in the business of figuring out which hypotheses are more probable than which others, again in the light of the evidence. Either way, science crucially involves thinking about the probabilities of hypotheses.

[1] I borrow this phrase from Richard Rorty's influential anthology of 1967, *The Linguistic Turn*, which documented the emphasis on language as a philosophical subject in the previous eighty years. Rorty got the expression from Gustav Bergmann (1906–1987).

Bayesianism was not the dominant philosophy of probabilistic inference that scientists themselves embraced in the twentieth century. Rather, the dominant mode of thought was frequentism. Frequentism does not have the simple unity that Bayesianism exhibits; rather, it is a varied collection of ideas about how observations should be used to evaluate hypotheses. Frequentism uses probability ideas in this enterprise just as Bayesianism does, but its basic idea is different. The first commandment of frequentism is: *thou shalt not talk about the probabilities that hypotheses have!* The claim that science has the job of assessing how probable different theories are may sound like an unremarkable truism, but this innocent-sounding remark is something that frequentists categorically reject.

The difference between frequentism and Bayesianism is often characterized in terms of what each philosophy takes the concept of probability to mean. The standard picture is that Bayesians think that probability means rational degree of certainty whereas frequentists define probability in terms of frequency. When you think about your probability of getting lung cancer, given that you smoked lots of cigarettes over many years, Bayesians take this probability to represent how confident you should be that you'll get cancer, given your history of smoking, whereas frequentists take the probability to represent how frequently heavy smokers get lung cancer. Viewed in this way, Bayesianism is about something subjective (= in the mind of a rational subject) and frequentism is about something objective (= out there in the external world). If the two philosophies have different subject matters, why is there conflict between them?[2] Why can't these partisan schools see that probability has both a subjective and an objective meaning (as Carnap 1950 recognized) with each *ism* going its own way? Why can't people just get along? The answer is that Bayesianism and frequentism fundamentally disagree about what the goals of science ought to be. There is more to the debate than a question about the meaning of the word "probability." But even the idea that each school is wedded to a single interpretation of probability is too simple.

On the one hand, there are situations in which Bayesian inferences can be carried out by using probabilities that are as objective as any frequentist could wish. If I tell you what the frequency is of tuberculosis in Wisconsin, that Susan

[2] There is another usage of this terminology, as when people claim that various norms are objective. Here the thought is that the norms are correct and non-arbitrary. Many Bayesians are objectivists in this sense.

lives in that state, and that her tuberculosis test came out positive (where the test procedure produces erroneous results with a certain frequency), you can calculate the probability that Susan has tuberculosis, given her test result. We'll see in a moment how Bayesians do this calculation. The present point is that the probabilities used in this Bayesian calculation are all about objective matters of fact. Bayesians can go to work on frequencies!

On the other hand, there are good reasons why the probabilities that frequentists discuss often should not be interpreted as frequencies. Frequentists are happy to talk about the probability that a fair coin has of landing heads if it is tossed. Fairness means that the value of this probability is ½. But fair coins often fail to have frequencies that match this probability. For example, suppose you toss a fair coin three times and then destroy it. The frequency of heads in the short lifetime of this coin will not equal 50 percent. For this simple reason, you can't equate probability with actual frequency. You may reply that the relevant frequency idea is hypothetical long-run frequency. Although a fair coin won't land heads 50 percent of the time if it is tossed just once, the suggestion is that if a coin is fair, then the frequency of heads will converge on 50 percent if you toss the coin again and again. What's wrong with that? Let us consider what "converge" means. Here is one interpretation:

> A coin has a probability of landing heads of ½ precisely when the frequency of heads will get closer and closer to 50 percent as the coin is tossed repeatedly.

This is false. It is possible for a fair coin to produce two heads in the first four tosses and three heads in the first five. There need be no lockstep, monotonic approach to 50 percent. We can replace this flawed suggestion with something that is true. Consider any small positive number you please; call it ε ("epsilon").

> A coin has a probability of landing heads of ½ precisely when the probability approaches 1 that the frequency of heads will be within ε of 50 percent as the number of tosses approaches infinity.

This is one version of the *law of large numbers*. Notice that the concept of probability appears on both sides of this biconditional. This is not a proper definition; it is circular. For this reason, the law of large numbers, though true, does not provide an *interpretation* of probability in the required sense.[3]

[3] Here's a third suggestion for defining probability in terms of frequency: a coin has a probability of landing heads of ½ precisely when the coin would have to land heads

Despite its name, frequentism as a philosophy of scientific inference has no commitment to interpreting probability in terms of the idea of frequency – either actual or hypothetical.

Although defining Bayesianism and frequentism in terms of their different interpretations of probability is too simple, it does contain an ounce of truth. Bayesians *often* equate probability with rational degree of certainty and frequentists *always* want probability to be more objective than this. But the heart of the matter is that the two philosophies propose different epistemologies, not different semantics. Frequentists want assignments of values to probabilities to have an "objective justification." It should be possible to defend one's assignments by citing frequency data or an empirically justified theory, for example. It isn't good enough to say "well, my probability assignment simply reflects how certain I am in the proposition in question." When I talk about objectivity in what follows, I have in mind this epistemic usage.

A probability primer and the basics of Bayesianism

Before discussing the partisan worlds of Bayesianism and frequentism, I'll begin with the mathematical core of the probability concept itself. This is something on which Bayesians and frequentists agree.

Probability assignments always rest on assumptions. For example, if you assume that the deck of cards before you is standard and that the dealer is dealing you cards "at random," you can conclude that the probability that the first card you are dealt will be an ace of spades is $\frac{1}{52}$ and that the probability that the first card you receive is either an ace or a jack is $\frac{8}{52}$. Without the assumptions mentioned, these probability assignments can be incorrect. I will make the role of assumptions explicit in my description of probability by adding a subscript "A" to the canonical axioms of probability theory described

50 percent of the time if it were tossed an infinite number of times. Although this biconditional is not circular, there still is a problem. It is not impossible for a fair coin to land heads each time it is tossed, even if it is tossed an infinite number of times. True, the probability of the infinite sequence HHHH . . . is zero. However, you can't equate impossibility with a probability of zero. The probability of *any* infinite sequence (including the alternating sequence HTHTHT . . .) is zero if the coin is fair.

by Kolmogorov (1950):

$$0 \leq \mathrm{Pr}_{\mathrm{A}}(H) \leq 1.$$

$$\mathrm{Pr}_{\mathrm{A}}(H) = 1 \text{ if A logically entails } H.$$

$$\mathrm{Pr}_{\mathrm{A}}(H \text{ or } J)$$
$$= \mathrm{Pr}_{\mathrm{A}}(H) + \mathrm{Pr}_{\mathrm{A}}(J) \text{ if A logically entails that } H \text{ and } J \text{ are incompatible.}$$

$\mathrm{Pr}_{\mathrm{A}}(H)$ represents the probability of the proposition H under the assumptions codified in the propositions A. Applying probability to a problem involves isolating a class of propositions that are to be evaluated. In the card example, the propositions concern the different cards you may be dealt, not whether it will rain tomorrow. Notice that probability in the above axioms is a mathematical *function*: it maps propositions onto numbers. Two different probability functions may assign different numbers to the same proposition. The model I just described says that the deck is standard and that cards are dealt at random, with the result that Pr_{A}(the first card you are dealt will be an ace of spades) $= \frac{1}{52}$. If we thought the deck was made of fifty-two such aces, we would use a different probability function, $\mathrm{Pr}_{\mathrm{B}}(-)$ according to which Pr_{B}(the first card you are dealt will be an ace of spades) $= 1$.

Here are three consequences of the axioms just stated that do not depend on what assumptions go into A: (i) Tautologies have a probability of 1 and contradictions have a probability of 0; (ii) If propositions H and J are logically equivalent, then $\mathrm{Pr}_{\mathrm{A}}(H) = \mathrm{Pr}_{\mathrm{A}}(J)$; (iii) $\mathrm{Pr}_{\mathrm{A}}(H) = \mathrm{Pr}_{\mathrm{A}}(H \& J) + \mathrm{Pr}_{\mathrm{A}}(H \& not J)$. This last equality follows from (ii) and the third axiom; it is called *the theorem of total probability*.

The third axiom describes how the probability of a disjunction is settled by the probabilities of the disjuncts if the disjuncts are incompatible with each other. But what if the disjuncts are not mutually exclusive? There is a general principle available here that you can visualize by thinking about probabilities in terms of the diagrams that John Venn (1834–1923) invented. Figure 2.1 shows a square in which each side has a length of one unit. Let's suppose that each point in the square represents a possible way the world might be. Each proposition that we might want to talk about can be associated with a set of points in the square – the set of possible situations in which the proposition is true. The area of the square is 1, which conveniently is also the maximum value that a probability can have. Tautologies are true in

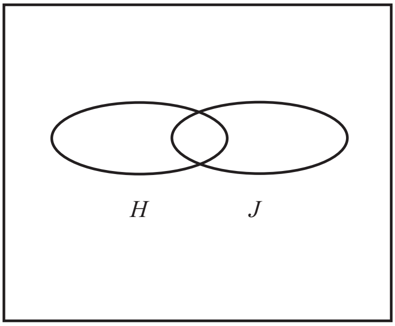


Figure 2.1

all possible situations; they fill the whole unit square. The figure represents propositions H and J as two ovals. The intersection of the two ovals – their area of overlap – represents the conjunction $H\&J$. Since there is a region of overlap, the two propositions are compatible with each other; there are situations in which both are true. I hope the Venn diagram makes it obvious that

$$\Pr(H \text{ or } J) = \Pr(H) + \Pr(J) - \Pr(H\&J).$$

The reason for subtracting $\Pr(H\&J)$ is to insure that the area of overlap is not double-counted. When $\Pr(H\&J) = 0$, the above equality reduces to the special case described in Axiom 3.

What can be said about the probability of conjunctions? This is where we need to define the concept of *probabilistic independence*:

> Propositions H and J are probabilistically independent in probability model A precisely when $\Pr_A(H\&J) = \Pr_A(H) \times \Pr_A(J)$.

When you flip a fair coin twice, the probability of getting a head on the first toss is $\frac{1}{2}$ and the probability of getting a head on the second is also $\frac{1}{2}$. The tosses are probabilistically independent; the probability of getting heads on both tosses is $\frac{1}{4}$. That is a contingent empirical fact about coin tossing; it is logically possible for tosses to be probabilistically dependent. Suppose we lived in a world in which there are two kinds of coins: 50 percent of the coins have two heads and 50 percent have two tails. You select a coin at random and toss it repeatedly. Under the assumptions stated, $\Pr_A$(Heads on the first toss) $=$ $\Pr_A$(Heads on the second toss) $= \frac{1}{2}$. However, it's also true that

Pr_{A}(heads on both the first and second tosses) $= \frac{1}{2}$. Independence fails. In this fanciful world, knowing the outcome on the first toss would give you information about what will happen on the second. In the real world, the tosses are independent; knowing the outcome of the first toss doesn't change the probability you assign to the second.

Probabilistic independence and logical independence are different. Propositions X and Y are logically independent precisely when all four conjunctions of the form $\pm X \& \pm Y$ are logically possible (i.e., non-contradictory). For example, "it is raining" and "you are carrying an umbrella" are logically independent of each other. However, if you follow the advice of accurate weather forecasts, these two propositions will be probabilistically dependent on each other. Consider any two propositions that are neither tautologies nor contradictions: if they are probabilistically independent, then they are logically independent, but the converse implication does not hold.

		color of boat on Tuesday		
		green (p = 0.2)	red (p = 0.3)	blue (p = 0.5)
color of boat on Monday	green (p = 0.2)			
	red (p = 0.3)			
	blue (p = 0.5)			

Here's a little exercise that involves thinking about how the probability of a conjunction is related to the probability of its conjuncts. It involves the example about sailboats mentioned in the previous chapter in the section on Copernicus and Ptolemy. My friend Susan saw a red sailboat on Lake Mendota on Monday, and on Tuesday she also saw a red sailboat. In the accompanying table I've listed probabilities for some sailboat colors on each of the two days. Note that the three probabilities for each day sum to one; I'm assuming that sailboats on the lake have no chance of being yellow. These probabilities are called *marginal probabilities* because they are written along the margins of the table. Now consider these hypotheses:

(ONE) Susan saw the same boat on both days.

(TWO) Susan saw one boat on Monday and a different boat on Tuesday.

The cells in the table represent conjunctions. For example, the cell in the upper right-hand corner represents the possibility that the sailboat seen on the first day is green *and* the one sighted on the second is blue. What probabilities does the TWO hypotheses dictate for the cells? What cell entries does ONE say are correct? Assume in both cases that sailboats don't change color from day to day. How does the concept of probabilistic dependence apply to what the two hypotheses say?

The truth value of a conjunction $H\&J$ is determined by the truth value of H and the truth value of J. The conjunction is true if H is true and J is true, and it is false otherwise. This is what logicians mean when they say that conjunction is a "truth-functional operator." We have just seen that the probability of the conjunction $H\&J$ isn't settled by the probability of H and the probability of J. If anything, it is the probabilities of conjunctions that settle the probability of a conjunct. Here I have in mind a fact I mentioned earlier, the theorem of total probability, which says that $\Pr(H) = \Pr(H\&J) + \Pr(H\&notJ)$.

Another concept that will be useful in what follows is *mathematical expectation*. You've encountered this before when you've heard discussion of the "life expectancy" of a baby born this year. As a first pass, this quantity can be understood as an average. If you say that the life expectancy for a baby born this year in the United States is 80 years, this means that 80 years will be the average lifespan of the individuals born this year. Let's get more precise by talking about probabilities and coin tosses. If you toss a fair coin ten times, there are eleven possible outcomes (0 heads, 1 head, 2 heads, . . . , 10 heads) and each of these has its own probability. The expected number of heads is defined as follows:

$$\begin{aligned}&\text{Expected}_{\text{A}}(\text{number of heads})\\&\quad = (0)\text{Pr}_{\text{A}}(\text{exactly 0 heads}) + (1)\text{Pr}_{\text{A}}(\text{exactly 1 head})\\&\qquad + (2)\text{Pr}_{\text{A}}(\text{exactly 2 heads}) + \cdots + (10)\text{Pr}_{\text{A}}(\text{exactly 10 heads})\\&\quad = \sum_{i=0}^{10} (i)\text{Pr}_{\text{A}}(\text{exactly } i \text{ heads}).\end{aligned}$$

Here A is the assumption that the coin is fair and you toss the coin ten times. It turns out that the expected value is 5. As you do this ten-toss experiment again and again, you can be more and more certain that the average number of heads across the different ten-toss repetitions is close to 5. This is the law of large numbers I mentioned earlier.

The expected number is often not the number you should expect. If you toss a fair coin three times, the expected number of heads is 1.5, but this doesn't mean that you should expect there to be 1.5 heads when you perform this experiment just once. In the experiment I described five paragraphs ago concerning a world in which all coins either have two heads or two tails, what is the expected frequency of heads if you toss a randomly chosen coin ten times? What is the frequency you should expect?

Although the axioms of probability that I have described always involve a relation between the assumptions that define the probability function and this or that proposition, I have yet to define the idea of "conditional probability." I have been talking about $\mathrm{Pr}_A(H)$, not about $\mathrm{Pr}_A(H \mid E)$. The latter is read as "the probability of H given E." Take care to understand what this means. It doesn't mean that E is true and that H therefore has a certain probability. Just as "if you toss the coin then it will land heads" does not assert that you toss the coin, so "Pr_A(the coin lands heads | you toss the coin) $= \frac{1}{2}$" does not say that you actually toss the coin. What it means is this: suppose for the moment that you have tossed the coin. You then are asked how probable it is that the coin will land heads, given that supposition. The value of the conditional probability is the answer to this question.

The concept of conditional probability can be introduced by saying how it is related to the notion of unconditional probability that is defined by our axioms:

$$\mathrm{Pr}_A(H \mid E) = \frac{\mathrm{Pr}_A(H \& E)}{\mathrm{Pr}_A(E)} \text{ if } \mathrm{Pr}_A(E) > 0.$$

This is called the *ratio formula*. If A says that E has a probability of zero, this "definition" of conditional probability offers no advice on what conditional probability means. I put "definition" in scare quotes because a (full) definition should provide necessary and sufficient conditions; the above statement provides only the latter. Some think that the conditional probability $\mathrm{Pr}_A(H \mid E)$ has no meaning when $\mathrm{Pr}_A(E) = 0$. I disagree. A coin can be fair even if you lock it in an impregnable safe so that the coin can never be tossed. Here Pr_A(the coin lands heads | you toss the coin) $= \frac{1}{2}$ even though Pr_A(you toss the coin) $=$ 0 (Rényi 1970; Hájek 2003; Sober 2008b). There is a second qualification that needs to be registered in connection with the ratio formula, which I'll discuss later. But for now it's worth noting that if $\mathrm{Pr}_A(H \mid E)$, $\mathrm{Pr}_A(H \& E)$, and $\mathrm{Pr}_A(E)$ all have values and $\mathrm{Pr}_A(E) > 0$, then the ratio formula must hold.

To illustrate the idea of conditional probability, let's return to the example of the deck of cards. As before, I'll assume that the deck is standard and that you are dealt cards at random. What is the value of Pr_A(the card you were just dealt is a heart | the card you were just dealt is red)? On the supposition that the card is red, the probability of its being a heart is $\frac{1}{2}$. The ratio formula delivers this result. Here's the argument:

$$\begin{array}{l} \text{Pr(the card is a heart and the card is red)} = \tfrac{1}{4}. \\ \text{Pr(the card is red)} = \tfrac{1}{2}. \\ \hline \text{Therefore, Pr(the card is Heart} \mid \text{the card is red)} = \tfrac{1}{2}. \end{array}$$

As an exercise, I suggest that you draw a Venn diagram of this example. You need to have an area of the diagram representing "the card is red" and another area representing "the card is a heart." And of course you need to consider the intersection of these two areas, which represents the conjunction of the two propositions. When you consider the conditional probability, you focus on the area of the diagram in which the card is red and determine what proportion of that area is occupied by the card's being a heart. Notice how hard this would be if the card had a probability of zero of being red!

I have used the word "assumption" to describe probability functions and the word "supposition" to describe conditional probabilities. These two terms may sound like synonyms but I am using them to pick out different things. In the example just described, I assumed that the deck is standard and that cards are dealt at random. I did not assign probabilities to those assumptions. With those assumptions in place, I asked you to consider the conditional probability Pr_A(the card you were just dealt is a heart | the card you were just dealt is red), which requires you to consider the supposition that the card is red. Assumptions define probability functions whereas suppositions come up within a given probability function when a conditional probability is being evaluated. We often use models that we believe are true and we often entertain suppositions that we think are false. I believe that the deck is standard and that the cards are dealt at random. In contrast, I do not believe that the card you were just dealt is red, though I wish to entertain that supposition in evaluating a conditional probability. Now that I have separated assumptions from suppositions, let me bring them back together. There is a numerical identity that connects the assumptions that a probability model makes and the suppositions that one entertains within a model. It is this:

$$\mathrm{Pr}_{A\&B}(H) = \mathrm{Pr}_A(H \mid B).$$

The values are the same, but the epistemic status of B is subtly different.

The "definition" of conditional probability makes it clear why $\mathrm{Pr}_\mathrm{A}(H\&J) \leq \mathrm{Pr}_\mathrm{A}(J)$, no matter what A is. Assuming that $\mathrm{Pr}_\mathrm{A}(J) > 0$, the inequality can be rewritten as $\mathrm{Pr}_\mathrm{A}(H \mid J)\mathrm{Pr}_\mathrm{A}(J) \leq \mathrm{Pr}_\mathrm{A}(J)$. Since probabilities are numbers between 0 and 1, the product of two probabilities cannot be greater than the value of either of them. This fact is relevant to the razor of silence that I discussed in the previous chapter. If you consider a conjunction $H\&J$ and slice away H (not by denying that H is true, but simply by declining to assert or deny it), the probability of what remains (J) cannot be less than the probability of the conjunction with which you began. In fact, if $\mathrm{Pr}_\mathrm{A}(H)$ and $\mathrm{Pr}_\mathrm{A}(J)$ are both positive and $\mathrm{Pr}_\mathrm{A}(H \mid J)$ is less than 1, the slicing away will increase probability. Silence reduces your risk of error. The razor of silence has a simple Bayesian rationale.

The fact that a conjunction can't be more probable than its conjuncts is anything but obvious to many people. In a much-cited psychology experiment, Tversky and Kahneman (1982) told their subjects the following story:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

The subjects then were asked which of the following statements is more probable:

- Linda is a bank teller.
- Linda is a bank teller and is active in the feminist movement.

Well over half of the subjects in the experiment said that the second statement is more probable than the first. This example is a warning: when you use the mathematical concept of probability, don't stumble into the mistake of committing "the conjunction fallacy"!

We now can use the ratio formula to derive Bayes's theorem. To simplify notation, I'll drop the subscript "A," but don't forget that a probability function is always built on a set of assumptions. Whenever I talk about a conditional probability $\Pr(X \mid Y)$, I'll assume that $\Pr(Y) > 0$. So let's start by describing each of $\Pr(H \mid E)$ and $\Pr(E \mid H)$ in terms of ratios of unconditional probabilities:

$$\Pr(H \mid E) = \frac{\Pr(H \& E)}{\Pr(E)} \qquad \Pr(E \mid H) = \frac{\Pr(E \& H)}{\Pr(H)}$$

These two equations can be rearranged to yield:

$$\Pr(H \& E) = \Pr(H \mid E)\Pr(E) \qquad \Pr(E \& H) = \Pr(E \mid H)\Pr(H)$$

The left-hand sides of these two equations are equal, since *H&E* is logically equivalent to *E&H*, which means that the right-hand sides must be equal to each other. Setting the right-hand sides of these equations equal and performing a little algebra yields Bayes's theorem:

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

Although the derivation of Bayes's theorem works for any propositions *H* and *E* you please, the typical application involves *H*'s being a "hypothesis" and *E*'s being "observational evidence." Bayes's theorem shows how the conditional probability $\Pr(H \mid E)$ can be computed from three other quantities. Notice that one of them is the unconditional probability $\Pr(H)$.

At the start of this chapter, I said that Bayesianism and frequentism are different philosophies of scientific inference. Does accepting Bayes's theorem place you knee deep in the former philosophy? Not so! The theorem is a mathematical truth – it follows from the axioms of probability and the "definition" of conditional probability. Frequentists do not challenge the correctness of this derivation. Rather, they challenge its *usefulness*. Frequentists think that it often isn't possible to think about $\Pr(E \mid H)$, $\Pr(H)$, and $\Pr(E)$ as objective quantities. However, they agree that *if* these three probabilities had objective values, the value of $\Pr(H \mid E)$ could be calculated by using Bayes's theorem. They also agree that if lions could fly, then zebras would need to watch out for aerial lion attacks.

In thinking about the probabilities that figure in Bayes's theorem, it is important to recognize that $\Pr(H \mid E)$ and $\Pr(E \mid H)$ are different quantities and therefore may have different values. Much heartache will be avoided by attending to this difference! In logic it is a familiar idea that a conditional and its converse are different, and that one can be true while the other is false. For example, consider

- If noisy gremlins are bowling in your attic, then you hear noise coming from your attic.
- If you hear noise coming from your attic, then noisy gremlins are bowling in your attic.

It is obvious that the first can be true while the second is false. In just the same way the following two conditional probabilities can have different values:

- Pr(you hear noise coming from your attic | noisy gremlins are bowling in your attic)
- Pr(noisy gremlins are bowling in your attic | you hear noise coming from your attic)

Personally, I think that the first probability has a high value and the second has a low one.

The quantity Pr(E) on the right-hand side of Bayes's theorem deserves a comment. Pr(E) is the unconditional probability of the evidence E. In our gremlin example, E is the proposition that you hear noises coming from your attic. You might think that Pr(E) should be high if there frequently are noises coming from up there and that it should be low if such noises are rare. I agree that frequencies often provide *evidence* that is relevant to estimating the value of Pr(E). But, as noted earlier, I don't want to *define* probability as frequency. So what does the unconditional probability of E mean? The theorem of total probability tells us that

$$\mathrm{Pr}(E) = \mathrm{Pr}(E\&H) + \mathrm{Pr}(E\,\&notH).$$

Using the "definition" of conditional probability (and assuming that all the relevant probabilities are positive), we can rewrite this as

$$\mathrm{Pr}(E) = \mathrm{Pr}(E \mid H)\mathrm{Pr}(H) + \mathrm{Pr}(E \mid notH)\mathrm{Pr}(notH).$$

This characterization of Pr(E) shows that the value of this quantity will sometimes be very different from the frequency with which E is true. The example of the world in which half the coins have two heads and half have two tails provides an example. You choose a coin at random and toss it repeatedly. Use the above equality to convince yourself that Pr(Heads) = 0.5. Yet, when you do the experiment, you obtain either 100 percent heads or 100 percent tails.[4]

When I explained earlier what probabilistic independence means, I did so by describing a relation among unconditional probabilities. Now that the idea of conditional probability has been introduced, I can define the idea of *conditional* independence. It parallels the unconditional concept already explained:

> X and Y are probabilistically independent of each other conditional on C in probability function $\mathrm{Pr_A}(-)$ if and only if $\mathrm{Pr_A}(X\&Y \mid C) = \mathrm{Pr_A}(X \mid C)\mathrm{Pr_A}(Y \mid C)$.

Here's an example from Mendelian genetics: the genotypes of two full siblings are independent of each other, conditional on the genotypes of their parents. For example,

> $\mathrm{Pr_M}$(sib 1 has *AA* & sib 2 has *AA* | mom has *AA* & dad has *Aa*) = $\mathrm{Pr_M}$(sib 1 has *AA* | mom has *AA* & dad has *Aa*) $\mathrm{Pr_M}$(sib 2 has *AA* | mom has *AA* & dad has *Aa*).

The "M" subscript on the probability function indicates that probabilities are assigned on the basis of the usual Mendelian model of inheritance. The probability on the left has a value of ¼ and the two probabilities on the right each have a value of ½. Notice that the above equality holds more generally; it holds for *any* genotypes that the two siblings, mom, and dad might have:

> For any genotypes G_1, G_2, G_3, and G_4, $\mathrm{Pr_M}$(sib 1 has G_1 & sib 2 has G_2 | mom has G_3 & dad has G_4) = $\mathrm{Pr_M}$(sib 1 has G_1 | mom has G_3 & dad has G_4) $\mathrm{Pr_M}$(sib 2 has G_2 | mom has G_3 & dad has G_4).

When this more general relation obtains, the parental genotype is said to *screen-off* each offspring genotype from the other. We can generalize from this genetics example and define screening-off as a relation that might obtain among any three variables X, Y, and Z:

[4] In this example, the unconditional probability of the evidence involves two possibilities – either H is true or *notH* is. But suppose there are n possible hypotheses $H_1, H_2, \ldots, H_n$, which are mutually exclusive and collectively exhaustive. What would $\Pr(E)$ mean in that case? The answer is a generalization of what I just said for the dichotomous case:

$$\begin{aligned}\Pr(E) &= \Pr(E \mid H_1)\Pr(H_1) + \Pr(E \mid H_2)\Pr(H_2) + \cdots + \Pr(E \mid H_n)\Pr(H_n)\\ &= \textstyle\sum_{i=1}^{n} \Pr(E \mid H_i)\Pr(H_i).\end{aligned}$$

$$Z \text{ screens-off } X \text{ from } Y \text{ precisely when, for any values } i, j, k,$$
$$\Pr(X = i \,\&\, Y = j \mid Z = k) = \Pr(X = i \mid Z = k)\Pr(Y = j \mid Z = k).$$

Here I'm using a notation that is standard in probability theory; "$X = i$" means that the variable X has the value i. As the genetics example suggests, Z screens-off Y from X precisely when Z screens-off X from Y. There is another, equivalent, definition of screening-off that you should know about. It says that Z screens-off Y from X if and only if $\Pr(X = i \mid Z = k) = \Pr(X \mid Z = k \,\&\, Y = j)$, for all i, j, k. You can prove the equivalence of these two ways of describing screening-off by using the "definition" of conditional probability (and assuming that all the conditioning propositions have positive probabilities). Just as parental genotype screens-off offspring genotypes from each other, it's also true that parental genotype screens-off grandparental genotype from offspring genotype. Screening-off often applies when a common cause has two (or more) effects, and it often comes up when you talk about a causal chain from Y to Z to X. "Often" does not mean *always*. Mom's genotype is a common cause of the genotypes of her two offspring. However, her genotype does not screen-off each from the other. See if you can figure out why this is so. And see if you can think of an example of a causal chain in which the proximate cause doesn't screen-off the distal cause from the effect.

Screening-off can be described informally in informational terms. If you know the parental genotype, the probability you assign to one offspring's genotype shouldn't be affected by learning the genotype of the other. And if you know the parental genotype, the probability you assign to an offspring's genotype shouldn't be affected by your learning the genotypes of the grandparents.

Conditional independence and unconditional dependence may sound like they are incompatible, but in fact they are not. Once again, genetics furnishes an example. As noted, the two offspring genotypes are independent of each other, conditional on the parental genotype. However, the fact that the two siblings have the same parents means that their genotypes will be unconditionally dependent:

$$\Pr(\text{sib 1 has } AA \,\&\, \text{sib 2 has } AA) > \Pr(\text{sib 1 has } AA)\Pr(\text{sib 2 has } AA).$$

Notice that this inequality makes no mention of what the parental genotype is. If you conditionalize on the parental genotype, the inequality turns into an equality! Not only are unconditional dependence and conditional independence not in conflict; I'll explain later in this chapter how conditional independence can be part of the explanation of unconditional dependence.

Translating the last displayed inequality into the language of conditional probability, we get both of the following:

$$\Pr(\text{sib 1 has } AA \mid \text{sib 2 has } AA) > \Pr(\text{sib 1 has } AA)$$
$$\Pr(\text{sib 2 has } AA \mid \text{sib 1 has } AA) > \Pr(\text{sib 2 has } AA).$$

This is what it means for the two genotypes to be *correlated*. Bayesians take these two inequalities to have an important epistemological significance. They gloss these inequalities by saying that the *AA* genotype of each sibling provides *confirmation* that the other sibling has the *AA* genotype. Bayesians define confirmation as follows:

Observation E confirms hypothesis H if and only if $\Pr(H \mid E) > \Pr(H)$.[5]

Disconfirmation gets defined in tandem:

Observation E disconfirms hypothesis H if and only if $\Pr(H \mid E) < \Pr(H)$.

If confirmation means probability raising and disconfirmation means probability lowering, then *evidential irrelevance* means that the observation leaves the probability of the hypothesis unchanged.[6] The Bayesian ideas of confirmation and disconfirmation entail that there is a symmetry between confirmation and disconfirmation:

E confirms H if and only if *notE* disconfirms H.

Convince yourself that this biconditional is correct when confirmation and disconfirmation are given Bayesian interpretations. And then convince

[5] Following Carnap (1950), Bayesians sometimes contrast the "incremental" concept of confirmation just described with one that is "absolute." The idea is that E absolutely confirms H precisely when $\Pr(H \mid E)$ is high. Notice that this can be true when E incrementally disconfirms H or is evidentially irrelevant to it in the incremental sense. I think it is unfortunate that the word "confirm" is used to denote a high value for $\Pr(H \mid E)$. I won't do so in what follows.

[6] In view of the fact that assigning a value to $\Pr(H \mid E)$ does not require that E be true, it is better to read the Bayesian definition of confirmation as explicating the following proposition: E, if true, would confirm H. A parallel point holds for disconfirmation.

yourself, by using Bayes's theorem, that confirmation is a symmetrical relation – if X confirms Y, then Y confirms X.[7]

The Bayesian definition of confirmation can be used to underscore my earlier point that $\Pr(E)$ should not be defined as the frequency with which E is true. Suppose Susan takes a tuberculosis test several times and it comes out positive every time. This might lead you to think that $\Pr(E) = 1$, where E says that Susan's test outcome is positive. However, if $\Pr(E) = 1$, E cannot confirm the hypothesis T, which says that Susan has tuberculosis. This entailment can be verified by consulting Bayes's theorem. To get things right, you need to see that $\Pr(E)$ is an average whose value is described by the theorem of total probability:

$$\Pr(E) = \Pr(E \mid T)\Pr(T) + \Pr(E \mid notT)\Pr(notT).$$

Doing so allows you to see that $\Pr(E)$ is less than one, which means that E can confirm H.

The next distinctively Bayesian idea I need to describe concerns how agents should change their probability assignments as new evidence rolls in. All the probabilities described in Bayes's theorem use the same probability function $\Pr_{\mathrm{A}}(-)$. The assumptions in A can be thought of as the assumptions that an agent is prepared to make at a given time. Suppose the agent learns (with certainty) that a proposition N is true, where N isn't something she already believed; N is news to her. Her set of assumptions has thereby been augmented. We need a rule that describes how the probabilities she assigned under her earlier probability function $\Pr_{\mathrm{A}}(-)$ are related to the probabilities she should assign under her later probability function $\Pr_{A\&N}(-)$. A rule that describes this relationship is called an updating rule.

Before you are dealt a card from the standard deck of cards that I keep talking about, you assign $\Pr_{\mathrm{A}}$(the card will be the Ace of hearts | the card is red) $=\frac{1}{26}$. Suppose you then learn that the card is red. Call this new piece of information N; N gets added to what you already assumed, namely A. So

[7] Bayesians have proposed different measures of *degree of confirmation*. These agree with the Bayesian definition of confirmation just described, but go beyond it, in that they assign numbers to represent how much the evidence confirms the hypothesis. These measures disagree with each other in that they are *ordinally non-equivalent* (Fitelson 1999). This means that there are Bayesian measures X and Y of degree of confirmation that have this property: $X(H_1, E) > X(H_2, E)$ while $Y(H_1, E) \leq Y(H_2, E)$.

what value should you assign to $\mathrm{Pr}_{\mathrm{A\&N}}$(the card will be the Ace of Hearts)? The *rule of updating by strict conditionalization* says that your new unconditional probability should be $\frac{1}{26}$. More generally, the idea is this:

> *The Rule of Updating by Strict Conditionalization:* $\mathrm{Pr}_{t2}(H) = \mathrm{Pr}_{t1}(H \mid N)$ if the totality of what you learned between t_1 and t_2 is that N is true.

This updating rule has two major limitations. First, it characterizes learning as the discovery that some proposition N is true. However, if I tell you that N has a probability of, say, 0.6, that new information isn't something that the machinery of strict conditionalization tells you how to take into account.[8] Second, strict conditionalization describes how you should change your assignments of probability when you add a proposition to your assumptions, but it doesn't tell you what to do if something you previously assumed turns out to be false. The rule of strict conditionalization represents learning as gaining certainties, where a certainty, once gained, can never be lost.[9]

The simple updating rule just described allows me to explain some standard vocabulary that is used in connection with Bayes's theorem. I have described $\mathrm{Pr}_{\mathrm{A}}(H \mid E)$ and $\mathrm{Pr}_{\mathrm{A}}(H)$ as the conditional and the unconditional probability of H, but it is customary to describe the former as H's *posterior* probability and the latter as H's *prior* probability. This temporal terminology is a bit misleading; it suggests that $\mathrm{Pr}_{\mathrm{A}}(H \mid E)$ is a probability assignment made later (after you've learned that E is true) whereas $\mathrm{Pr}_{\mathrm{A}}(H)$ is a probability assignment made earlier (before you learn that E is true). In fact, the A subscript shows that both these probability assignments hold true under a single set of assumptions – the assumptions that an agent makes at a given time. And remember that you don't need to think that E is true to consider the value of $\mathrm{Pr}_{\mathrm{A}}(H \mid E)$! What is true is that the old conditional probability $\mathrm{Pr}_{\mathrm{A}}(H \mid E)$ is where the new unconditional probability $\mathrm{Pr}_{A\&E}(H)$ comes from (under the rule of strict conditionalization) when you learn that E is true. The old conditional probability gives rise to a new unconditional probability. Don't let the temporal labels "posterior" and "prior" confuse you. $\mathrm{Pr}_{\mathrm{A}}(H \mid E)$ is the posterior probability of H in the sense that its value is the same as the value of $\mathrm{Pr}_{A\&E}(H)$, once you learn that E is true. Notice that the former

[8] Jeffrey (1965) develops a theory of updating for this more general notion of learning.

[9] Titelbaum (2013) develops a Bayesian model for losing certainties.

probability doesn't involve the assumption that E is true, but the latter one does.

Bayes's theorem allows you to compute how gaining new evidence E should lead you to change your degree of confidence in the hypothesis H. The posterior probability may have a different value from the prior. However, there are two cases in which no such change is possible. If $\Pr(H) = 1$ (or 0), then $\Pr(H \mid E) = 1$ (or 0), no matter what E is. The two extreme probability values (0 and 1) are "sticky." This is why Bayesians advise you to be extremely circumspect about assigning a hypothesis a prior of 0 or 1. In doing so, you are saying that no future experience could make it reasonable for you to change how confident you are in H.

One more bit of terminology can now be introduced. I used the gremlin example to illustrate the difference between $\Pr(H \mid E)$ and $\Pr(E \mid H)$. We are now calling the first of these H's posterior probability. The second also has a name – it is called H's *likelihood*. This terminology, due to R. A. Fisher, is unfortunate. In ordinary English, talking about the probability of H and the likelihood of H are two ways of saying the same thing. In the technical parlance that is now canonical, the two come apart. To avoid confusing them, keep gremlins firmly in mind; when you hear the noise in your attic, the gremlin hypothesis has a high likelihood but a low probability. In what follows, when I say "likelihood," I will be using the term's technical meaning.

I now can complete my sketch of Bayesianism by describing an important consequence of Bayes's theorem. Suppose hypotheses H_1 and H_2 are competing hypotheses. We have some observational evidence E and we want to know which of these hypotheses has the higher posterior probability. If you write Bayes's theorem for each of these hypotheses (please do so!), you can derive the following equation, which is called *the odds formulation of Bayes's theorem*:

$$\frac{\Pr(H_1 \mid E)}{\Pr(H_2 \mid E)} = \frac{\Pr(E \mid H_1)}{\Pr(E \mid H_2)} \frac{\Pr(H_1)}{\Pr(H_2)}.$$[10]

This says that the ratio of posterior probabilities equals the likelihood ratio multiplied by the ratio of prior probabilities. Notice that the unconditional probability of the observations, $\Pr(E)$, has dropped out. Notice also that this version of Bayes's theorem says that there is exactly one way that an observation E can lead you to change how confident you are in H_1 as compared

[10] "Odds" is a word from gambling; it refers to the ratio of posterior probabilities. If this ratio is, say, 20-to-1, that means that H_1 is twenty times as probable as H_2.

with H_2. If the ratio of posteriors is to differ from the ratio of priors, this must be because the likelihoods differ. And the more the likelihood ratio departs from 1, the more the ratio of posterior probabilities departs from the ratio of priors.

The odds formulation of Bayes's theorem makes it easy to see how a hypothesis with a very low prior probability can have its probability driven above 0.5 by several favorable observations, even when one such observation is not enough to push the hypothesis over the top. Consider Susan and her positive tuberculosis test. Suppose the prior probability of Susan's having tuberculosis is 0.001. She then takes a tuberculosis test that has the following property:

$$\Pr(\text{positive test outcome} \mid \text{Susan has tuberculosis}) = 0.96$$

$$\Pr(\text{positive test outcome} \mid \text{Susan does not have tuberculosis}) = 0.02$$

The odds formulation of Bayes's theorem allows you to compute the ratio of posterior probabilities from the numbers we have at hand. The likelihood ratio is 48. The ratio of the priors is $\frac{1}{999}$. So the ratio of the posterior probabilities is $\frac{48}{999}$. This last number means that the posterior probability of Susan's having tuberculosis is $\frac{48}{999+48}$. This is way less than $\frac{1}{2}$, but it is bigger than $\frac{1}{1000}$. The positive test result has increased Susan's probability of having tuberculosis, but not by a whole lot. Now suppose that Susan takes the test a second time and again gets a positive result. Since the two test results are independent of each other, conditional on each of the two hypotheses, the odds formulation of Bayes's theorem will take the following form:

$$\frac{\Pr(H_1 \mid E_1 \& E_2)}{\Pr(H_2 \mid E_1 \& E_2)} = \frac{\Pr(E_1 \mid H_1)}{\Pr(E_1 \mid H_2)} \frac{\Pr(E_2 \mid H_1)}{\Pr(E_2 \mid H_2)} \frac{\Pr(H_1)}{\Pr(H_2)}.$$

The product of the two likelihood ratios is (48)(48) = 2304. Given the ratio of the priors, the ratio of the posterior probabilities is now $\frac{2304}{999}$, so the probability of tuberculosis is now $\frac{2304}{999+2304}$, which is about 0.69. The *single* positive test outcome doesn't entail that Susan probably has the disease, but the *two* positive outcomes together have that implication. People often think that if they take a reliable tuberculosis test and get a positive outcome, then they probably have tuberculosis. Kahneman and Tversky (1985) call this the *base rate fallacy*; the mistake is the failure to take account of the prior probability of tuberculosis.

When I introduced the odds version of Bayes's theorem, I mentioned that the likelihood ratio represents the sole vehicle in the Bayesian framework whereby new evidence can modify your relative confidence in competing hypotheses. It will be useful to have a principle that isolates this unique role. Hacking (1965) calls the following *the law of likelihood*:

> Evidence E favors hypothesis H_1 over hypothesis H_2 if and only if $\Pr(E \mid H_1) > \Pr(E \mid H_2)$.

When the evidence favors H_1 over H_2 in this sense, the ratio of posterior probabilities exceeds the ratio of priors.

The law of likelihood isn't a deductive consequence of the odds formulation of Bayes's theorem. The theorem is a mathematical fact, but the law is not a truth of mathematics; *favoring* isn't a concept that gets used in the axioms of probability. So perhaps we should regard the law as a proposed explication of the ordinary language concept of favoring. If we do so, we must conclude that the law is flawed. Suppose a talented weather forecaster looks at today's weather conditions and concludes that there probably will be snow tomorrow. The forecaster might summarize this finding by saying that the present weather conditions favor snow tomorrow over no snow tomorrow. Here the word "favoring" is being used to describe an inequality between *probabilities*, not a *likelihood* inequality; what is being claimed is that Pr(snow tomorrow | today's weather conditions) > Pr(no snow tomorrow | today's weather conditions). So if the law of likelihood is an explication of the word "favoring," it is flawed (Sober 2008b). An alternative interpretation of the law of likelihood is better. We can regard the law as a stipulation; the term "favoring" is being used to mark the fact that likelihood inequalities have a special epistemic significance, with no pretense that the law captures every proper use of the word "favoring" in ordinary English.[11] It is not for nothing that Bayesians have come to call the likelihood ratio "the Bayes factor."[12]

It is a consequence of the law of likelihood that the evidence at hand may favor an implausible hypothesis over a sensible one. When you observe

[11] Stipulations are often said to be "arbitrary." But within the Bayesian framework, there is nothing arbitrary about the claim that the likelihood ratio plays a special epistemic role. What is arbitrary is using the word "favoring" to name that role. This, by the way, reveals one limitation of the idea that philosophy's sole aim is to explicate concepts that already have names in ordinary language.

[12] Fitelson (2011) argues that Bayesians should reject the law of likelihood; I reply in Sober (2011d).

that the card you are dealt is an ace, the law says that this observation favors the hypothesis that the deck is made entirely of aces over the hypothesis that the deck is normal (since $1 > 4/52$). This may sound like an objection to the law, but there is a reply. Your doubts about the first hypothesis stem from information you had before you observed the ace, not from what you just observed (Edwards 1972). Likelihood comparisons are supposed to isolate what the evidence says, not to settle which hypotheses are more probable than which others.

Another feature of the law of likelihood is that it says that an observation can favor one hypothesis over another even when neither hypothesis predicts the observation. Suppose $\Pr(E \mid H_1) = 0.001$ and $\Pr(E \mid H_2) = 0.000001$. Neither hypothesis "predicts" E in the sense of saying that E is more probable than not, but the fact remains that E discriminates between the two hypotheses. Asking what a hypothesis "predicts" is a highly imperfect guide to interpreting evidence.

To keep things simple, I have treated the law of likelihood as a Bayesian idea, and I have talked about two philosophies of scientific inference – Bayesianism and frequentism. In fact there is a third camp; there are non-frequentists who are critical of Bayesianism (Edwards 1972; Royall 1997). These "likelihoodists" think that the law of likelihood stands on its own; they think that its justification does not depend on the role that likelihoods play in the odds formulation of Bayes's theorem. The motivation for likelihoodism is illustrated by the following example. When Arthur Stanley Eddington observed the bending of light during a solar eclipse, this was widely regarded as strong evidence favoring Einstein's general theory of relativity over the classical physics of Newton. Likelihoodists represent this in terms of the relationship between two likelihoods:

$$\begin{aligned}&\Pr(\text{Eddington's data on the solar eclipse} \mid \text{general relativity theory})\\&\quad > \Pr(\text{Eddington's data on the solar eclipse} \mid \text{classical mechanics}).\end{aligned}$$

You don't need to think about the prior probability of either theory to see that this inequality is true.[13]

[13] You can see here why likelihoodists don't like the "definition" of conditional probability as a ratio of unconditional probabilities. Likelihoodists want likelihoods to "make sense" even when priors do not.

Although likelihoodists aren't Bayesians, there is a formal connection between the law of likelihood and Bayesian confirmation theory:

$$\Pr(E \mid H) > \Pr(E \mid notH) \text{ if and only if } \Pr(H \mid E) > \Pr(H).$$

E favors *H* over *notH* (in the sense of the law of likelihood) precisely when *E* confirms *H* (in the sense of Bayesianism). I suggest that you prove this biconditional. Doesn't this formal connection of the two *ism*'s force likelihoodists to admit that they are Bayesians under the skin? Not really. Besides eschewing prior probabilities, likelihoodists think that assigning a value to $\Pr(E \mid notH)$ often lacks an objective justification. It is clear enough what the probability was of Eddington's observations of the solar eclipse, given general relativity. However, the probability of those observations, given the negation of general relativity, is more opaque. The negation of general relativity is a vast disjunction, covering all possible alternatives to general relativity, even ones that have not yet been formulated. The likelihood of *notGTR* therefore takes the following form:

$$\Pr(O \mid notGTR) = \sum_i \Pr(O \mid A_i)\Pr(A_i \mid notGTR).$$

The likelihood of *notGTR* is a weighted average of the likelihoods of all the alternatives (the A_i's) to *GTR*; to compute this average, you need to know how probable each A_i is, given *notGTR*. The negation of the general theory of relativity is an example of what philosophers of science call a "catchall hypothesis." Likelihoodists restrict their epistemology to "specific" theories – to general relativity and Newtonian mechanics, for example. So there are *two* reasons why likelihoodists aren't Bayesians: they don't want to talk about the prior and posterior probabilities of theories, *and* they don't want to talk about the likelihoods of catchalls (Sober 2008b).[14]

[14] Can the objection to Bayesianism that focuses on its need for prior probabilities be dealt with by appealing to various theorems concerning "the washing out of priors"? The idea here is that agents who start with very different prior probabilities and then interpret the evidence in the same way (because they agree on the values of the likelihoods) will end up agreeing on the posterior probabilities; their different starting points don't matter in the long run. The mathematical arguments being appealed to here are correct, but the problem is that they are asymptotic. When agents who have different priors confront a finite data set, they will disagree about the posteriors, often dramatically; what would happen in the infinite long run doesn't change that point. Think about Susan's single tuberculosis test.

It may be helpful to think of the difference between Bayesianism and likelihoodism in terms of the distinction between *private* and *public*. Bayesianism is a philosophy for individual agents who each want to decide how confident they should be in various hypotheses. Likelihoodism is an epistemology for the public world of science; it aims to isolate something objective on which agents can agree despite the fact that they differ in terms of their prior degrees of confidence in the hypotheses under consideration. Agents need prior and posterior probabilities to live their lives, but science needs something that in an important way transcends individual differences. This suggestion does not deny that scientists are agents.

Bayesianism comes in many forms, but to organize ideas let's lump these variants together under a single banner: *computing the posterior probabilities of hypotheses is always an attainable goal*. Likelihoodists claim that this is often impossible to achieve. When Bayesianism fails, likelihoodists hold that discovering which of several specific hypotheses the evidence favors is an attainable goal. Likelihoodism's goal is more modest than Bayesianism's. In a sense, likelihoodism is an *attenuated* Bayesianism; likelihoodism describes what remains of Bayesianism when some of it is stripped away. I have yet to describe what frequentism embraces as its attainable goal. In fact, I think there is no such thing; frequentism is too big a tent for that to be true. However, I have mentioned that frequentists have something negative in common. They want to use probabilistic tools in scientific inference without ever assigning probabilities to hypotheses. Later in this chapter I'll discuss one variety of frequentism and outline its goals and methods.

The present lay of the land is that most theorists about inference are *monists*; they sign up under a single *ism* and swear allegiance to it 100 percent of the time. I am inclined to be more pluralistic. I think that Bayesian, likelihoodist, and frequentist ideas all have their place. The attainable goals in scientific inference vary from problem to problem.

Ockham's razor for Bayesians

The odds formulation of Bayes's theorem has great significance for our investigation of Ockham's razor. For Bayesians, parsimony is not rock bottom; rather it is Bayes's theorem that is fundamental. This means that if Bayesians are going to show that a simpler theory S has a higher posterior probability than a theory C that is more complex, they must show that S has the higher likelihood or that it has the higher prior probability (or both). In saying this, I am not endorsing Bayesianism as a true and complete epistemology of science. Rather, I am stating an *if*: *if* you are a Bayesian and you think that simplicity is epistemically relevant, there are just two stories you can tell about why this is so. Bayesians of course have the option of scoffing at the relevance of parsimony, and some have done so.

Two kinds of prior probability

When Bayesians talk about prior probabilities, this often involves substantive empirical background assumptions. For example, when Susan takes a tuberculosis test and the test comes out positive, you may want to figure out how probable it is that she has tuberculosis, given this observation. Calculating this posterior probability requires that you assign a value to a prior probability. What probability should you assign to her having tuberculosis before you take account of her test outcome?

As already noted, what is a prior probability at one time is often identical in value to an earlier posterior probability. The prior probability at time t_2, Pr_{t2}(S has tuberculosis), will have the same value as the posterior probability at the earlier time t_1, Pr_{t1}(S has tuberculosis | S lives in Wisconsin) if the only relevant fact you learn between t_1 and t_2 is that Susan lives in Wisconsin. If you also know that the frequency of tuberculosis in the state this year is about 0.00001 (approximately 60 cases in a population of about 6000000), you may want to assign your prior at t_2 a value of 0.00001. This sort of prior probability is different from what Bayesians term a "first prior." The first prior of Susan's having tuberculosis must be based on no empirical evidence at all. Some Bayesians think that a proper theory of scientific inference requires that one assign first priors to hypotheses. Here the assumptions that constitute the probability function $\mathrm{Pr}_A(-)$ are merely the logical truths. Although it isn't weird to assign a prior probability to Susan's having tuberculosis on the assumption that Susan lives in a state where the frequency of tuberculosis is 0.00001, it is hard to understand how a prior probability can be assigned to this proposition on the assumption of tautologies alone. Yet, many Bayesians think this is necessary.

The traditional solution to this problem, which many Bayesians now reject, is to appeal to the *principle of indifference*. This principle says that if there are n exclusive and exhaustive propositions (called a *partition*), and you have no evidence that tells you which are more probable and which are less, then your first prior should be that each alternative has a probability of $\frac{1}{n}$. The trouble with this principle is that it leads to contradiction. In the case of Susan, if you think that there are just two possibilities (Susan has tuberculosis, Susan does not have tuberculosis), the principle of indifference assigns each a first prior of $\frac{1}{2}$. But if there are three possibilities (Susan has severe tuberculosis, Susan has tuberculosis that isn't severe, Susan has no tuberculosis), then the principle tells you to assign each a prior of $\frac{1}{3}$. The proposition that Susan does not have tuberculosis receives a prior of $\frac{1}{2}$ or $\frac{1}{3}$, depending on how you slice the cake. And, of course, there are many other partitions that can be considered as well. If the principle is somehow restricted by singling out a unique method of slicing up logical space, the principle is saved from contradiction, but it becomes arbitrary.

Most Bayesians now eschew the principle of indifference and so are reluctant to invoke the idea of first priors.[15] But what becomes of Bayesianism if first priors are abandoned? If calculating the posterior probability Pr(Susan has tuberculosis | Susan's tuberculosis test was positive) requires that you have a value for the prior probability Pr(Susan has tuberculosis), and if this prior is based on an earlier posterior, then that earlier posterior must itself involve a prior. We go back and back, but presumably not to infinity. Don't we need a starting point on which to base all subsequent calculations? And if that starting point can't be rationally defended, doesn't that mean that all the subsequent calculations on which it is based are irrational?

There is an alternative picture of Bayesianism that relieves it of the responsibility of defending first priors. The picture just described is *foundationalist*; it says that we need to defend a set of first priors on which all subsequent probabilities are to be built. The alternative is furnished by Otto Neurath's (1921) metaphor of the ship (made famous by Quine 1960). Neurath said that scientists are like sailors who must repair their ship while keeping it afloat. They never are able to put the boat in drydock; they can't disassemble the boat completely and then rebuild it from scratch on an ideal rational basis. The best the sailors can do is to try to improve the ship by replacing parts piecemeal. Viewed in this way, the task of Bayesianism is not to defend a set of first priors and then calculate what we should think about the world on that

[15] Maybe the word "most" is an overstatement, in view of the widespread Bayesian practice of talking about "informationless priors."

basis as new observations are made. Rather, the proper picture is that we now have numerous beliefs about the world. Our task is to take new observations into account so as to improve our system of beliefs. We don't start from zero; we start from where we are.

There is a second analogy that also illustrates this non-foundationalist version of Bayesianism. The goal of deductive logic is to tell you what you can and what you cannot deduce from a set of premises. Logic does not have the job of telling you what the premises are with which you should begin. Similarly, Bayesianism has the goal of telling you how you should adjust your degrees of beliefs as you obtain new evidence, given that you now have various degrees of confidence in different propositions. It tells you how to change where you are; it doesn't tell you where you should begin (Howson 2001).

As you can tell, I think that Neurath-style Bayesianism is more defensible than the foundationalist variety. But it is well to have both options in mind as we consider those Bayesian approaches to Ockham's razor that seek to connect simplicity with prior probability. We'll come back to non-first priors in a while. For now, let's take first things first.

Jeffreys's simplicity postulate

Harold Jeffreys (1891–1989) was an influential probabilist and geophysicist who played an important role in developing Bayesian approaches to scientific reasoning. Jeffreys thought that using Bayes's theorem requires first priors. By definition, first priors can't be justified by observations; Jeffreys believed in addition that they can't be justified *a priori*, either. His idea was to use simplicity to assign first priors to candidate hypotheses. In the preface to his 1931 book *Scientific Inference* (see also pp. 47–48), he says this:

> The chief guiding principle of both scientific and everyday knowledge [is] that it is possible to learn from experience and to make inferences from it beyond the data directly known by sensation . . . In the present work the principle is frankly adopted as a primitive postulate and its consequences are developed. It is found to lead to an explanation and a justification of the high probabilities attached in practice to simple quantitative laws.

The nature of this "leading" is something we need to examine.

Building on earlier work that he did with Dorothy Wrinch (see, for example, Wrinch and Jeffreys 1921), Jeffreys puts great weight on the fact that for any finite body of observations, there are infinitely many hypotheses that might, in principle, be the explanation of those observations. How are we to assign prior probabilities to these competing hypotheses? Jeffreys sees that the principle of indifference leads to a disastrous result. The disaster he has in mind is not the problem I mentioned before – that the principle delivers contradictory verdicts. Rather, his point is that if you have infinitely many pair-wise incompatible hypotheses and assign each of them the same prior probability, you must assign each a prior of zero.[16] But a probability of zero is sticky. If you embrace the postulate that it is possible to learn from experience, you need to assign these hypotheses *un*equal priors. Jeffreys's idea was to order these infinitely many hypotheses in terms of their simplicity and then assign them probabilities. For example, if you assign the simplest hypothesis a prior of ½, the next simplest hypothesis a prior of ¼, and the n^{th} simplest hypothesis a probability of $(½)^n$, each of the hypotheses in this infinite series receives a positive probability, and the sum of this infinite series is 1.

Jeffreys's formulations of his simplicity postulate went through a series of refinements. He thinks that counting the *adjustable parameters* in a hypothesis is relevant to assessing how complex it is, but he usually wanted other features of the hypothesis to count as well.[17] To understand what an adjustable parameter is, consider the following problem. You place a kettle on your stove and heat it to various temperatures and note how much pressure there is in the kettle each time. By doing this, you obtain several pairs of observations; each pair represents a temperature value and an associated pressure value. You can represent these pairs as points on Cartesian coordinates where the x-axis represents temperature and the y-axis represents pressure. You now confront an example of what philosophers call *the curve-fitting problem.* Which curve should you draw to represent the general relation of temperature and pressure in your kettle? Here are two hypotheses about the form that the true but unknown curve will take:

(LIN) $y = a_0 + a_1 x$

(PAR) $y = a_0 + a_1 x + a_2 x^2.$

[16] If you assign each of the members of an infinite set the same positive probability, the sum of these will exceed one, which violates the axioms of probability.

[17] Jeffreys (1931, p. 39; 1939, p. 47) says that the complexity of a law is the sum of the number of adjustable parameters and the absolute values of the integers that occur in it.

LIN says that the curve is a straight line, but it doesn't tell you which straight line is the true one. PAR says that the curve is a parabola, but it is silent on which parabola describes the relation of temperature and pressure in your kettle. The a terms in these hypotheses are the adjustable parameters. Statisticians call LIN and PAR "models" because they contain adjustable parameters. It is important to see that LIN and PAR differ from hypotheses that specify a unique straight line or parabola, such as the following:

(L) $y = 3 + 4x$
(P) $y = 2 + 5x + 6x^2$

In these two equations, the adjustable parameters in LIN and PAR have been replaced by specific values; the adjustable parameters, as it were, have been "adjusted." Notice that LIN has fewer adjustable parameters than PAR, but that the L and P have the same number of adjustable parameters, namely zero.

How are LIN and PAR related to specific straight lines and specific parabolas? LIN and PAR both make existence claims; they can be expressed more long-windedly as follows:

(LIN) There exist numbers a_0 and a_1 such that $y = a_0 + a_1x$.
(PAR) There exist numbers a_0, a_1, and a_2 such that $y = a_0 + a_1x + a_2x^2$.

LIN is a disjunction of the infinitely many straight lines in the x, y plane; PAR is an infinite disjunction of all the many parabolas. The fact that the a terms are bound to existential quantifiers shows that the parameters in a model are creatures of that model. The a_1 term in LIN refers to the slope of a straight line but the a_1 term in PAR does not. Although it isn't unusual to express models like LIN and PAR by using overlapping terminology, it would be just as correct (and maybe less misleading) to use a_0 and a_1 in LIN and a_2, a_3, and a_4 in PAR. The variables x and y are a different story; the two models uses these terms to refer to the same things out there in nature – the physical magnitudes of temperature and pressure as these apply to your kettle.

Now let's consider the infinite ascending hierarchy of polynomial models. There is "$y = a_0$", then LIN, then PAR, and so on. If you assume that the true model about your kettle is a polynomial, you can apply Jeffreys's simplicity postulate by assigning a prior probability of ½ to the first, ¼ to the second, and so on. The sum of this infinite series is 1.

What justifies assigning higher prior probabilities to simpler hypotheses? Jeffreys says that this comports with good scientific practice, but he didn't think that that the simplicity postulate is to be based on our observations of scientific practice and our impressions about what counts as "good." He says that our assumption that it is possible to learn from experience "leads" to the simplicity postulate, and his prose suggests that the simplicity postulate is supposed to be a "consequence" of the assumption. In his 1921 publication with Wrinch, he is more cautious. Wrinch and he say (p. 390) that "the most natural ways of well-ordering" the hypotheses are ones in which simpler hypotheses receive higher priors. The "naturalness" adverted to here isn't much of a justification. If your simplicity ordering is H_1,H_2, $H_3, \ldots$, you could follow the Jeffreys/Wrinch suggestion about priors, but you also could assign priors in a way that is not lockstep with simplicity. For example, you could let $\Pr(H_3) > \Pr(H_2) > \Pr(H_1) > \Pr(H_6) > \Pr(H_5) > \Pr(H_4)$, and so on. Learning from experience *does* require priors that are unequal if there are infinitely many alternative hypotheses, but there are lots of assignments of priors that achieve this.

Jeffreys's simplicity ordering has been criticized for not coinciding with our intuitive judgments about which hypotheses are simpler and which are more complex (see, for example, Ackermann 1963). I have taken a different angle. My main point is that even if Jeffreys's simplicity ordering coincides with our intuitive judgments concerning which hypotheses are simpler than which others, he gives no compelling reason for having this simplicity ordering dictate an ordering of prior probabilities.

Jeffreys wants his simplicity postulate to be *global*, not *local*. He doesn't look at inference problems one at time and in each case assign prior probabilities to the small handful of competing hypotheses that scientists take seriously. Rather, he wants his postulate to encompass *all* inference problems and *all* conceivable hypotheses. This is why he thinks he must consider infinitely many hypotheses. What is more, he needs this infinity to be denumerable, so that the machinery of assigning positive prior probabilities can be put to work without missing any candidate hypothesis. And he needs to ignore the fact that new scientific theories sometimes involve new concepts and new mathematical frameworks. Unfortunately, the idea that we now can enumerate the set of all possible future scientific theories is illusory.[18] And

[18] Just before he states his simplicity postulate, Jeffreys (1931, p. 36) says this: "so far as there is a problem, it would concern future attempts at specifying a wide range

even if we cut these grand ambitions down to size, there is trouble enough for Jeffreys's approach.

It is an interesting feature of Jeffreys's simplicity postulate that it places no upper bound on how complex the universe might be. In the infinite ascending hierarchy of polynomials, each polynomial gets a positive prior. Descartes, Leibniz, Newton, and Kant all thought that nature is simple and that this claim about the world is presupposed by the scientist's using simplicity to evaluate candidate theories. Jeffreys's simplicity ordering also has an implication about nature, but it is more modest. He is committed to the thesis that nature is *probably* simple; for example, in the probability assignment we have considered, the prior probability that one of the first n polynomials is true is $1 - (½)^n$. Accepting Jeffrey's simplicity postulate does not commit you to the stronger thesis that some of his predecessors felt obliged to embrace.

Popper's objection to Jeffreys's postulate

Karl Popper (1902–1994) was critical of Bayesianism in general and of Jeffreys's simplicity postulate in particular. He vehemently disagreed with Bayesians about the goals of science. Popper denied that science is in the business of finding theories that are probably true. Rather, he maintained that science should aim at formulating "bold conjectures" and then subjecting those conjectures to rigorous empirical test. Popper also disagreed with Jeffreys about the relationship of simplicity and probability. Whereas Jeffreys held that simpler theories are more probable, Popper argued that precisely the opposite is true.

Popper's point can be grasped by thinking about LIN and PAR. LIN is simpler than PAR in the sense that LIN has fewer adjustable parameters (since

of laws that might not admit classification into any of the types at present known to pure mathematics." He then says that "the class of all differential equations of finite order and degree, with rational coefficients, certainly includes all laws of classical physics; and they form an enumerable set . . . Further, the quantum theory modifies the classical differential equations in standard ways, and the quantum relations can hardly be more numerous than the classical ones." He then ignores the problem he described about future science and says that "the set of all possible forms of scientific laws is finite or enumerable, and their initial probabilities form the terms of a convergent series of sum 1." If Jeffreys's point is that the entire history of science on our planet will address only a finite number of theories, he is right, but then there is no need to think about an infinite series. And even if the number is finite, we still need to describe the theories so as to assign them prior probabilities.

$2 < 3$). But is LIN more probable than PAR? In his 1934 *Logik der Forschung* (whose English translation, *The Logic of Scientific Discovery*, appeared in 1959), Popper points out that LIN is a special case of PAR. LIN can be expressed equivalently as a conjunction:

$$(y = a_0 + a_1 x + a_2 x^2) \quad \text{and} \quad (a_2 = 0).$$

Note that the first conjunct of this conjunction is PAR. So LIN says that PAR is true *and* that $a_2 = 0$ is true as well. LIN entails PAR, since a conjunction entails each of its conjuncts. This means that LIN cannot be more probable than PAR; it is impossible for a conjunction to be more probable than one of its conjuncts; to think otherwise is to commit the conjunction fallacy described earlier in connection with Linda the feminist bank teller. According to Popper, Jeffreys got things backwards; for Popper simpler statements are *less* probable than more complex statements.

There is a disconnect between Jeffreys's position and Popper's criticism. Although Jeffreys does tell you to count adjustable parameters, he says something more when he lists his polynomials and assigns them priors; he takes the items on the list to be *mutually incompatible*. This is crucial to his idea that the sum of their prior probabilities cannot be greater than 1. In contrast, when Popper talks about examples like LIN and PAR, he takes these models to be *mutually compatible*; in fact, one of them logically entails the other. Statisticians describe this relationship by saying that LIN is *nested* inside of PAR. LIN is a special case of PAR; as we have just seen, LIN is obtainable from PAR by setting the parameter $a_2 = 0$. Jeffreys is implicitly assuming that the adjustable parameters in the models he considers are all non-zero; Popper makes no such assumption.

Popper is right that Jeffreys's simplicity postulate runs into the trouble if the models considered are nested. However, Popper's objection disappears if we consider only models that are not (Howson 1988). For example, instead of comparing LIN to PAR, you can compare LIN to the following model:

(PAR*) $\quad y = a_0 + a_1 x + a_2 x^2$, where $a_2 \neq 0$.

Although the axioms of probability theory entail that LIN can't have a higher prior probability than PAR, they leave open whether LIN has a higher prior probability than PAR*.

This reformulation of Jeffreys's simplicity postulate (wherein the postulate is limited to non-nested models) rescues it from Popper's objection, but that does not mean that the reformulated postulate is plausible. Indeed, the suggestion that we should take LIN to have a higher prior probability than PAR* is hard to swallow. Before you see any data at all from the kettle on your stove, you are supposed to think that it is more probable that $a_2 = 0$ than that $a_2 \neq 0$. This is like saying, before you drop a super-sharp dart onto a straight line that extends infinitely in two directions, that the dart has a higher probability of landing at zero than it has of landing non-zero.

Although Popper's objection to Jeffreys is less devastating than Popper imagined, Popper's focus on nested models is interesting. It may seem weird to consider two hypotheses as competitors when one of them entails the other, but this will make more sense when we turn to frequentism later in this chapter. For now, it is worth noticing a simple consequence of the fact that LIN is more parsimonious than PAR. Since the two are nested, you can't consistently accept LIN and deny PAR, nor can you consistently accept LIN and say that you have no commitment concerning the status of PAR. Sometimes a simpler theory is logically stronger than one that is more complex.[19] Neither the razor of denial nor the razor of silence applies to the case of nested models.

Popper, falsifiability, and corroboration

Given Popper's claim that simpler theories are *less* probable than more complex theories, why does he think that scientists should value simplicity? Popper's answer, in Chapter 7 of *The Logic of Scientific Discovery*, is that simpler theories are more falsifiable. Consider how many observations it would take to falsify LIN. Two observations are not enough; for any two points, a straight line can always be found that passes through them. It takes at least three points to falsify LIN. What about PAR? For any three points, a parabola can be found that passes through them. It takes at least four points to falsify PAR. So LIN is more falsifiable than PAR.

But what's so great about greater falsifiability? As you can see from the example of LIN and PAR, the more falsifiable hypothesis is easier to refute if it

[19] Terminology: The statement *S* is said to be "logically stronger" than the statement *W* precisely when *S* logically entails *W*, but not conversely. Stronger and weaker don't mean better and worse; they simply separate statements that say more from those that say less. In a Venn diagram, the region in which *S* is true is part of the region in which *W* is true.

is false. Gathering three data points is less work than gathering four. But that is a pragmatic fact; it is easy *for us* to refute a highly falsifiable theory if it is false. So what is the *epistemic* relevance of the fact that LIN is more falsifiable and simpler than PAR? Popper's answer is that theories that are more falsifiable say more than theories that are less falsifiable. Popper thinks that science is right to prize scientific theories that make bold, contentful claims. Newton's universal law of gravitation is, for Popper, a shining example. Newton made a claim about *all* the matter in the universe, no matter when or where it exists. He didn't meekly limit himself to proposing a theory for the small corner of the universe that we human beings happen to occupy. Newton's third Rule of Reasoning in Philosophy has a Popperian ring: "the qualities of bodies, which admit neither intension nor remission of degrees, and which are found to belong to all bodies within the reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever."

Popper is famous for saying that falsifiability solves the *demarcation problem*, which is the problem of saying how science differs from non-science. Popper's solution is that scientific statements are falsifiable, while unscientific statements are not. Falsifiable statements need not be false; a falsifiable statement is one that can be refuted by a finite body of observations *if* it is false. Popper understood falsification in terms of deduction. If a theory *T* deductively entails that an observation statement *O* should be true, and *O* turns out to be false, then you can deduce that *T* is false. The following argument is deductively valid:

If *T* then *O*
notO
———
notT

A falsifiable theory has deductive consequences concerning what we should observe. The twin of Popper's famous claim that scientific statements are falsifiable is his claim that they are not verifiable. Whereas the previous argument form is deductively valid, the following one is not:

If *T* then *O*
O
———
T

This argument form is fallacious; it is called *the fallacy of affirming the consequent.*[20] Just as falsifiability separates science from non-science, so degree of falsifiability separates scientific statements from each other. In science, more falsifiability is better than less, and some falsifiability is better than none at all.[21]

In claiming that more falsifiable theories are "better" than ones that are less falsifiable, Popper meant that we should prefer a logically stronger hypothesis over a weaker hypothesis when both deductively entail a set of observations. To bolster this thesis, Popper defined a mathematical concept that he called *degree of corroboration.* This may sound like another name for *degree of confirmation*, but "corroboration" and "confirmation" have taken on technical meanings that place them at odds with each other. Philosophers of science now usually use the term "confirmation" to denote the Bayesian concept I described earlier, but Popper intended his concept of corroboration to be part of his anti-Bayesian philosophy. Popper thinks that the Bayesian approach to confirmation is a dead end.[22] For Popper, evidence never supports hypotheses and it never supplies a good reason to accept them.

Popper (1959, p. 17; 1983, p. 240) defines $\mathrm{COR}_B(H \mid E)$, the degree to which evidence E corroborates hypothesis H (under background assumptions B), as follows:

$$\mathrm{COR}_B(H \mid E) = \frac{\Pr_B(E \mid H) - \Pr_B(E)}{\Pr_B(E \mid H) - \Pr_B(E \& H) + \Pr_B(E)}$$

[20] Notice that where Popper sees an asymmetry, Bayesians find a symmetry; as noted earlier, the Bayesian definition of confirmation entails that O confirms H if and only if not$O$ disconfirms H.

[21] A complication: the above description of falsifiability makes it sound as if Popper thinks that falsifiable theories deductively entail observations *all by themselves*. In fact, Popper (1959, p. 38) denies this and insists that the typical situation is that a theory issues in observational predictions only when it is supplemented with auxiliary assumptions. The right picture is that $T\&A \rightarrow O$, not that $T \rightarrow O$. This logical point is what philosophers of science now call Duhem-Quine thesis. Duhem took the logical point to have a holistic epistemological consequence; see Sober (2004) for criticism of this holism.

[22] Popper (1959, p. 375) claims that universal generalizations in an infinite domain must all have prior probabilities of zero. From this, it follows via Bayes's theorem that their posterior probabilities must also be zero, no matter what the evidence happens to be. Popper's claim is right when it is directed at one of Carnap's (1950) systems of inductive logic, but it isn't true of all versions of Bayesianism, not by a long shot.

In view of Popper's anti-Bayesianism, this is not a measure that he should have endorsed. The measure is Bayesian (Kreuth 2005, p. 121) since it requires a value for $\Pr_B(E\&H)$. Popper (1983, p. 240) says that the denominator in his measure of degree of corroboration was introduced because it gives the measure various desirable formal properties, but the fact remains that this is a measure that Popper should never have touched with a stick.

There is another glitch in Popper's measure of corroboration. It fails to deliver the result he wants – that a logically stronger hypothesis is always more corroborated than a logically weaker hypothesis when both entail the evidence. When $S \rightarrow W \rightarrow E$, $\text{COR}_B(S \mid E)$ and $\text{COR}_B(W \mid E)$ have the same numerators, since $\Pr_B(E \mid S) = \Pr_B(E \mid W) = 1$, but $\text{COR}_B(W \mid E)$ has the smaller denominator, since $\Pr_B(E\&W) > \Pr_B(E\&S)$. It follows that $\text{COR}_B(W \mid E) > \text{COR}_B(S \mid E)$, which is the opposite of what Popper wants.

Does this mean that we should try to repair Popper's definition so that it captures his idea that E corroborates S more than E corroborates W whenever $S \rightarrow W \rightarrow E$?[23] I don't think so. In saying this, I am using the term "corroboration" in its ordinary non-technical meaning, so even if Popper stipulates that logically stronger theories are always more strongly "corroborated," that doesn't matter. Suppose the data from the kettle on your stove consist of 20 data points that fall exactly on a straight line. LIN entails what I just told you about the data, but so does any conjunction of which LIN is a conjunct. For example, consider the following hypothesis:

(J) LIN & your kettle was made in Wisconsin.

J entails the observational report just as LIN does, but it is hard to see why the fact that J is logically stronger than LIN should force you to say that J is more strongly corroborated than LIN is. Suppose the linearity of your twenty data points is strong evidence *against* the hypothesis that the kettle was made in Wisconsin (since Cheese Heads are almost always non-linear thinkers). To make matters worse (for Popper), let's place J side-by-side with

(K) LIN & your kettle was not made in Wisconsin.

[23] Almost all Bayesian measures of degree of confirmation now on the market have the property that when $S \rightarrow W \rightarrow E$, $\text{DOC}(S,E) \leq \text{DOC}(W,E)$ (William Roche, personal communication).

J and K both entail LIN, so Popper will have to say that both are better corroborated than LIN is.

Let's not throw the baby out with the bathwater. Popper says that logically stronger hypotheses are *always* better corroborated than logically weaker hypotheses when both entail the observations. The comparison of LIN, J, and K throws doubt on that thesis. Even so, let's hold on to the idea that logically stronger hypotheses are *sometimes* better in a sense of "better" that is epistemically relevant. We will examine an epistemic framework that delivers that result when we turn to frequentism.

Popper's characterization of simplicity

Even if Popper could show that logically stronger hypotheses are better corroborated than logically weaker alternatives when both entail the evidence, there is an additional difficulty that besets his attempt to show that simplicity is epistemically relevant. His saying that simpler theories are more falsifiable makes sense when the example is LIN versus PAR, but there are other examples in which things go badly awry. We have already seen an example: proposition J is more falsifiable than LIN, but LIN seems to be the simpler hypothesis.

Another context in which Popper's view of simplicity seems mistaken involves probabilistic theories. I mentioned earlier that if you toss a coin *n* times, any sequence of heads and tails that you obtain is logically compatible with the hypothesis that the coin is fair, no matter how big *n* is. Judging simplicity by degree of falsifiability, Popper (1959, p. 141) concludes that all probabilistic theories are equally and infinitely complex. This conclusion is absurd. It also is absurd to say that probabilistic statements are unscientific, which is what Popper's proposed solution to the demarcation problem entails.[24]

Let's modify LIN and PAR to produce a pair of probabilistic theories that seem to differ in their simplicity. As stated, LIN says that the data you draw from your kettle must fall on a straight line and PAR says that the data you

[24] Popper (1959, p. 191) recognizes that his criterion of falsifiability founders when it is applied to probability statements. His response is to recommend a convention – that we regard a probability statement as refuted by observations when the statement says that the observations are very improbable. Popper notes that choosing a criterion for how improbable is improbable enough is arbitrary. For criticism of this repair of the demarcation criterion, see Sober (2008b, pp. 48–58, p. 130).

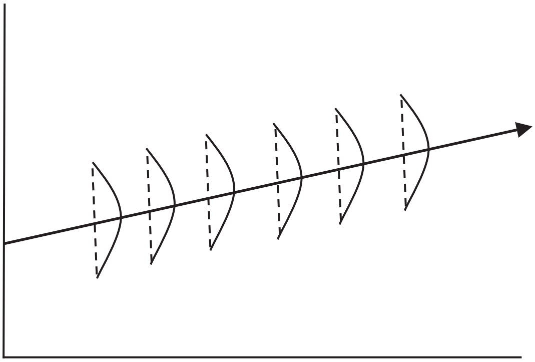


Figure 2.2

draw must fall on a parabola. Let's modify these two models by adding an error term to each:

(LIN$_e$) $y = a + bx + e$

(PAR$_e$) $y = a + bx + cx^2 + e$

These new models no longer say that *x*-values determine *y*-values. Rather, they say, for each *x* value, that there is an *expected y*-value around which there can be variation. Here I am using the concept of mathematical expectation that I explained earlier. The error term "*e*" usually represents a bell-shaped distribution. Whereas an instance of LIN is a straight line, an instance of LIN$_e$ is a straight line with a bell curve drawn around it, as shown in Figure 2.2. The figure is three-dimensional; the bell curve sticks up from the page and represents the probability distribution of the different *y*-values that can occur for a given *x*-value.

LIN$_e$ and PAR$_e$ are not falsifiable (in Popper's deductive sense of that term) by any finite data set. But surely they differ in their complexity. The obvious idea is something we have from Jeffreys; if you count adjustable parameters, you get a finite number for each model and the number for LIN$_e$ is smaller. The error term is just another adjustable parameter, not something that renders the model infinitely complex.[25] Of course, the question then arises of why this difference between LIN$_e$ and PAR$_e$ is epistemically relevant. Jeffreys had

[25] Another peculiarity in Popper's account of simplicity is to be found in his treatment of trigonometric functions like $y = \sin(x)$. These are infinitely complex if we measure complexity as Popper does. For any *n* data points, there is a sine curve that fits them perfectly. But surely we don't want to say that the sine model and its ilk

an answer, but it doesn't seem very convincing. Perhaps another account can do better. Keep that possibility in mind when we exit Bayesianism. For now, I'll continue exploring what Bayesianism can say about parsimony.

Parsimony and non-first priors

Jeffreys's goal was not to *justify* Ockham's razor; rather, he wanted to *use* simplicity to assign prior probabilities. If he had a good argument for his simplicity postulate, we could use that argument to provide a Bayesian justification for one version of Ockham's razor. But he did not, so we cannot.

Jeffreys was interested in "first priors," but, as noted earlier, prior probabilities are often understood in another way. There are "non-first priors." The prior $\Pr_A(H)$ will be a first prior if A consists only of logical truths, but if A includes empirical assumptions, the prior will be non-first. It is the latter sort of prior that is involved in setting Susan's prior for having tuberculosis at around $\frac{1}{100{,}000}$ because that is the frequency of the disease in Wisconsin. I noted before that there is a connection between both sorts of priors and the razor of silence: $\Pr_A(H \& J) \leq \Pr_A(J)$, regardless of what A is. Can something similar be said concerning the razor of denial?

Let's consider a modification of the story about Susan. Suppose you observe Susan's symptoms and consider two hypotheses. The first says that she has tuberculosis while the second says that she has disease *X*. Suppose the observed symptoms have the same probability under the two hypotheses; that is, the two likelihoods Pr(Susan's symptoms | Susan has tuberculosis) and Pr(Susan's symptoms | Susan has disease *X*) are equal. If tuberculosis is far more common in Wisconsin than disease *X*, it is reasonable to hold that Pr(Susan has tuberculosis) > Pr(Susan has disease *X*). This difference in the prior probabilities is enough to allow you to conclude that the tuberculosis hypothesis has the higher posterior probability. This is a consequence of the odds formulation of Bayes's theorem.

In describing this new example, I didn't mention parsimony. But you can imagine someone appealing to Ockham's razor to defend the claim that the hypothesis that Susan has tuberculosis is "better" than the hypothesis that she has disease *X*. The suggestion would be that the tuberculosis hypothesis

are infinitely complex. Once again, an attractive option is to measure simplicity by counting adjustable parameters.

is more parsimonious because it invokes a fairly common disease. Since you can explain the symptoms by invoking this hypothesis, there is no need to reach for the more esoteric hypothesis that Susan has disease *X*. Here "greater parsimony" is being used to indicate "higher prior probability." In this case, we have a good reason (our frequency data) for saying that one hypothesis has a higher prior probability than another; the priors are non-first.

It may sound far-fetched and potentially misleading to invoke Ockham's razor to defend the tuberculosis hypothesis. Why mention the razor if the point is really about prior probabilities? And why think that tuberculosis is a simpler hypothesis than disease *X*? I don't disagree with the sentiment behind these questions. I mention the example because it is useful to be alert to this way of talking about Ockham's razor. In 1966, the biologist George C. Williams (1926–2010) published a very influential book called *Adaptation and Natural Selection.* The book was a multi-faceted critique of hypotheses of group selection. Biologists beginning with Darwin had explained the existence of self-sacrifice and cooperation by hypothesizing that it was the upshot of competition among groups: groups of altruists do better than groups of selfish individuals in the struggle for existence. Williams argued that this is a huge mistake. He began his attack by announcing

> a ground rule – or perhaps doctrine would be a better term – that adaptation is a special and onerous concept that should be used only when it is really necessary. When it must be recognized, it should be attributed to no higher a level of organization than is demanded by the evidence. In explaining adaptation, one should assume the adequacy of the simplest form of natural selection, that of alternative alleles in Mendelian populations, unless the evidence clearly shows that this theory does not suffice.[26] (Williams 1966, pp. 4–5)

Williams calls this a parsimony principle (p. 18), but he does not clarify why it should be adopted; indeed, there are multiple interpretations that are worth considering. The one of interest here involves a claim about non-first priors. To describe this interpretation of Williams's ground rule, I want to consider an

[26] Williams's talk of higher and lower should remind you of Morgan's formulation of his canon, which I discussed in Chapter 1. Are the two principles connected or does "higher" have a different meaning in the two contexts?

influential argument against group selection that appeared two years before Williams's book.

John Maynard Smith (1920–2004) was an important warrior in the battle against group selection. In a paper he published in 1964, he assessed the hypothesis of group selection by constructing what came to be called the "haystack model." Imagine a field that contains numerous haystacks. Suppose that each haystack gets settled by a single pregnant mouse. She gives birth and her progeny then mate with each other, and the grandoffspring of the original foundress mate with each other, and so on, for numerous generations. Each mouse is either altruistic or selfish. Suppose that individuals who have the genotype *AA* or *Aa* are selfish, while those that are *aa* are altruistic. This means that a parental pair can give birth to both altruistic and selfish offspring. The selfish mice outcompete the altruists who live in the same haystack, so that by the end of all those generations, the only haystacks that contain altruists are ones that were founded by an altruistic female who had mated with an altruistic male. At this point, the individuals exit their haystacks and mate at random with each other. Then fertilized females found their own colonies (again, one foundress per haystack) and the process begins anew. If this cycle is iterated numerous times, the upshot is that it is almost certain that we'll find at the end that altruism has gone to 0%. The haystack model was widely interpreted as showing that the evolution of altruism is very improbable if group selection is the process at work.[27]

Now consider a team of biologists post-1966 who are starting to study some trait in a population (not mice) and are interested in the question of whether the trait evolved by group selection. If the biologists are persuaded by Maynard Smith's thought experiment, they may assign a very low prior probability to the hypothesis that the trait evolved by group selection. They do so without examining the biological details that are specific to the trait and the population in question. Perhaps they think that Maynard Smith's argument fleshes out the parsimony argument that Williams formulated.

Thanks to Williams and Maynard Smith, many biologists came to believe in the 1960s that group selection hypotheses aren't just *wrong* – they are naïve, confused, and subversive of the scientific enterprise. This thoroughly negative

[27] Williams (1966) does not cite Maynard Smith (1964), but Williams (pp. 113–117) offers his own reasons for thinking that group selection will rarely be effective.

position is now much less influential. Part of the change is that parsimony is no longer thought to be a magic bullet that disposes of all group selection hypotheses once and for all. The subject now is much more data-driven. Most biologists now agree that the hypothesis that a trait in a species evolved because of group selection has to be assessed by looking at evidence that is specific to that trait and that species. Sweeping arguments of the sort that Williams constructed have faded in their influence.[28] As for Maynard Smith's haystack model, it now is clear that the model is loaded with assumptions that *a priori* bias the case against group selection (Wade 1978; Sober and Wilson 1998, pp. 67–71). Friends and foes of group selection hypotheses now agree that invoking two causes (individual selection and group selection) rather than just one (individual selection only) must be supported by evidence; indeed, the same point holds for invoking one cause instead of two. The playing field, in this sense, is now level. This does not mean that Ockham's razor has no bearing on the choice between one cause and two; we'll get to that. But for now, we should place the interpretation I have sketched of Williams's parsimony argument in our inventory of razors. Talk of parsimony is sometimes a surrogate for claims about (non-first) prior probabilities.

Likelihoods and common causes

Unification was a prominent theme in the previous chapter's survey of the history of Ockham's razor. A theory U that unifies two sets of observations O_1 and O_2 seems to be simpler than the disunifying theory $T_1\&T_2$ (where T_1 explains O_1 and T_2 explains O_2). The question is whether this difference in simplicity is epistemically relevant. I'll consider this problem when I get to frequentism, but for now we are still in the world of Bayesianism. From the

[28] Although Williams's book contributed to the widespread conviction in the 1960s that group selection hypotheses can be dismissed *a priori*, the book also cites empirical reasons for rejecting specific hypotheses of group selection. An example is his discussion of sex ratio. And Williams concedes that there is at least one empirically substantiated case of group selection, citing Lewontin and Dunn's (1960) discussion of a segregator-distorter t-allele in the house mouse. It is puzzling how Williams could advance strong global arguments against group selection, given his positive assessment of this case. If the evolution of the *t*-allele in the house mouse involved group selection, maybe there are other traits out there in nature that did the same thing.

odds formulation of Bayes's theorem, you know that there are two factors that are relevant to determining whether $\Pr(U \mid O_1 \& O_2) > \Pr(T_1 \& T_2 \mid O_1 \& O_2)$; these are the priors of the two theories and their likelihoods. Having just discussed how Ockham's razor is related to the first of these considerations, I now turn to the second.

The term "theory" gets used in a variety of ways. Sometimes it is restricted to grand generalizations like Newton's theory of gravitation. At other times it is applied more widely, even to intellectual objects of quite modest proportions. To get things started, let's consider two "theories" of this second sort. They come from a book that Wesley Salmon (1925–2001) published in 1984 called *Scientific Explanation and the Causal Structure of the World*. You, an instructor in a philosophy class, assign your students the task of writing an essay on a topic you describe. Two of your students then turn in papers that are word-for-word identical. Not wishing to jump to a hasty conclusion, you consider two hypotheses:

(CC) The two students searched the Internet together and found a file that they agreed to plagiarize.

(SC) The two students worked separately and independently.

The CC theory postulates a common cause; SC postulates two separate causes. The observed matching of the two papers favors CC over SC, and the law of likelihood explains why. The evidence favors CC over SC because

$$\Pr(\text{the papers match} \mid \text{CC}) \gg \Pr(\text{the papers match} \mid \text{SC}).$$

The matching is something one would expect if the plagiarism hypothesis were true. In contrast, if the separate cause hypothesis were true, the matching would almost be a miracle; matching isn't *impossible* under that hypothesis, but it is *very* improbable. According to SC, the matching is a coincidence; according to CC, it is anything but.[29] I'll soon analyze this likelihood

[29] A similar example is provided by Alfred Wegener's argument for continental drift. Wegener (1880–1930) noticed that the wiggles in the eastern continental margin of South America correspond rather exactly to the wiggles in the western margin of Africa. He also learned that the distribution of geological formations running down the one matches the distribution running down the other and that the distribution of organisms on the two coasts – both fossilized and extant – also show a detailed correlation. The pattern is "just as if we were to refit the torn pieces of a newspaper by matching their edges and then check whether the lines of print run smoothly across"

argument in more detail; what I've just said about it is a first pass, intended to convey the intuitive idea. I hope my brief description of Salmon's example reminds you of material from the previous chapter – the difference between Ptolemy's and Copernicus's explanations of various astronomical regularities and Whewell's discussion of consilience of induction by analogy with two witnesses who agree.

The plagiarism example is interesting because the hypothesis with the higher likelihood also is the one that postulates fewer causes. This suggests that Ockham's razor may sometimes have a likelihood justification. The word "sometimes" is important here. It is there to leave room for the possibility that some applications of Ockham's razor are justified for reasons having nothing to do with the law of likelihood; it also leaves room for the possibility that some applications have no justification at all.

Hans Reichenbach (1891–1963), who was Salmon's teacher at UCLA, thought that postulating common causes plays an important role in scientific and everyday thinking. In his book *The Direction of Time*, Reichenbach (1956, p. 157) enunciates a principle that he calls

> *The Principle of the Common Cause*: If an improbable coincidence has occurred, there must exist a common cause.

Reichenbach gives several brief examples that resemble Salmon's example about plagiarism. He begins with these two.

- Both lamps in a room go out suddenly. We regard it as improbable that by chance both bulbs burned out at the same time, and look for a burned-out fuse or some other interruption of the common power supply.
- Several actors in a stage play fall ill showing symptoms of food poisoning. We assume that the poisoned food stems from the same source – for instance, that it is was contained in a common meal – and thus look for an explanation for the coincidence in terms of a common cause.

Reichenbach then says that "chance coincidences, of course, are not impossible: the bulbs may burn out simultaneously, the actors become ill simultaneously for different reasons. The existence of a common cause is therefore in

(Wegener 1928, p. 77). Wegener argued that this systematic matching should not be dismissed as a mere coincidence. His preferred alternative was that the continents had once been in contact and then had drifted apart. Wegener's pieces of newspaper are like Salmon's student essays.

such cases not absolutely certain, but only probable" (pp. 157–158). Reichenbach's point is that his formulation of the principle of the common cause is too strong – the phrase "must exist" should be replaced with "probably exist." Reichenbach then adds "this probability is greatly increased if coincidences occur repeatedly." He illustrates this point by providing two more examples.

- Two geysers which are not far apart spout irregularly, but throw up their columns of water always at the same time. The existence of a subterranean connection of the two geysers with a common reservoir of hot water is then practically certain.
- The fact that measuring instruments such as barometers always show the same indication if they are not too far apart is a consequence of the existence of a common cause – here, the air pressure.

Notice that the first two examples are one-off: the two lamps each go out once and the actors all get sick on a single day. The third and fourth examples involve repetitions; in each there is a temporal series of events that exhibits a pattern. In the first pair, two event tokens have the same property; in the second, there is an association between two event types.[30]

When Reichenbach talks about "improbable coincidence," he means *correlation*. Modifying the geyser example slightly, let's imagine that each geyser spouts once each week for a single hour. So each geyser's frequency of spouting is 1 hour out 24×7 hours $= \frac{1}{168}$. Now suppose that whenever the one geyser spouts, the other always does too. The striking fact about these frequencies is that

$$\begin{gathered}\text{freq(geyser 1 spouts \& geyser 2 spouts)}\\ > \text{freq(geyser 1 spouts)freq(geyser 2 spouts).}\end{gathered}$$

The inequality is true because $\frac{1}{168} > (\frac{1}{168})(\frac{1}{168})$. This fact about frequencies may lead you to think that the following probabilistic inequality is true:

$$\begin{gathered}\text{Pr(geyser 1 spouts at time } t \text{ \& geyser 2 spouts at time } t)\\ > \text{Pr(geyser 1 spouts at time } t)\text{Pr(geyser 2 spouts at time } t).\end{gathered}$$

[30] Terminology: the type/token distinction can be grasped by considering the ambiguity in the statement "you and I own the same shirt." This might mean that there are two shirts that are similar and we each own one. Or it may mean that there is a single shirt and we are co-owners. In the first case, "same shirt" means *same type of shirt*; in the second, it means *same token shirt.*

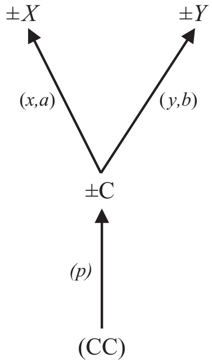


Figure 2.3

Notice that I am taking pains to separate frequencies from probabilities; this echoes a point I made at the start of this chapter, where I talked about the difference between a coin's probability of landing heads and the frequency with which it actually does so. When Reichenbach says that the behaviors of the two geysers exhibits an "improbable coincidence," he means that they spout simultaneously more often than one would expect if they were probabilistically independent of each other. In order to mark the difference between the frequency inequality and the probabilistic inequality, I will henceforth use the term "association" for the former and "correlation" for the latter.

If we interpret Reichenbach's "improbable coincidence" to mean *correlation*, it becomes clear that his principle of the common cause, as stated, requires a further modification. Two events can be correlated just because the one causes the other. What Reichenbach intended is the following: *if there is a correlation between two events, then (probably) one causes the other or they have a common cause.* When Reichenbach describes the geyser example, he says that the two events are simultaneous. If cause must precede effect, then Reichenbach's principle entails that the correlated events probably have a common cause.

How does Reichenbach justify his principle? We have already seen that he presents several examples that are intended to persuade the reader that the principle is sensible. But Reichenbach does something more; he argues that a common cause explanation of *X* and *Y*, if formulated in a certain way, will *entail* that *X* and *Y* are correlated. Figure 2.3 represents a common cause explanation that connects three variables (±*X*, ±*Y*, ±*C*), each of which comes in two states. Suppose we're talking about the two geysers. At any moment the geyser represented by $\pm X$ is either erupting or is not; ditto for the geyser represented by $\pm Y$. And at any moment the postulated underground source, represented by $\pm C$, is either in a high pressure or a low pressure state. Whereas each variable is represented by a letter preceded by "±", the two states of a variable correspond to two propositions; when I use X, Y, and C without the "±" prefix, I am talking about propositions, not variables. I'll represent the common cause hypothesis that Reichenbach has in mind by using the following notation:

$$\begin{aligned} &\Pr_{CC}(X \mid C) = x \quad \Pr_{CC}(X \mid notC) = a \\ &\Pr_{CC}(Y \mid C) = y \quad \Pr_{CC}(Y \mid notC) = b \\ &\Pr_{CC}(C) = p. \end{aligned}$$

All of these probabilities are creatures of the common cause hypothesis. The assumptions I'll describe that shape the $\Pr_{cc}(-)$ probability function all say that there is a common cause of $\pm X$ and $\pm Y$. C is the proposition that the common cause is in one state; the other possibility is that the common cause is in the state described by *notC*. It is a mistake to think that the proposition *notC* asserts that X and Y have no common cause!

There is more to Reichenbach's common cause hypothesis than the bare claim that $\pm X$ and $\pm Y$ have a common cause $\pm C$. The additional content involves three assumptions:

Assumption 1 (screening-off): $\Pr_{CC}(X\&Y \mid \pm C) = \Pr_{CC}(X \mid \pm C)\Pr_{CC}(Y \mid \pm C)$.[31]
Assumption 2 (nonzero): $\Pr_{cc}(C)$ and $\Pr_{cc}(notC)$ are both positive.
Assumption 3 (positive correlation of the cause with each effect):

$$\Pr_{CC}(X \mid C) > \Pr_{CC}(X \mid notC) \quad \text{and} \quad \Pr_{CC}(Y \mid C) > \Pr_{CC}(Y \mid notC).$$

What we need to show is that these three assumptions entail

$$(*) \qquad \Pr_{CC}(X\&Y) > \Pr_{CC}(X)\Pr_{CC}(Y).$$

Notice that (*) says that X and Y are unconditionally dependent, whereas Assumption 1 says that they are conditionally independent. As promised earlier, I'll now explain how conditional independence can be part of the explanation of unconditional dependence.

[31] $\Pr(X\&Y \mid \pm C) = \Pr(X \mid \pm C)\Pr(Y \mid \pm C)$ means that $\Pr(X\&Y \mid C) = \Pr(X \mid C)\Pr(Y \mid C)$ and $\Pr(X\&Y \mid notC) = \Pr(X \mid notC)\Pr(Y \mid notC)$.

The theorem on total probability allows (*) to be rewritten as

$$\mathrm{Pr}_{\mathrm{CC}}(X\&Y \mid C)\mathrm{Pr}_{\mathrm{CC}}(C) + \mathrm{Pr}_{\mathrm{CC}}(X\&Y \mid notC)\mathrm{Pr}_{\mathrm{CC}}(notC) > \mathrm{Pr}_{\mathrm{CC}}(X)\mathrm{Pr}_{\mathrm{CC}}(Y).$$

With the screening-off Assumption 1, this entails

$$\mathrm{Pr}_{\mathrm{CC}}(X \mid C)\mathrm{Pr}_{\mathrm{CC}}(Y \mid C)\,\mathrm{Pr}_{\mathrm{CC}}(C) + \mathrm{Pr}_{\mathrm{CC}}(X \mid notC)\,\mathrm{Pr}_{\mathrm{CC}}(Y \mid notC)\mathrm{Pr}_{\mathrm{CC}}(notC)$$
$$> \mathrm{Pr}_{\mathrm{CC}}(X)\mathrm{Pr}_{\mathrm{CC}}(Y),$$

which translates into our algebraic symbolism as follows:

(**) $xyp + ab(1-p) > [xp + a(1-p)]\,[yp + b(1-p)]\,.$

This last inequality expands to

$$xyp + ab(1-p) > xyp^2 + xbp(1-p) + ayp(1-p) + ab(1-p)^2$$

and this rearranges to

(***) $xyp\,(1-p) + abp(1-p) > xbp(1-p) + ayp(1-p).$

Assumption 2 allows us to divide by $p(1-p)$ and thus to simplify this inequality to

$$xy + ab > xb + ay,$$

which rearranges to

$$x(y-b) > a(y-b),$$

which becomes

$$(x-a)(y-b) > 0.$$

This inequality follows from Assumption 3. QED. So Reichenbach's result takes the form of a conditional: if there is a common cause ±C of ±X and ±Y that satisfies the three assumptions, then ±X and ±Y will be correlated. This conditional is a truth of mathematics. For future reference, I will call it *Reichenbach's theorem*.[32]

[32] Reichenbach's theorem generalizes to variables that are not dichotomous. Suppose E_1, E_2, and C are (possibly non-dichotomous) events where C screens-off E_1 from E_2. Consider any state x (of E_1) and any state y (of E_2) for which the following two conditions hold: (i) for all states c and c' of C, if $\mathrm{Pr}(E_1 = x \mid C = c) > \mathrm{Pr}(E_1 = x \mid C = c')$, then $\mathrm{Pr}(E_2 = y \mid C = c) > \mathrm{Pr}(E_2 = y \mid C = c')$, and (ii) there exist at least two states c_1 and c_2 for C for which $\mathrm{Pr}(C = c_1) > 0, \mathrm{Pr}(C = c_2) > 0$, and $\mathrm{Pr}(E_1 = x \mid C = c_1) > \mathrm{Pr}(E_1 = x \mid C = c_2)$. Then it follows that $\mathrm{Pr}(E_1 = x \text{ and } E_2 = y) > \mathrm{Pr}(E_1 = x)\,\mathrm{Pr}(E_2 = y)$.

Let's suppose that we are considering an example like Reichenbach's geysers in which we are convinced by abundant observations that two events are correlated. Does it follow that Reichenbach's common cause model is true? No it does not. We need to steer well clear of the *fallacy of affirming the consequent*, which I mentioned earlier in the discussion of Popper. Reichenbach's theorem tells us that if Assumptions 1–3 are true, then the two events will be correlated. We know that the events are correlated. It does not follow that Assumptions 1–3 are true.

Given this simple point, let's consider a more modest question: does the correlation of the two effects, plus the fact that this correlation is a consequence of Reichenbach's three assumptions, show that there *probably* exists a common cause that conforms to his three assumptions? To conclude this, you would need to show that the model's posterior probability is high (or at least that it is greater than 0.5). But nowhere in Reichenbach's argument are prior or posterior probabilities for the common cause hypothesis described. Reichenbach's *argument* is not Bayesian, though his principle of the common cause asserts that there probably exists a common cause.

Still more modestly, we can ask whether the correlation is *evidence* for the common cause model. If hypothesis H entails the evidence statement E, there is a very general circumstance in which E confirms H in the Bayesian sense of confirmation described earlier. Bayes's theorem entails that if $0 < \Pr(E) < 1$, $0 < \Pr(H) < 1$, and $H \rightarrow E$, then $\Pr(H \mid E) > \Pr(H)$. Look at Bayes's theorem and convince yourself that this is right. It is tempting to invoke this fact to argue that the correlation of the two events confirms Reichenbach's common cause hypothesis. Unfortunately, there is a fly in the ointment.

The fly is the distinction between association and correlation. It is the former that we observe, not the latter. The correlation that I derived (following Reichenbach) involves probabilities that are internal to the common cause model. This is why I used the "CC" subscript throughout. The quantities $\Pr_{\mathrm{CC}}(X\&Y)$, $\Pr_{\mathrm{CC}}(X)$, and $\Pr_{\mathrm{CC}}(Y)$ are all defined within the common cause model, and the inequality $\Pr_{\mathrm{CC}}(X\&Y) > \Pr_{\mathrm{CC}}(X)\Pr_{\mathrm{CC}}(Y)$ is a creature of that model. This inequality does not express an "observation" if that term is supposed to mean a fact we can ascertain without using the model in question.[33]

[33] Another reason to avoid treating a probabilistic correlation as an observation within a Bayesian format is that it requires one to talk about second-order probabilities; one needs to assign a probability to a probabilistic proposition. Bayesianism does not prohibit this, but it would be nice to avoid this complication.

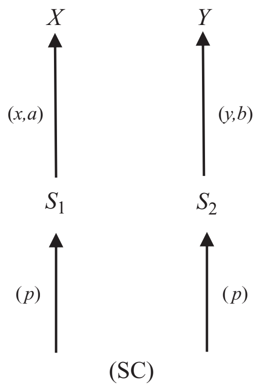


Figure 2.4

And here's the punch line: if what you observe is an association (not a correlation), you can no longer say that the common cause model *logically entails* what you observe. Perhaps the model says that the observed association was very probable, but that isn't enough to show that the observation confirms the model. So the Bayesian argument in the previous paragraph, which purports to show that the correlation is evidence for the common cause model, does not work.

This does not mean that Reichenbach's theorem is pointless; the theorem just needs to be supplemented. As the example of plagiarism suggests, we can describe a circumstance in which a common cause model has a higher likelihood than a separate cause model. The task at hand is to do something that Reichenbach's theorem does not do; we need to construct a competing model that postulates separate causes. One such model is depicted in Figure 2.4. For now, you can ignore the letters labelling branches in that figure; they will come in handy later. The separate cause model I want to consider is given by three assumptions:

Assumption 4 (the separate causes are probabilistically independent of each other):

$$\Pr_{SC}(\pm S_1 \,\&\, \pm S_2) = \Pr_{SC}(\pm S_1)\Pr_{SC}(\pm S_2).$$

Assumption 5 (the separate causes together screen-off each effect from the other):

$$\Pr_{SC}(X\&Y \mid \pm S_1 \,\&\, \pm S_2) = \Pr_{SC}(X \mid \pm S_1 \,\&\, \pm S_2)\Pr_{SC}(Y \mid \pm S_1 \,\&\, \pm S_2).$$

Assumption 6 ("locality" – each component cause screens-off the other):

$$\Pr_{SC}(X \mid \pm S_1) = \Pr_{SC}(X \mid \pm S_1 \,\&\, \pm S_2) \quad \text{and}$$

$$\Pr_{SC}(Y \mid \pm S_2) = \Pr_{SC}(Y \mid \pm S_2 \,\&\, \pm S_1).$$

This separate cause model entails that *X* and *Y* will be uncorrelated. Since the common cause model entails that *X* and *Y* will be positively correlated, you can check the data to see which model is more in accord with what we observe. If you observe that *X* and *Y* are positively associated, you can conclude that the data favor the common cause over the separate cause model (in the sense of the law of likelihood):

Pr(*X* and *Y* are positively associated | the CC model defined by Assumptions 1–3) >
Pr(*X* and *Y* are positively associated | the SC model defined by Assumptions 4–6).

Notice that it isn't any old common cause model and any old separate cause model that we are talking about here. On the contrary, it is the highly specific common cause model that Reichenbach constructed and the highly specific separate cause model that I just constructed that are being compared.

When we say that the eruption of Reichenbach's two geysers involves an "association," we are offering a summary description of a detailed chronicle that records what happens each hour to each geyser over some extended period of time. The claim of association is logically weaker than the raw data. We have seen that the observed association favors the common cause over the separate cause model (once various assumptions are put in place). However, this is not the only kind of data that can do so. Notice that each model treats each of the observations in the raw data as the product of a single process. For example, the common cause model says that the probability of the first geyser's erupting on January 7, 2014 between 10 a.m. and 11 also describes that geyser's probability of erupting on February 15, 2014 between 3 p.m. and 4, and so on for all the hours of all the days in the data set. The common cause model imposes the same pattern on the second geyser. The common cause model also says that the probability of a geyser's erupting at a later time is not influenced by whether it erupted earlier. These features of the common cause model also apply to the separate cause model. That is, both models assume that the processes generating the data we have on the two geysers are *i.i.d.* – they are *independent and identically distributed*. Here "independent" means probabilistically independent and "identically distributed" means that the same probability distribution attaches to each time. Perhaps the i.i.d. assumption is plausible for the geyser example, but sometimes it is unsatisfactory; when it is unsatisfactory, we need to reconsider how common cause and separate cause models can differ in likelihood.

An example in which *i.i.d.* is an implausible assumption has already been discussed; it is Salmon's example of the student papers that match. A word's probability of appearing in a given place in an essay depends on which place you're talking about. For example, "or" has a lower probability of being the first word than it has of being the thirtieth. In addition, the probability of a word's appearing later is often influenced by whether the word appeared earlier. Here we can learn from Chekov's quip that a loaded gun that appears in the first act of a play must be fired by the end: if "gun" appears early in an essay, this raises the probability that the word will appear later. Even if Reichenbach's example of the two geysers is reasonably treated as *i.i.d.*, Salmon's example of the two student essays is not.

species 1	$+T_1$	$+T_2$	$+T_3$	$+T_4$	$+T_5$	$+T_6$	$+T_7$	$+T_8$
species 2	$+T_1$	$+T_2$	$+T_3$	$+T_4$	$+T_5$	$+T_6$	$+T_7$	$+T_8$

To move beyond the world of *i.i.d.*, we do not need to start from scratch; the Reichenbachian common cause model and the separate cause model that I constructed can still be used, but we need to rethink what the data are. As an example, suppose we consider two species and wonder whether they have a common ancestor. The common ancestry hypothesis postulates a common cause and the separate ancestry hypothesis postulates separate causes. To evaluate these two hypotheses, we score each species for eight dichotomous characters.[34] Suppose we obtain the data shown in the accompanying table. Since both species are in the + state for all eight characters, the association is zero in the sense that

$$\begin{aligned}&\text{freq(species 1 is + and species 2 is +)} = 8/8 = \\ &\quad \text{freq(species 1 is +)freq(species 2 is +)} = (8/8)(8/8).\end{aligned}$$

However, this zero association does not force us to conclude that the data fail to favor common ancestry over separate ancestry. Whether the data speak in favor of common ancestry depends on facts about each of the eight characters. These characters may be very different from each other. It may be completely

[34] Here I use terminology from biology. A *character* is a variable, and it has various *character states*. The older philosophical terminology is *determinable and determinate*.

artificial to think that a species' scoring a + on one character is "the same" as its scoring a + on another.[35]

Instead of focusing on the observation of a zero association, we can focus on the characters we observe one by one. For each character, we can assess whether it favors common ancestry over separate ancestry. We therefore have eight observations, one for each of the eight traits ($\pm T_1, \pm T_2, \ldots, \pm T_8$) in our data set. Each has the form:

> Species 1 has trait T_i & species *2* has trait T_i.

Here I am assuming that the two species are token objects.[36] The present problem resembles the problem of explaining why the two student essays in Salmon's example are in the same state. The common ancestry hypothesis says that there exists a most recent common ancestor of species *X* and *Y* and that it has some state or other for each of the eight traits. To bring the common cause hypothesis (described by Assumptions 1–3) and the separate cause hypothesis (described by Assumptions 4–6) into contact with each other in terms of what each says about a single trait that is exhibited by both *X* and *Y*, we need one more assumption. It is here that the letters labeling branches in Figures 2.3 and 2.4 become relevant:

> *Assumption 7* ("cross-model homogeneity"): The parameters used in the common cause model have the same values as the counterpart parameters that are used in the separate cause model.

Fleshed out in this way, the separate cause hypothesis "mimics" much of what is asserted by the common cause hypothesis. With this added assumption, we can derive the following result concerning each of the eight characters shown in the table of data:

$$\begin{aligned}&\Pr(\text{the two species have } +T_i \mid \text{common ancestry})\\&\quad > \Pr(\text{the two species have } +T_i \mid \text{separate ancestry}),\ \text{ for each } i.\end{aligned}$$

[35] If it makes no sense to say that a + on one character is "the same" as a + on the other, then the data described in the table can be recoded so that four of the matches involve + states and the other four involve − states. Now we have a strong positive association. This underscores the point that it is sometimes the character states that matter one by one, not their "association."

[36] I am here assuming that species are individuals (or, at least, are historical objects), a view defended by Ghiselin (1974) and Hull (1978).

Using the algebraic representation we have from before, this inequality gets expressed as

$$xyp + ab(1-p) > [xp + a(1-p)]\,[yp + b(1-p)]\,.$$

This is the same inequality (**) I used earlier to express the idea that the common cause model entails that X and Y are unconditionally correlated; now it says that the observation that species 1 and species 2 both have trait $+T_{\mathrm{i}}$ is more probable under the common ancestry hypothesis than it is under the hypothesis of separate ancestry. Notice that the two hypotheses have the same likelihoods given the observation of a single species:

$$\begin{aligned}&\Pr(\text{species } i \text{ has trait } T \mid \text{common ancestry})\\&\quad = \Pr(\text{species } i \text{ has trait } T \mid \text{separate ancestry }) \text{ for } i = 1, 2.\end{aligned}$$

It is the *conjunction* "species 1 has trait T & species 2 has trait T" that is more probable under the one hypothesis than it is under the other.[37] Returning to the data set in which the two species match on each of the eight characters scored, we can say that there are eight "votes" in favor of common ancestry and zero against. And just as each match on a dichotomous character favors common ancestry over separate ancestry, a mismatch would have the opposite evidential significance.[38]

These results are qualitative, not quantitative. When two species have the same trait, this favors common ancestry, but that does not mean that all shared traits provide the same strength of evidence. Darwin anticipates this point in *the Origin* when he says that

> adaptive characters, although of the utmost importance to the welfare of the being, are almost valueless to the systematist. For animals belonging to two most distinct lines of descent, may readily become adapted to similar conditions, and thus assume a close external resemblance; but such

[37] This is also a consequence of Myrvold's (2003) Bayesian account of unification, although there are other respects in which the Reichenbachian likelihood inequality and Myrvold's account diverge.

[38] It is a mathematical fact that if $\Pr(E \mid \mathrm{CA}) > \Pr(E \mid \mathrm{SA})$, then $\Pr(notE \mid \mathrm{CA}) < \Pr(notE \mid \mathrm{SA})$. However, it is not a mathematical fact that if $\Pr(E \mid \mathrm{CA}) > \Pr(E \mid \mathrm{SA})$ then $\Pr(F \mid \mathrm{CA}) < \Pr(F \mid \mathrm{SA})$ when E and F are incompatible. So do not conclude that since X and Y's sharing trait T favors *CA* over *SA*, then even the slightest difference between X and Y must have the opposite evidential significance. The shift from a dichotomous variable to one that has more than two states is all-important.

> resemblances will not reveal – will rather tend to conceal their blood-relationship to their proper lines of descent. (Darwin 1859, p. 427)

The fact that sharks and dolphins are both shaped like torpedoes provides scant evidence for their having a common ancestor, since natural selection would promote the evolution of this trait whether or not the two groups have a common ancestor. The situation changes when we consider traits that are useless in one or both of the groups considered. This is why the tailbone shared by monkeys and humans is telling. I have called the point that Darwin is making in this passage *Darwin's Principle* (Sober 2008b, 2011a): adaptive similarities provide negligible evidence for common ancestry whereas neutral or deleterious similarities provide stronger evidence.

To provide a formal representation of Darwin's Principle, we need to define a new concept. The law of likelihood provides a criterion for whether an observation favors one hypothesis over another, but it says nothing about how much favoring is going on. A natural supplement is to use the likelihood ratio to quantify strength of favoring (Royall 1997):

> The degree to which E favors H_1 over H_2 is given by the likelihood ratio $\frac{\Pr(E \mid H_1)}{\Pr(E \mid H_2)}$.[39]

We can use the likelihood ratio to represent Darwin's point:

$$\frac{\Pr(X \text{ and } Y \text{ have trait } T \mid \textit{Common Ancestry})}{\Pr(X \text{ and } Y \text{ have trait} T \mid \textit{Separate Ancestry})} \approx 1 \text{ when } T \text{ is adaptive for both } X \text{ and } Y.$$

$$\frac{\Pr(X \text{ and } Y \text{ have trait } T \mid \textit{Common Ancestry})}{\Pr(X \text{ and } Y \text{ have trait } T \mid \textit{Separate Ancestry})} \gg 1 \text{ when } T \text{ is not adaptive for both } X \text{ and } Y.$$

There is nothing especially biological about Darwin's idea here. In the plagiarism example, there are many similarities that unite the two student essays. The fact that both essays are divided into paragraphs is scant evidence for the

[39] The law of likelihood demands that degree of favoring depend just on the two likelihoods, but why opt for the ratio, rather than, say, the difference? One reason is that more and more evidence often reduces the magnitudes of each likelihood, so that the likelihoods of competing hypotheses grow closer together as each approaches zero. But even when this happens, the likelihood ratio can increase without bound, and in such cases the growing body of evidence seems to favor one hypothesis over the other ever more profoundly (Sober 2008b, p. 33).

plagiarism hypothesis. The identical misspelling of various words in the two essays is more weighty.[40]

Returning to the data set of eight matches that characterize the two species, we can use the likelihood ratio measure to say the following: if the different characters are independent of each other (conditional on common ancestry and conditional on separate ancestry), then the eight matchings together favor common ancestry over separate ancestry more strongly than does any single match taken by itself. Since the likelihood ratio for each character has a value greater than one, the product of eight such ratios is itself a likelihood ratio that must be bigger than any of the eight.

It is important to see that the assumptions that characterize the common cause and the separate cause models are substantive. Change them and the common cause model may not have the higher likelihood. For example, the greater likelihood of the common cause over the separate cause model can be nullified by dropping the assumption that all probabilities in the model are non-zero. If you assume that $\Pr_{CC}(C) = 0$ or 1, and leave all the other assumptions in place, the two models have identical likelihoods; have a look back and see what happens to it when p is assigned a value of 0 or 1. For the common cause and the separate cause hypotheses to differ in likelihood, it is key that those hypotheses *not* specify the states of the postulated causes. For another example, suppose that the two causes postulated by the separate cause model are perfectly correlated; that is, $\Pr_{SC}(S_2 \mid S_1) = 1$ and $\Pr_{SC}(S_2 \mid not S_1) = 0$. If that's the only change you make in the models, the result is again that the two models have identical likelihoods. This means that if you think that the common cause model is "better" than the separate cause model even when the postulated causes are deterministic, and even when the postulated separate causes are perfectly correlated, you need to shop elsewhere.

It isn't just that a likelihood inequality can be changed to an equality by jiggling assumptions; the inequality can be reversed. That is, there are assumptions that entail that the matching of two effect favors separate causes over a common cause. Imagine a population of organisms in which each

[40] Darwin's contrast between adaptive and non-adaptive similarities needs to be refined. For example, it turns out that some adaptive similarities provide stronger evidence for common ancestry than neutral similarities provide; see Sober and Steel (2014).

parental pair has one daughter and one son. Suppose that generations do not overlap; this means that all the individuals in the parental generation reproduce simultaneously and then die. We then observe two organisms from the offspring generation and note that both are female. This similarity is evidence *against* their being siblings. Here's a similar example: for many years in Britain, the rule of primogeniture dictated that the eldest son would inherit all of the parental fortune (and the aristocratic title, if any). Given this, the observation that two men are both very rich (or that both are dukes) is evidence against their being brothers. In these two examples, an observed similarity is evidence *against* a common cause hypothesis.

Another case in which the likelihood inequality "flips" can be found by considering Assumption 3, the assumption of positive correlation. If we observe that species X and Y both have trait T, this assumption says that a state of the postulated common cause that raises the probability of X's having T also raises the probability of Y's having T.[41] It isn't essential that the probabilities that pertain to X have the same values as those that pertain to Y. What is required is just that $(x - a)$ and $(y - b)$ have the same sign. This assumption is substantive. Imagine several businesses, each with a single boss and a number of employees. You observe that two employees are both unhappy with their jobs. Is that shared trait evidence that the two employees have the same boss? If Reichenbachian assumptions hold, it is. But suppose that each boss is either strict or not strict, where these traits have the following property: if X's boss is strict, this raises the probability that X will be unhappy, but if Y's boss is strict, this lowers the probability that Y will be unhappy. Now the unhappiness of X and Y is evidence that they have *different* bosses.

Similar issues arise in connection with Assumption 7 (cross-model homogeneity). Notice that the probability parameter (p) assigned by the common cause model to C's having the plus state is exactly the same as the parameter assigned by the separate cause model to S_1's and S_2's having that state. This expresses the assumption that the probability a cause has of being in a given state is independent of whether that cause has two effects or just one. Suppose we observe that two individuals in the offspring generation have the same genotype. Does this evidence favor the common cause hypothesis that they are siblings over the hypothesis that they are not? Assumptions 1–7 entail the

[41] This specific assumption can be seen as a special case of the broader and vaguer idea of uniformity of nature, which I discussed in Chapter 1 in connection with Hume.

former conclusion, but if we retain 1–6 and change assumption 7, we can get the opposite result. For example, suppose the genotype exhibited by both of the individuals in the offspring generation is evidence that each came from a parental pair of a given genotype, and it is known that parental pairs with that genotype rarely have more than one offspring. Change Assumption 7 in this way and the matching of two effects will favor the hypothesis of separate causes over the common cause hypothesis.

Salmon's example of student plagiarism can be embellished so that Assumption 7 is violated. The common cause hypothesis says that the two students go to the Internet together and find a paper there that they agree will be the one that they both will copy. Suppose they do an Internet search, turn up a thousand candidate essays, and then choose one of them at random to be their source. The common cause hypothesis therefore says that the probability of their having chosen the essay they do is $\frac{1}{1000}$. What does the separate cause hypothesis say? I formulated it like this: the two students work in isolation from each other and don't plagiarize at all. But if this is what the separate cause explanation says, there is no reason to think that the parameter p that the common cause model deploys should be carried over to the separate cause model. If we use new parameters p_1 and p_2 for the separate cause model, we'll need to add a new assumption about how these two parameters are related to p if we are to decide whether the common cause model has the higher likelihood. This problem would not arise if the separate cause hypothesis were reformulated to say that each student separately decides to go to the web and chooses an essay at random from 1,000 candidates; now the cross-model homogeneity assumption is satisfied.

In Chapter 1, I briefly discussed Newton's second rule of reasoning in philosophy: *Therefore to the same natural effects we must, as far as possible, assign the same causes.* Newton does not say that similarities *always* have a common cause explanation, but he thinks it is good epistemic practice to treat them that way. Another thing that Newton does not say is that this preference is *a priori*, but it is easy to take this conclusion away from his prose. I have been arguing in the preceding pages that the *ought* connecting observed similarity and postulated common cause is more contingent, depending as it does on the substantive assumptions that structure the common cause and the separate cause models that are in competition. Sometimes it is true that the common cause hypothesis has a higher likelihood than the separate cause hypothesis, but there also are cases in which the likelihoods are equal, and still others in which it is the separate cause explanation that wins the likelihood competition. Common cause explanations are, in an intuitive sense, more parsimonious because they postulate fewer causes, but whether parsimony is epistemically relevant, and how it is relevant, depend on the background assumptions that are in place.

Ockham's razor is often thought to regulate *existence claims*. Understood in this way, the principle of parsimony tells you to regard such claims as *guilty until proven innocent*. You should begin with the parsimonious assumption that *X*s do not exist and should abandon that position only if your observations force you to do so. The likelihood justification of the idea that common cause explanations are superior to separate cause explanation shows why it is a mistake to characterize Ockham's razor in terms of "existence claims" and leave it at that. True, the common cause explanation postulates one cause and the separate cause explanation postulates two. But the common cause explanation postulates one common cause whereas the separate cause explanation postulates none. So what should you count – the number of causes or the number of common causes? The epistemology just described provides an answer to this question; without an epistemological analysis, the problem is ill-formed. There are many things to count and many ways to count them.

On similarity

There is a long history of thinking about common cause versus separate cause explanations as competing hypotheses that seek to explain similarity. As just noted, Newton talks about explaining "the same effects" by postulating the same causes, and he was neither the first person nor the last to put things that way. This formulation opens a big can of worms. Which true descriptions should you use to describe events or objects? For any two things you please, you can find descriptions according to which the things are "the same" and other descriptions that say they are different. In evolutionary biology, this problem gets formulated by asking what counts as a trait. Do robins and penguins have the same trait (which we call "wings") even though the two structures exhibit many morphological differences? If you describe the forelimbs of robins and penguins in enough detail, they will be different; if you describe them in less detail, they will be "the same." Does the inference of common ancestry depend on what we choose to describe?

The Reichenbachian analysis of common cause versus separate cause explanation shows that this problem is less monumental than it might first appear. Suppose an individual *a* has trait *T* and a distinct individual *b* has *T* as well. Whether this observed matching favors common cause over separate cause depends on the models that flesh out these two hypotheses. With assumptions 1–7, you get the result that the common cause model has the higher likelihood. But it is equally true that if you change the two models, you can get the opposite result – that the matching favors separate over common cause. The same point holds if you move to a finer-grained description. Instead of using the dichotomous trait $\pm T$, you can consider an *n*-state character *C* (whose states are $C_1, C_2, \ldots, C_n$) on which $\pm T$ supervenes. If *a* and *b* are in the same *C*-state, is that evidence favoring common ancestry? If they are in different *C*-states, is that evidence with the opposite epistemic significance? To answer these questions, you need to construct a model of character evolution. This will settle the matter. The epistemic significance of observed sameness or difference is not intrinsic; it depends on background assumptions. In the history of thinking about common cause explanations, observed similarity seemed to be the golden road that leads to common cause explanations; we have seen that observed differences can get you there too. Similarity got the problem going, but it is a ladder that can be kicked away once it has been climbed.[42]

A three-way Reichenbachian distinction

I have treated the comparison of common cause and separate cause hypotheses as an epistemological issue. My question was whether common cause explanations have higher likelihoods than separate cause explanations. This contrasts with a metaphysical question: if two event types are correlated and neither causes the other, must they have a screening-off common cause? What now gets called *Reichenbach's principle of the common cause* says they must. The metaphysical and the epistemological issues are distinct. Even if Reichenbach's principle has counterexamples (as I'll argue shortly), it may still be

[42] The *principle of total evidence* dictates that you should use all the similarities and differences you know about in evaluating common cause and separate cause hypotheses. However, it makes no sense to treat different descriptors as independent of each other when they are not. If *a* and *b* match on a dichotomous trait $\pm T$ but don't match on an *n*-state character *C* on which $\pm T$ supervenes, what should you do? Since the state of *C* entails the state of $\pm T$, but not conversely, you should use the former.

true that in a wide range of cases, common cause explanations have higher likelihoods than separate cause explanations. And both Reichenbach's metaphysical principle and my epistemological thesis about likelihoods are distinct from the mathematical result I have called Reichenbach's theorem. This theorem is true, full stop.

It is easy to confuse Reichenbach's theorem and his principle of the common cause, but it also is easy to see how they differ. Suppose that *X* does not cause *Y* and that *Y* does not cause *X*. The two propositions to scrutinize are these:

> (Reichenbach's theorem) If *X* and *Y* have a common cause and assumptions 1–3 are true, then *X* and *Y* are correlated.
>
> (Reichenbach's principle) If *X* and *Y* are correlated, then they have a common cause that obeys assumptions 1–3.

The one is the converse of the other. Reichenbach's principle played no role in my conditional justification of the Reichenbachian likelihood inequality. Part of my justification for the inequality comes from assumptions 1–3, which are part of Reichenbach's theorem. However, the theorem is strictly about a common cause model; it says nothing about what a separate cause model would look like. This is why I added assumptions 4–7 to assumptions 1–3 to derive the likelihood inequality.

Although Reichenbach's principle isn't part of the epistemology I have described, it is worth a comment. People have doubted this principle based on considerations from quantum mechanics (Van Fraassen 1982), but there are problems for the principle that arise in more mundane contexts. Following the lead of Yule (1926), I constructed the following counterexample to Reichenbach's principle: if sea levels in Venice and bread prices in Britain both have tended to increase monotonically over the past two hundred years, then years with higher than average sea levels will tend to be years with higher than average bread prices (and *vice versa*). If you draw 2,400 samples from each of these processes (choosing a day at random from each month of each year and doing a pair of measurements), you will find that there is an association between the two variables in your data. If the sample is large enough and the association is strong enough, you can conclude that there is a probabilistic *correlation* here, not just a frequency *association*. And yet it is perfectly possible (even plausible) that the correlation is due to two separate causes, one of them at work in Britain, the other at work in Venice (Sober 1988, 2001, 2008b).[43] This is why I think that Reichenbach's metaphysical principle is false.[44]

It may seem strange to talk of "higher than average" sea levels and bread prices in connection with Reichenbach's principle of the common cause, but, in fact, this idea is already part of Reichenbach's examples. Consider his two geysers. You keep hourly records for a good long time on each geyser. You code spouting with the number 1 and no spouting with 0. The average score for a geyser across the time period you are considering is simply its frequency of spouting. A geyser's score for an hour in which it is spouting is higher than its average score. Although geyser spouting is a dichotomous trait whereas sea levels and bread prices are quantitative, concepts of association and correlation apply to both.

Meek and Glymour (1994, p. 1006) say that the Venice-Britain counterexample dissolves once it is realized that such dependencies are due "either to an unobserved common cause or to an unrepresentative sample or to mixing populations with different causal structures and different probability distributions." I think there is plausibly no screening-off common cause in the Venice-Britain example and that the sample isn't unrepresentative, since the association in the data will persist if new data are drawn. As for the third option that Meek and Glymour mention, Reichenbach's principle aims to tell you how to infer a causal model from data; if you already *know* what the true causal model is, there is no need for his principle. What is more, the principle says that an unconditional correlation between two variables entails that they are causally connected. It does not save the principle to point out that a separate cause model can give rise to dependencies. Arntzenius (2010) has a different reply to the counterexample; he says that a screening-off common cause can be constructed by having the common cause be a composite of facts about Venice and facts about Britain. This, I think, is a common cause only in name.[45] I have a similar reaction to the suggestion that the screening-off common cause in the Venice/Britain case is "the passage of time."

[43] For other bizarre examples of "spurious correlations," go to www.tylervigen.com.

[44] Reichenbach's principle of the common cause follows from what is now called "the causal Markov condition" (Spirtes *et al.* 2001; Pearl 2009), so my doubts about the former also are doubts about the latter.

[45] However, I do see the interest of considering a principle that says that the whole state of the universe at time t_1 renders conditionally independent the different events that happen at time t_2.

Bayesian Ockham's razor

The Reichenbachian argument concerning the likelihood comparison of a common cause (*CC*) and a separate cause (*SC*) model had the following result:

> Given Assumptions 1–7, Pr(X and Y have trait T | CC) > Pr(X and Y have trait T | SC), for each possible assignment of values to the five probabilistic parameters in the two models.

Assumption 7 (of "cross model homogeneity") allows you to draw a stronger conclusion, which holds for *any* probability distribution assigned to the values the parameters might have:

> Given Assumptions 1–7, Pr(X and Y have trait T | CC) > Pr(X and Y have trait T | SC).

Notice that you are now comparing the likelihoods of models that contain adjustable parameters. This result is something that Bayesians can accept, but it is not *very* Bayesian. It involves no consideration of prior probabilities, and it assigns no probability distribution to any of the five parameters. Likelihoodists who are critical of Bayesianism should find nothing to dislike in this result.

The approach that statisticians now call "Bayesian Ockham's razor" is, unsurprisingly, more substantive in its Bayesian commitments than this Reichenbachian inequality.[46] There still are no prior probabilities, but there are probability distributions over parameters. The basic idea can be described by using the example of the sailboat sightings that I described in Chapter 1 at the end of the section on Ptolemy and Copernicus. Earlier in the present chapter I used that example to talk about the probabilities of hypotheses. Now I'll put it to a different use. As you'll recall, during a week last summer, Susan went to Lake Mendota each day and each day she reported that she saw a red sailboat. How does this information bear on the following two hypotheses?

(ONE) There is a single sailboat that was on Lake Mendota each day during the week and no other boats were on the lake then.

(SEVEN) There are seven sailboats that were on Lake Mendota that week, one each day, and no other boats were on the lake that week.

[46] For discussion of Bayesian Ockham's razor, see Box and Tao (1973), Berger (1985), Jefferys and Berger (1992a), MacKay (2003), and Henderson *et al.* (2010). The methodology of minimum description length (MDL) is closely related to this Bayesian idea; see Grünwald (2007) for discussion.

Notice that neither of these hypotheses says anything about the color of sailboats. With the right data, you could evaluate these two hypotheses for their probabilities, but that isn't what Bayesian Ockham's razor is used to do. Rather, it is used to evaluate the likelihoods of hypotheses. But what probability does each hypothesis assign to Susan's seven observational reports? Neither hypothesis says enough to answer this question on its own. What we need are some assumptions. For example, suppose that, on any given summer day,

$$\Pr(b \text{ is red} \mid b \text{ is a sailboat on Lake Mendota}) = \tfrac{1}{10}.$$

Since Susan (S) has keen eye sight and an interest in sailboats, let's assume that

$\Pr(S \text{ thinks there is a red sailboat on the lake} \mid \text{there is a red sailboat on Lake Mendota}) \approx 1$

$\Pr(S \text{ thinks there is a red sailboat on the lake} \mid \text{there is no red sailboat on Lake Mendota}) \approx 0.$

If what S perceives on one day is independent of what she perceives on another (conditional on what's out there on the lake), these numbers have the consequence that the two hypotheses have different likelihoods:

$$\Pr(\text{On each day, } S \text{ sees a red sailboat on the lake} \mid \text{ONE}) \approx \tfrac{1}{10}$$
$$\Pr(\text{On each day, } S \text{ sees a red sailboat on the lake} \mid \text{SEVEN}) \approx \left(\tfrac{1}{10}\right)^7.$$

The more parsimonious hypothesis has the higher likelihood and the likelihood ratio is a million to 1.[47]

It is important that ONE and SEVEN do *not* say that the sailboats are red. The epistemology of the comparison changes if you change the hypotheses to:

(ONE+) There is a red sailboat that was on Lake Mendota each day during the week and no other boats were on the lake then.

(SEVEN+) There are seven red sailboats that were on Lake Mendota that week, one each day, and no other boats were on the lake that week.

Now the two likelihoods are both close to 1 and so is the likelihood ratio. The difference between comparing ONE and SEVEN and comparing ONE+ and

[47] MacKay (2003) discusses a similar example: you see a tree in a field and want to figure out whether there is one box behind the tree, or two. The assumption of cross-model homogeneity plays a role in his analysis, just as it does in my analysis of ONE versus SEVEN.

SEVEN+ shows that it isn't enough to count the sailboats that the competing hypotheses postulate; if you do that, the assessment of ONE versus SEVEN will be the same as the assessment of ONE+ versus SEVEN+.

In assessing the likelihoods of ONE and SEVEN, what would happen if we replaced $\frac{1}{10}$, 1, and 0, with other numbers? Will the likelihood ratio still be greater than unity? Its exact value will change, but as long Assumptions 1–7 are true, the more parsimonious hypothesis will have the higher likelihood. This follows from the Reichenbachian inequality I derived earlier. What does Bayesian Ockham's razor add to this qualitative result? Bayesians who use this tool want point values for the likelihoods of each hypothesis, or at least a point value for the likelihood ratio. To get this stronger result, you need stronger assumptions. This is what makes Bayesian Ockham's razor more Bayesian than the more modest Reichenbachian argument.

In discussing ONE and SEVEN, I used the following observation to evaluate the hypotheses: on each day of the week in question, *S* reported seeing a red sailboat on Lake Mendota. Let us now consider a logically weaker description of the observations: *S* reported seeing a sailboat each day and her reports agreed on the color of the sailboat sighted. This agreement is something that ONE predicts but SEVEN only accommodates. Here "prediction" does not mean *entailment*; it means that the hypothesis says the agreement is highly probable. SEVEN treats the agreement as a mere coincidence. I hope this point reminds you of the discussion of Copernicus and Ptolemy in Chapter 1.

		color of Tuesday's sailboat		
		green	red	blue
color of Monday's sailboat	green	p_1	p_2	p_3
	red	p_4	p_5	p_6
	blue	p_7	p_8	p_9

Proponents of Bayesian Ockham's razor often point out that models that contain fewer adjustable parameters postulate a narrower range of possibilities over which probabilities must be distributed (Jefferys and Berger 1992a; Mackay 2003). ONE and SEVEN don't explicitly mention adjustable parameters, but given Susan's seeing red, we can bring this detail about her observations to bear on how we formulate the competing models. Since the page you are now reading is two-dimensional, I'll shift from Susan's seven reports to what she says on Monday and Tuesday. We therefore need to consider

(ONE-MT) There is a single sailboat that was on Lake Mendota on both Monday and Tuesday and, on each day, it was the only boat on the lake; there is a color c that that sailboat has.

(TWO-MT) There was a single sailboat on Lake Mendota on Monday and a different single sailboat out there on Tuesday; there is a color c_{M} that Monday's sailboat has and a color c_{T} that Tuesday's sailboat has.

ONE-MT has one adjustable parameter (c) and TWO-MT has two (c_{M} and c_{T}). To talk about the likelihoods of these two models, each model must assign values to the nine cells in the accompanying 3×3 table. For simplicity I assume that there are just three possible sailboat colors.

The models agree that the nine p's in the table sum to one. The ONE-MT model says that the only entries that can have positive probability are the ones on the main diagonal (p_1, p_5, and p_9); off-diagonal cells in the table have probabilities of zero. TWO-MT involves no such restriction; it must assign values to all nine cells of the table. So TWO-MT says there are more possibilities than ONE-MT admits. These structural facts do not entail that ONE-MT must assign a higher value to p_5 than TWO-MT does. However, with additional assumptions, you can obtain that result. For example, if you invoke the principle of indifference, you can have ONE-MT assign probabilities of $\frac{1}{3}$ to each main diagonal cell and TWO-MT assign probabilities of $\frac{1}{9}$ to those same cells. Another possibility arises if the two models agree on what the marginal distribution is. Perhaps you have good frequency data that supports the conclusion that on any given day the probability of a Lake Mendota sailboat's being green is g, its probability of being red is r, and its probability of being blue is b (and that none of these is equal to 0 or 1). This entails that ONE-MT will assign higher values to the entries in the table's main diagonal than TWO-MT assigns, since $g > g^2$, $r > r^2$, and $b > b^2$. This second strategy is the one that typically gets used when Bayesian Ockham's razor is put to work. In my view, this argument based on the marginal probabilities is stronger than the one that invokes the principle of indifference. Both arguments are responses to the fact that it isn't logically inevitable that ONE-MT assigns higher values to the main diagonal in the table than TWO-MT does. If sailboats were chameleons (changing their color from day to day), all bets would be off.

I argued earlier in this chapter that Bayesianism provides natural and unproblematic solutions to some inference problems but strains at the bit when it addresses others. Assigning a prior probability to Susan's having tuberculosis is on one side of this divide while assigning a prior probability to Newton's law of gravitation is on the other. The same division of cases arises in connection with Bayesian Ockham's razor. Computing the likelihoods of ONE-MT and TWO-MT is pretty straightforward. Now let's turn to a harder problem about likelihoods. Beginning in the nineteenth century, astronomers observed that each time the planet Mercury goes round the Sun, the point at which the planet is closest to the Sun changes its location a bit. This "precession of the perihelion" was something that Newton's laws of motion and gravitation (*N*) could account for, but only partly. Once the gravitational influences of the Sun and the known planets were taken into account, some of the precession was explained, but there was 43 seconds of arc per century that remained a puzzle. Some physicists proposed that there exists a heretofore unobserved planet ("Vulcan") between Mercury and the Sun that exerts a gravitational influence on Mercury. When efforts to observe the planet failed, Vulcanism faded as a viable option. Other physical mechanisms were then proposed, all within the framework of *N*. Instead of a single solid planet between Mercury and the Sun, maybe there are "matter rings." Or maybe the Sun isn't a perfect sphere, and that is the culprit. None of these ideas panned out. The one suggestion that remained afloat involved modifying *N*. Simon Newcomb (1895) suggested that Newton's inverse square law of gravitation be replaced by a law in which the exponent is $2+\varepsilon$ where ε is an adjustable parameter. Call this new theory N^*. The story then took a dramatic turn when Einstein proposed his general theory of relativity (*GTR*) in 1915. Einstein's theory predicted a precession that was remarkably close to the observed value. This was widely viewed as strong evidence favoring Einstein's *GTR* over Newton's *N*. But what about Newcomb's suggestion that the law of gravitation should have an exponent of $2+\varepsilon$? What does N^* predict?

Jefferys and Berger (1992a) answer this question by assuming that the value of ε has a normal probability distribution that is centered on 0. The standard deviation of the distribution can be larger or smaller; that turns out not to matter. Regardless of what the standard deviation is, the following likelihood inequality holds:

$$\begin{aligned}&\Pr(\text{Mercury's precession} \mid GTR)\\&\quad > \Pr(\text{Mercury's precession} \mid N^* \,\&\, \varepsilon \text{ is normally distributed with a mean of } 0).\end{aligned}$$

Jefferys and Berger show, in addition, that the likelihood ratio of *GTR* to N^* (with ε normally distributed and centered on 0) is at least 27; the exact value depends on the standard deviation you choose. They see this likelihood analysis as a victory for Bayesian Ockham's razor. The problem with their argument is that there is no reason why the probability distribution of ε should be centered on zero (Sober and Forster 1992). True, Newton's theory N set $\varepsilon = 0$. But why should Newcomb's N^* retain this point value as the mean of a new distribution? Jefferys and Berger (1992b, p. 213) answer this question as follows:

> We assert that, prior to seeing the Mercury data, one would have no reason to differentiate between positive and negative values of ε – hence symmetry. And *a priori*, we feel that most scientists would favor smaller – as opposed to larger – values of ε, for the simple reason that no specific larger value is distinguished by the theory, and even modest values of ε would certainly have been observed in other experiments. It is crucial to remember that we are talking about reasonable "prior" opinions – beliefs that a reasonable scientist would hold prior to seeing the data concerning Mercury's perihelion.

I sense an appeal to the principle of indifference in the first sentence; my reply is that having no reason to assume asymmetry is not a reason to assume symmetry. I also balk at thinking that the problem is to find a distribution for ε that characterizes how one should think about Newcomb's N^* "prior to seeing the Mercury data." Newcomb suggested his N^* in response to data on Mercury's precession. Why not use that data to construct a distribution for ε?

Bayesian Ockham's razor works better on the sailboat problem than it does on Newcomb's N^*. You can gather frequency data on Lake Mendota sailboat colors and use that evidence to ground assumptions about the marginal probabilities in the 3×3 table. In contrast, it is unclear how observation or theory would allow you to justify a value for the average likelihood of N^*.

Frequentism and adjustable parameters

In the preceding section, I investigated the circumstances in which a common cause explanation has a higher likelihood than a separate cause explanation. It now is time to remove our Bayesian hats and step into the world of frequentism, leaving the law of likelihood behind. In this brave new world, Ockham's razor has an epistemic justification, but the justification does not derive from its mirroring likelihoods. In fact, the typical situation in this new setting is that parsimony and likelihood *clash*.

Consider a simple example: you drive south from Madison on a country road and stop your car when you see two fields of corn – one on the left side of the road, the other on the right. Each field contains thousands of corn plants. Let's call the average height in the first field m_1 and the average height in the second m_2. You sample a hundred plants from each field and compute the average heights in your two samples; the two sample means are s_1 and s_2. Suppose their values are 56 and 52 inches, respectively. You want to use these data to evaluate two models:

(NULL) $m_1 - m_2 = 0$.
(DIFF) There exists a number d such that $m_1 - m_2 = d$.

I call the first model NULL because it says that there is no difference in the two mean heights.[48] My label for DIFF is a bit of a misnomer, however, since this model does not *require* that the mean heights differ; it merely *allows* that this might be so. DIFF contains a single adjustable parameter (d) whose value can be estimated from the data. The maximum likelihood estimate of d is 4 inches; this is the estimate that makes the observed difference between the sample means most probable.[49] Notice that DIFF has a flexibility that NULL does not possess. When you observed the difference in sample means, it might have comported very poorly with what NULL says; in contrast, any observed difference in sample means can be accommodated by DIFF.

DIFF may strike you as a tautology. Well, it is a *near*-tautology. DIFF entails that there are two fields of corn plants, and that isn't a tautology. But the rest of what DIFF says sounds pretty empty – if there are two fields of corn, then the two fields have mean heights, and there is a number that represents the difference in mean heights. How boring this model is! Even so, DIFF can be

[48] Notice that my use of "null" has nothing to do with the motives and beliefs of investigators. This contrasts with using "null hypothesis" to refer to the hypothesis you want to refute or to the one you should believe unless evidence is produced to the contrary.

[49] That is, $\Pr[(s_1 - s_2) = 4 \mid (m_1 - m_2) = 4] > \Pr(s_1 - s_2) = 4 \mid (m_1 - m_2) = x]$, for any $x \neq 4$. And here's a fine point: the probability of this precise difference in the sample means is zero under the hypothesis that $(m_1 - m_2) = 4$; we need to talk about its probability *density* or about the probability that the observed difference is *approximately* 4 inches.

used to make predictions and these predictions may differ from the ones that issue from NULL. DIFF makes predictions in the following sense. If you replace the adjustable parameter d in this model with that parameter's maximum likelihood estimate, the result is a fitted model, which I'll call L(DIFF). DIFF is an infinite disjunction, covering all the many values that d might have. L(DIFF) is the likeliest disjunct in this disjunction. L(DIFF) makes a prediction about what new data drawn from the two populations will be like. DIFF makes a prediction about new data by being fitted to the data at hand. NULL also makes a prediction about new data, but it does not need to be fitted to the old data to do so.

How should you evaluate these two models? In terms of fit-to-data, NULL can't do better than DIFF. The two models tie precisely when $(s_1 - s_2) = 0$; otherwise, DIFF will fit the data better. Fit-to-data reflects likelihoods. In our example, $(s_1 - s_2) = 4$ inches and L(DIFF) says that the parameter $d = 4$. This fitted model confers on the observation a higher probability than NULL confers. Another comparison we might make of the two models concerns their probabilities of being true. Since NULL entails DIFF, NULL can't be more probable, no matter what the data are. So in terms of fitting the old data and in terms of probability, NULL can't be better than DIFF. But what about the model's ability to accurately predict new data when fitted to old? Let us call this property of the model its *predictive accuracy* (Forster and Sober 1994). When a model is repeatedly fitted to old data sets and then judged by how well it predicts new ones, there will be some variation in its performance. Predictive accuracy is an expectation; it describes what the *average* performance in this prediction task would be. Predictive accuracy isn't the same as fit-to-old-data, nor is it the same as the model's probability of being true.

Notice that NULL is more parsimonious than DIFF if we measure parsimony by counting adjustable parameters. Model builders in the different sciences know that fit-to-data can be improved by making a model more complex, but they also know from bitter experience that highly complex models often do a poor job of predicting new data when fitted to old. When a complex model founders in this way, it is said to have *over-fitted* the data. To avoid over-fitting, scientists attend to how parsimonious a model is; fit-to-data is relevant to estimating how predictively accurate a model will be, but it is not the whole story.

In a paper published in 1973, the Japanese statistician Hirotugu Akaike (1927–2009) made the role of parsimony explicit by proving a remarkable theorem about predictive accuracy:

> *Akaike's theorem*: An unbiased estimate of model M's predictive accuracy is $\log\{\Pr[\text{data} \mid L(M)]\} - k$.[50]

Akaike proved that this theorem follows from a set of assumptions; I'll describe these soon. For now, let's get clear on what the theorem says. The theorem says there are two considerations that should influence your estimate of predictive accuracy: (i) the logarithm of the likelihood of the fitted model L(M), and (ii) k, which is the number of adjustable parameters that the model contains. Akaike's result led to the formulation of AIC, the Akaike Information Criterion (AIC), which is a method for scoring models. A model's AIC score is the quantity $\log\{\Pr[\mathit{data} \mid \mathrm{L}(M)]\} - k$. Since log-likelihoods go up as likelihoods go up, a model's score is improved by its fitting the data better. But the model's score is also improved if the model has a low value for k. AIC imposes a penalty for model complexity. In our comparison of NULL and DIFF, DIFF does better in terms of fit-to-data, but worse in terms of parsimony. So which model has the better AIC score? Since DIFF has one more adjustable parameter than NULL, the answer is

$$\text{AIC(DIFF)} > \text{AIC(NULL) if and only if}$$
$$\log\{\Pr[\text{data} \mid L(\text{DIFF})]\} - \log\{\Pr[\text{data} \mid L(\text{NULL})]\} > 1.$$

It is logically inevitable that the difference in the log-likelihoods is positive. It is not inevitable that the difference is greater than 1; that depends on the data. For DIFF to have the better AIC score, it must fit the data *sufficiently* better than NULL does to overcome the fact that DIFF is more complex.

I formulated NULL and DIFF so that each addresses the *difference* in the mean heights of the two populations. These two models have 0 and 1 adjustable parameters, respectively. I could just as easily have described the models by having them be about the mean heights in each of the two populations. The ONE model says that there is a single number that characterizes the mean heights in the two populations. The TWO model says that there is a mean height in the first population and possibly a different mean height in the second. The ONE model *unifies* the two populations whereas the TWO model

[50] The quantity that Akaike talked about is $2k - 2\log\{\Pr[\text{data} \mid L(M)]\}$; this is -2 times the quantity I cite. Akaike's quantity is an estimate of a model's predictive *in*accuracy. In Akaike's formulation, models are better if they have smaller AIC scores; in the formulation I am using, models are better if their AIC scores are bigger. This, of course, is just a terminological difference.

does not. Unification is achieved by the simple device of having the same adjustable parameter apply to both populations.

Although the difference in parsimony between ONE and TWO is modest, it is easy enough to tweak the example to make the difference larger. Instead of there being two fields of corn, let there be ten. Now consider a model that unifies the ten fields by saying that they have the same mean height. A competing model fails to unify, since it assigns a different adjustable parameter to each field. Now there is a bigger difference in how parsimonious the two models are (1 versus 10 adjustable parameters). The ten-parameter model must fit the data *way* better than the one-parameter model does if the more complex model is to receive the better AIC score.

Akaike's theorem gave rise to AIC, but the two should not be confused. The theorem is a theorem; it follows mathematically from the assumptions that Akaike identifies. However, this mathematical result doesn't automatically settle the question of whether you should use AIC to estimate predictive accuracy. It isn't just that there is room to wonder whether Akaike's assumptions are satisfied in a given real-world inference problem. In addition, there is the question of whether an estimator's being unbiased is a sufficient reason to use it. I think it is pretty clear that lack of bias is insufficient. An unbiased estimator of a quantity is "centered" on the quantity's true value. Your kitchen scale is an unbiased estimator of the weight of an apple if repeatedly weighing the apple would tend to yield estimates whose average is the apple's true weight. An unbiased estimator can spit out estimates that differ from each other; some are too high while others are too low. It is the long-run average that matters.[51] This means that unbiased estimators may differ in their *variances*; some may tend to produce a wider spread of estimates than others. Now consider two estimators: the first is unbiased and has a very large variance, while the second is just a little bit biased and has a very small variance. If you are going to weigh your apple just once, you may want to use the second

[51] More formally, an estimator is a function $f(-)$ that takes observations $O_1, O_2, \ldots, O_n$ as inputs and outputs an estimate. An estimator is unbiased if the expected value of $f(O_1, O_2, \ldots, O_n)$ is the true value of the quantity being estimated, for any value for n. What Akaike in fact proved about AIC is something weaker than what I say above; he showed only that AIC is an *asymptotically* unbiased estimator. That is, as the number of observations is increased, the bias in AIC's estimate of a model's predictive accuracy converges on zero.

estmator. This is enough to show that the fact that an estimator is unbiased does not suffice to show that you should use it. So Akaike's result, by itself, does not suffice to justify your using AIC. This, of course, leaves it open that there are other mathematical results that close the gap.

On what assumptions does Akaike's theorem depend? One of the assumptions is that new data are drawn from the same (unknown) underlying reality as old. This deserves to be called a uniformity of nature assumption, with a nod to Hume (Forster and Sober 1994). This assumption has two parts, which can be described by considering the example of the kettle on your stove where the curve-fitting problem was to evaluate LIN_e and PAR_e. First, there is the idea that all data sets are produced by the same true (but unknown) curve. The other part of the uniformity of nature assumption concerns how you obtain your data. You obtain your data by choosing different x-values and then seeing what the y-values are that go with them. There is one probability that you'll choose x-values in this interval, another probability that you'll choose x-values in that interval, and so on. The second part of the uniformity of nature assumption says that different data sets have their x-values chosen from a single probability (density) distribution. This second part has important implications concerning the kind of problem that AIC is able to address. Suppose all your present data are chosen by looking at x-values that are between 0 and 100 and you want to decide which model will do better at predicting y-values in some very different range of x-values (say $1{,}000 < x < 2{,}000$). The assumptions that go into proving that AIC is an unbiased estimator do not apply to such problems of *extrapolation* (Forster 2000).[52]

A second assumption in the proof that AIC is unbiased concerns "regularity;" repeated estimates of the value of an adjustable parameter in a model must form a bell-shaped distribution.

[52] The connection of AIC with the distinction between interpolation and extrapolation can be seen by considering the relation of AIC to a different model selection criterion, cross validation. In take-one-out cross validation, you have n data points and you fit each candidate model to $n-1$ of them and then see how accurately the fitted model predicts the datum that you omitted. You do this n times, with a different data point omitted each time. The average accuracy that a model has in this process is its cross-validation score. This procedure makes no overt use of parsimony, but it turns out to be asymptotically equivalent to AIC (Stone 1977). Notice that in curve fitting, $n-2$ of the n steps in the cross-validation procedure involve problems of interpolation.

A third assumption that is needed to prove Akaike's result is that one of the models being considered is true.[53] This assumption can be relaxed in practice if one cares only about ordering models for their predictive accuracy (and not about estimating the absolute value of each model's predictive accuracy) and one of the candidate models has a special case that is *close* to the truth. It is unproblematic to assume that one of the candidate models is true in the case of NULL and DIFF, but the assumption is unsatisfactory in many model selection problems. Scientists often find themselves considering a set of candidate models that all contain *idealizations*; this means that all are false. This is why the TIC criterion (named for Takeuchi 1976) is theoretically important. TIC, like AIC, subtracts a "penalty" from the log-likelihood of the fitted model. However, instead of k, the TIC penalty is provided by a "matrix trace function," which can be estimated from the data, though doing so typically requires a very large data set. The unbiasedness of TIC can be derived without the assumption that one of the candidate models is true. TIC reduces to AIC when one of the candidate models is true. Burnham and Anderson (2002, p. 65, p. 96, p. 369) say that AIC is a good approximation of TIC when the task is to compare candidate models; they say that k, the number of adjustable parameters, is a "parsimonious" estimate of the matrix trace function. Parsimony (in the sense of number of adjustable parameters) is not part of the TIC apparatus.

As noted, applying AIC to a model requires that you obtain maximum likelihood estimates of the model's adjustable parameters. In particular, the maximum likelihood estimate for each parameter must be unique. To see what this means, suppose you have a single data point from your kitchen kettle and want to use AIC to ascertain how predictively accurate LIN will be. There are infinitely many straight lines that pass through this one data point. What would it mean to fit LIN to this one datum and then use "the" fitted model to predict a new datum? There is no such single fitted model picked out here, so AIC does not apply. Statisticians say of such cases that the model is not "identifiable." This has practical implications for how complex a model can be if AIC is to be able to score it. If you have 500 observations from your kettle, then a super-complex model that contains, say, 1,000 adjustable parameters will be beyond your reach. Statisticians often suggest that AIC be applied to a model only when you have at least forty observations for every

[53] Forster and Sober (1994, p. 29) neglect to mention this important assumption.

adjustable parameter; the minimum is more modest if you use a modification of AIC called $\mathrm{AIC_c}$ (Burnham and Anderson 2002, p. 66). Both AIC and $\mathrm{AIC_c}$ impose a limit – not on how complex *nature* can be, but on how complex a *model* can be and still be evaluable given the data you have at hand. No such impediment is to be found in Bayesianism, where it is the average likelihood of a model, not the likelihood of the most likely member of the model, that matters.

How many causes for a single effect?

I now want to apply model selection ideas to a simple problem that is central to thinking about Ockham's razor (Forster and Sober 1994). Let there be a single observed effect. Is it better to explain this by postulating two causes or just one? Ockham's razor claims that one is better than two. What does AIC say about this claim?

As an example, let's consider how teacher experience and class size in elementary school affect a child's probability of being admitted some years later to college.[54] The probability of admission to college under each of four possible "treatments" is represented in the accompanying two-by-two table. In this table, b is the baseline probability, t is the difference made by having experienced teachers, s is the difference made by having small class size, and i is an interaction term that I'll explain shortly. Each of s, t, and i can be positive, negative, or zero.

Pr(admission to college | ±class size & ±teacher experience)		
	Small class size	Large class size
Teachers experienced	$b+t+s+i$	$b+t$
Teachers inexperienced	$b+s$	b

Now let's consider five possible models. Each model assigns a zero or a question mark to each of s, t, and i. When a model assigns a question mark to a parameter, this means that the parameter is an adjustable parameter in that model. The simplest model has no adjustable parameters because it assigns zeroes across the board:

[54] For an overview of studies of the effect of class size reduction on educational outcomes, see Whitehurst and Chingos (2011).

(Null)	$t = 0$	$s = 0$	$i = 0$
(Teacher experience only)	$t = ?$	$s = 0$	$i = 0$
(Class size only)	$t = 0$	$s = ?$	$i = 0$
(Additive two causes)	$t = ?$	$s = ?$	$i = 0$
(Interactive two causes)	$t = ?$	$s = ?$	$i = ?$

If parsimony means paucity of adjustable parameters, you can assess how complex a model is by counting question marks.

You can test these models against each other by doing an experiment in which children are randomly assigned to each of the four treatment cells for their elementary school experience. However, you may not need to contemplate running this morally questionable experiment if nature has already done this for you by inducing chance fluctuations in class size and in teacher experience in several schools over several years. Either way, you need to determine, years later, the frequency with which students are admitted to college in each cell of the accompanying table.

Freq(admission to college | ±class size & ±teacher experience)		
	Small class size	Large class size
Teachers experienced	f_1	f_2
Teachers inexperienced	f_3	f_4

The interactive model will be able to fit these data perfectly, regardless of what the four frequencies turn out to be. Not so for the other models: for each of them, there are possible data sets that the model can fit only imperfectly. In general, the more adjustable parameters, the better able a model is to fit the data. Model selection criteria like AIC give weight to both the likelihoods of fitted models (which is what fit-to-data reflects) and to parsimony in assessing which models should be expected to be more predictively accurate than which others. Depending on the data, AIC will conclude that the simplest model is the best, the worst, or somewhere in between. As usual, parsimony is relevant, but it is not the whole story.

The additive two-cause model has this name because of what it means for that model to set the interaction term i equal to 0. Consider a student who occupies the lower right-hand cell. This student is in large classes and has inexperienced teachers. If that individual had been placed in small classes taught by inexperienced teachers, the effect would be to change the student's probability of getting into college by s. On the other hand, if the student in the lower right-hand cell had been in large classes taught by experienced teachers, the effect would be to change the student's probability by t. Question: what difference in the student's probability of gaining admission to college would there be if the student in the lower-right cell had been placed in the upper-left cell? If the model is additive, the effect is to change the student's probability by $s+t$. The effect of changing both factors is just the sum of changing one factor without changing the other and changing the other factor without changing the one. If the interaction term i is non-zero, the system is non-additive; in technical parlance, there is an "interaction" of the two causes.

Ockham's razor is often described by saying that postulating fewer causes is "better" than postulating more. Although model selection makes sense of this idea, the present example brings out a respect in which this formulation of the razor is imperfect. Both the additive and the interactive models postulate two causes. Yet, they are unequally parsimonious if we assess parsimony by counting adjustable parameters. From the point of view of model selection theory, these two models differ from each other in the same way that the additive two-cause model differs from each of the single-cause models. What matters is number of adjustable parameters, not number of causes (Forster and Sober 1994, p. 16).[55]

Earlier in this chapter, I analyzed Williams's (1966) parsimony argument against group selection by understanding it as a claim about prior probabilities. The present analysis of why a one-cause model can be better than a two-cause model provides a new perspective on his argument. Understood in terms of prior probabilities, the parsimony argument against the hypothesis that a trait is a group adaptation does not involve looking at any data that are specifically about the trait in question. In contrast, a model selection format allows you to compare a model that postulates both individual and group selection to explain a trait's frequency in a population with a model that postulates individual selection only. Now data on the trait are central and priors play no role. The greater parsimony of the model that refuses to invoke group selection is relevant, but fit-to-data matters too (Sober and Wilson 1998).[56] This interpretation of Williams's parsimony argument has the additional advantage that it explains the first part of his ground-rule, that

[55] And if our fundamental focus is on model *comparison*, what matters is the *difference* between the number of parameters in one model and the number in another; the absolute values aren't key.

[56] This, in effect, is the methodology that Lewontin and Dunn (1960) put to work in their argument that group selection influenced the evolution of a segregator-distorter gene in the house mouse.

adaptation is an “onerous concept.” A hypothesis of neutrality (according to which all traits have the same fitness value) is a null model and it therefore is simpler than a model that invokes natural selection (and thereby claims that there is variation in fitness).[57]

Bayesian model selection

I so far have discussed model selection theory as a frequentist endeavor, but there is a subject out there called Bayesian model selection theory. The difference between the two is not that Bayesians assign prior probabilities to models (something that frequentists abhor). Rather, the difference is that model selection criteria like AIC have the goal of estimating predictive accuracy whereas Bayesian model selection aims at estimating the likelihood of a model. To understand what is challenging about this latter task, let’s return to the DIFF model of the two fields of corn. When you observe that the sample means differ by 4 inches, how probable is that outcome according to DIFF? That all depends. If DIFF asserted that the means in the two fields differ by 4.1 inches, the model would say that the observation is pretty probable. However, if DIFF asserted that the fields differ by 50 inches, then the model would say that the observed difference of 4 is very improbable. In fact, the model doesn’t assert either of these things; it leaves open how much the two fields will differ. This means that you need to average over a great many possibilities to compute the likelihood of DIFF:

$$\begin{aligned}&\Pr(\text{sample means differ by 4 inches} \mid \text{DIFF})\\ &\quad = \sum\nolimits_i \Pr(\text{sample means differ by 4 inches} \mid \mathrm{D}_i)\Pr(\mathrm{D}_i \mid \text{DIFF}).\end{aligned}$$[58]

Here D_1, $\mathrm{D}_2, \ldots \mathrm{D}_\mathrm{n}$ are the possible differences in mean height that might characterize the two fields. To compute the likelihood of the model, you need to know how probable each D_i is according to DIFF.

[57] It is interesting to compare Williams’s thesis that neutrality should be one’s “default assumption” with Mayr’s (1983, p. 326) contention that it is natural selection that should be presumed innocent until proven guilty. Mayr was appealing to the accumulation of past experience, wherein traits of previously unknown function were repeatedly discovered to make a fitness difference. Williams’s outlook fits into frequentism, Mayr’s into Bayesianism.

[58] This summation should of course be an integral since difference in mean heights is a continuous quantity.

Schwarz (1978) identified a set of assumptions that allowed him to derive an estimate of the average likelihood of a model. He called this the Bayesian Information Criterion (BIC):

$$\mathrm{BIC}(M) = 2\log[\Pr(\mathrm{data} \mid L(M)] - (k)\mathrm{Log}(n).$$

Schwarz shows that BIC(M) is an unbiased estimate of $2\log[\Pr(\mathrm{data}|\, M)]$.[59] Here k is the number of adjustable parameters in the model and n is the number of observations you have in your data set. BIC resembles AIC in two respects; it takes account of the likelihood of the likeliest member of the model and it subtracts a penalty term. Notice that the penalty term in BIC is bigger than the one in AIC. This has led many to think that the two criteria are in competition, and therefore to ask which one is better. In fact, this is an apples and oranges question, since the two criteria have different goals (Wasserman 2000). AIC does not aim at estimating the average likelihood of a model! Even so, it is striking that both AIC and BIC penalize models for their complexity.

Without going into the details of what assumptions are involved in Schwarz's derivation or in other more recent Bayesian forays into the same terrain, we can identify a kind of assumption that must figure in any such derivation. Frequentists usually are unwilling to commit to probability distributions over the parameters that a model introduces. In the case of DIFF, they have no problem talking about the likelihood of L(DIFF). But to estimate the likelihood of DIFF, you need to know how probable the different values of $\mathrm{D_i}$ would be if DIFF were true. The definition of BIC does not overtly represent the need for this information, and this may foster the illusion that using BIC does not require it. But there is no getting around the fact that assumptions of this sort are necessary if you are going to estimate the likelihood of a model that contains adjustable parameters. You can't get something from nothing.

Modern Bayesians don't pretend that the answer to this question about DIFF can be known *a priori*. Given information about the mean heights in other pairs of adjacent corn fields, you may be willing to commit to assumptions that allow you to calculate the likelihood of DIFF. If you have evidence that backs up these assumptions, I have no qualms about doing the Bayesian

[59] Just as I did in my description of AIC, I have multiplied Schwartz's formulation by -1. The effect is that bigger BIC scores mean higher likelihoods.

calculation of the model's average likelihood. What bothers me is simply making up assumptions that allow the Bayesian calculation to go forward.

Notice that the BIC scores for NULL and DIFF, given the *n* observations you made, are the same regardless of whether the two models are about the two fields of corn on opposite sides of a road south of Madison or are about the average weights of the stones in two piles of rock in China (where the observations are 50 and 54 kilograms). The fact that BIC treats these two problems as "the same" indicates that substantive assumptions are at work in each. Wasserman (2000, p. 99) says that using BIC to compare two models will yield results that approximate what would happen if you used a prior distribution for the values of adjustable parameters that is due Harold Jeffreys (a "Jeffreys prior") when a constant is assigned a certain value. Wasserman says that this prior suffices to justify the BIC ordering, not that it is necessary. Even so, his point serves to emphasize the fact that BIC depends on assumptions about the prior distribution.[60]

The world of model selection

The preceding pages provide only a brief glimpse of the ideas that have been developed in the statistical literature on model selection. I mentioned AIC, TIC, and BIC, but there are additional criteria out there.[61] The mathematical underpinnings of these different criteria are a rich area of research. Many practitioners believe that choosing a model selection criterion depends on

[60] The Jeffreys prior for a parameter is often "improper," meaning that the sum of the probabilities for all possible values of the parameter is greater than one. For example, consider a parameter that describes the mean height in a population of corn plants where the variance in heights in the population is known. The Jeffreys prior assigns a value of 1 to every mean height from 0 to infinity. When Bayesians use improper priors, they do so because they want an "informationless" prior that will allow them to obtain posterior probabilities. Improper priors can't represent rational degrees of belief.

[61] Statistical (or "machine") learning theory is another framework in which parsimony (as measured by the VC-dimension) is relevant to assessing a model's predictive accuracy. It differs dramatically from the Bayesian and frequentist ideas I have described in this chapter. VC stands for Vladimir Vapnik and Alexey Chervonenkis, the theory's inventors. The VC dimension of a model isn't the same as the number of adjustable parameters, but both penalize models for being "too flexible" in their ability to accommodate data. Vapnik was inspired to develop the theory by reading Popper's treatment of degree of falsifiability. For details, see Vapnik (2000), von Luxburg and Schölkopf (2009), Kulkarni and Harman (2011), and Cherkassky (2013).

the specifics of the problem one wishes to address. These details, interesting though they are, are beyond the scope of this book; see Forster (2000) for further discussion. What is important here is something that several model selection criteria have in common. Whether they aim at estimating a model's predictive accuracy or its likelihood, they agree that parsimony, as measured by the number of adjustable parameters in a model, is relevant to making those estimates. It isn't just that parsimony *seems* relevant; there are mathematical arguments for why it *is* relevant. Ockham's razor is alive and well in statistics.[62]

How the two parsimony paradigms differ

In this chapter I have described two "paradigms" in which the epistemic relevance of parsimony can be made clear. In the first, more parsimonious theories have higher likelihoods. In the second, parsimony is relevant to estimating a model's predictive accuracy.[63] According to the first paradigm, parsimony is epistemically relevant because the data favor simpler theories over theories that are more complex when "favoring" is interpreted in terms of the law of likelihood. According to the second, parsimony is relevant because the number of adjustable parameters in a model helps you estimate its predictive accuracy. These two sorts of epistemic relevance are dramatically different. In the first, greater parsimony goes *hand-in-hand* with greater likelihood. In the second, more parsimonious models, when fitted to the data, often confer *lower* probabilities on the data at hand; here parsimony and likelihood *clash*.

In the Reichenbachian argument, assumptions 1–7 entail that the common cause hypothesis has a higher likelihood than the separate cause hypothesis.

[62] Two foundational questions are worth mentioning here, though I won't explore them. The first concerns what a parameter is. When we say that a linear model has two adjustable parameters (the y-intercept and the slope), why shouldn't we say, instead, that it has just one, since a pair of numerals can be represented as a single numeral by interweaving? The other problem is discussed in Forster and Sober (1994) under the heading of the "subfamily problem." Although the NULL model of the two fields of corn may receive a better AIC score than DIFF does, NULL can't score better than *L*(DIFF). So how does AIC explain why we should use NULL rather than *L*(DIFF) to predict new data? Both questions are discussed in Sober (2008b, pp. 99–102 and pp. 93–95).

[63] I mentioned a third paradigm, according to which more parsimonious hypotheses have higher prior probabilities, but I think that idea plays third fiddle to the other two.

The two hypotheses contain adjustable parameters, but there is no need to estimate their values. The likelihood inequality follows from the assumptions, regardless of what values the parameters have. AIC is different. The models it evaluates *do* contain adjustable parameters whose values must be estimated from the data. AIC never computes the likelihood of a model, but it does compute the likelihood of a fitted model; it is the value of Pr[data | L(M)] that affects the model's AIC score, not the value of Pr(data | M). However, computing the value of Pr[data | L(M)] is not an end in itself in AIC. The likelihood of the fitted model is relevant to estimating the model's predictive accuracy, but so too is the number of adjustable parameters.

How do Bayesian Ockham's razor and BIC relate to the two parsimony paradigms? Both concern likelihoods, and so they fall under the first heading. BIC is a device for estimating the likelihood of a model that contains adjustable parameters, a project in which AIC takes no interest. Bayesian Ockham's razor is a bit different from BIC; it tells you how to *calculate* the likelihood of a model, not just *estimate* its value. Whereas BIC has a likelihood term and a penalty term for model complexity, Bayesian Ockham's razor does not have these two components. Rather, it describes a technique for computing average likelihoods that has the consequence that more parsimonious models will often have higher likelihoods.

The two parsimony paradigms differ in the way they provide parsimony with an epistemic justification, but they differ in another way as well. They differ in how one should think about the alternative hypotheses with which a given hypothesis competes. The law of likelihood does not impose any restrictions on which hypotheses can be "competitors," but it is entirely natural to require that competing hypotheses be incompatible with each other. If a coin lands heads eleven times in twenty tosses, it makes perfect sense to compare the hypothesis that the coin is fair ($p = 0.5$) with the hypothesis that the coin is highly biased in favor of heads ($p = 0.9$). It is decidedly odd to compare the first of these hypotheses with the hypothesis that the coin's probability of landing heads is somewhere between 0.4 and 0.6. Yet, in explaining the ABCs of AIC, I considered the models NULL and DIFF. These models are nested; one of them entails the other, and thus they are compatible.[64] Shouldn't proper

[64] There are other frequentist ideas that allow nested models to be compared. An example is the likelihood ratio test. And it doesn't just *permit* models to be nested; it *requires* that they be. The likelihood ratio test implicitly assigns a role to parsimony, in that a model with n adjustable parameters is rejected only if a model with $n + 1$ parameters fits the data, not just better, but *significantly* better.

competitors be incompatible with each other? Well, if the goal is to find hypotheses that are true, maybe they should be. But the goal of AIC is to estimate predictive accuracy. Since nested models can differ in their predictive accuracies, there is nothing amiss in treating them as competitors. In discussing Popper's idea of corroboration, I held out the hope that a logically stronger hypothesis can be "better" than a logically weaker hypothesis. AIC makes room for this Popperian idea.

AIC does not require competing models to be nested. For example, instead of comparing NULL and DIFF, we could have compared the NULL hypothesis with

(DIFF*) There exists a number $d \neq 0$ such that $(m_1 - m_2) = d$.

I noted earlier that NULL can't fit the data better than DIFF; at best, they tie. However, NULL *can* fit the data better than DIFF*, but the difference will be arbitrarily small. If the observed difference in the sample means is exactly 0 inches, NULL fits this observation perfectly, whereas L(DIFF*) will say that the difference in the population means is 0.000000 . . . 0000000000001. NULL is now just a *tiny* bit more likely than L(DIFF*). However, the questions remains of whether NULL or DIFF* will be more accurate predictors, and that isn't settled by the likelihoods.[65] From the point of view of using AIC to estimate predictive accuracy, it makes no difference whether you compare NULL and DIFF, or NULL and DIFF*.

It may seem almost pointless to gather data if you are virtually certain from the get-go that NULL is false and DIFF is true. However, there can be a real point to using AIC and other model selection criteria in this circumstance. With respect to the two fields of corn, I am virtually certain that NULL is false. I am as certain about this as I am about just about anything. DIFF, on the other hand, is a near tautology; I am very sure that it is true. If the goal were to separate true models from false ones, there would be little reason to look at data. But there is a substantial reason to do so if the goal is predictive accuracy. AIC and other model selection criteria that aim at estimating predictive accuracy thus provide an opening for an instrumentalist philosophy of science. Instrumentalism says that the goal of scientific inference is to find

[65] Notice that the AIC scores of DIFF and DIFF* will be indistinguishable in practice, which means that it doesn't matter which one you use if you want to compare either with NULL. Nested and non-nested models receive the same treatments in AIC, but the difference makes a big difference in a Bayesian framework.

theories that are predictively accurate; scientific realism says that the goal of scientific inference is to find theories that are true.[66]

This does not mean that realists should conclude that AIC is irrelevant to their concerns. AIC's connection with instrumentalism is consistent with AIC's also having a connection with realism. This connection derives from the fact that AIC estimates which of two models, when fitted to the data, will (on average) be closer to the truth, where closeness is measured by Kullback-Leibler distance (Burnham and Anderson 2002).[67] The surprising fact is that a false model can be closer to the truth (in this sense) than a true one. In the next section I will try to remove the air of paradox that surrounds this fact. For now, the point is that realists who want to find fitted models that are approximately true should not look upon AIC with disdain. AIC scores assist realists in this task. AIC is thereby connected to both instrumentalism and to realism. When model M_1 has a better AIC score than model M_2, this is a reason to think that M_1 will be more predictively accurate than M_2 *and* that L(M_1) is closer to the truth than L(M_2) is. This is why I like the following slogan: *instrumentalism for models, realism for fitted models* (Sober 2008b, pp. 96–99).

The motivation for AIC is often described in terms of "noise" and "signal." Consider the example of curve-fitting. The data you have before you is the joint product of the true (but unknown) underlying curve and the fact that your observations are subject to error (Figure 2.2). The true curve is the signal, and the error is the noise. So far so good, but now consider an additional claim: your goal is to separate the noise from the signal in your data, and that is what AIC helps you do. This addition paints too simple a picture of AIC's relation to realism and instrumentalism. To say that the point of AIC is to separate noise from signal makes it sound like the goal of AIC is to find the truth. There is something to this, but there is an instrumentalist aspect

[66] Instrumentalism and realism are each different from Van Fraassen's (1980) constructive empiricism, which says that the goal is to find theories that are true in what they say about observables (Sober 1999a, p. 26).

[67] The Kullback-Leibler distance from a candidate probability distribution *c* to the true distribution *t*, where each has *k* discrete states, is

$$\sum_{i=1}^{k} t_i \log\left[\frac{t_i}{c_i}\right]$$

The KL distance is a "directed" distance; the distance from *c* to *t* (where *t* is true) can differ from the distance from *t* to *c* (where *c* is true).

of AIC that should not be ignored. The goal is to find predictively accurate models, not models that are true. And although noise in the data is extremely common, this is not a conceptual requirement for AIC to do its work. Even when data points must fall on the true underlying curve, a simpler model that fits the data worse may be more predictively accurate than a more complex model that fits the data better.

Why a false model can be more predictively accurate (and closer to the truth) than a true one

It may seem strange that a false model can be more predictively accurate than a true one, and it may seem absurd that a false model can be closer to the truth than a true one, so let me try to demystify both of these ideas. My remarks in this section aren't about estimating predictive accuracy or closeness to the truth; they concern what predictive accuracy and closeness to the truth are.

Consider the following two hypotheses about the coin that sits on the table in front of you; neither contains an adjustable parameter:

(TRUE) The coin is i.i.d. and $p = 0.5$.
(FALSE) The coin is i.i.d. and $p = 0.8$.

TRUE is true and FALSE is false. Imagine a series of experiments. In each, you toss the coin ten times and record the frequency of heads. How predictively accurate will TRUE and FALSE each be? Whether the series is long or short, there is a chance that the average frequency of heads across this series of ten-toss experiments will be closer to 80 percent than it is to 50 percent even though $p = 0.5$ is the true value. However, this *probably* won't happen. When I introduced AIC, I used "predictive accuracy" as a label for the *expected* performance of a model, not as a description of a model's actual performance in a single experiment. In this example, nothing seems strange or absurd: *TRUE is more predictively accurate than FALSE, and TRUE is closer to the truth than FALSE!*

Matters change when you consider models that contain adjustable parameters. Still assuming that the coin is i.i.d. with $p = 0.5$, you now consider the following two hypotheses:

(*T*) The coin is i.i.d. with adjustable parameter p.
(*F*) The coin is i.i.d. with $p = 0.501$.

T is true and *F* is false. *T* has one adjustable parameter and *F* has none. The model *T* is an infinite disjunction; it is true because one of its disjuncts is true, but the fact remains that many of *T*'s disjuncts are very far from the truth. In contrast, *F*, though false, is very close to the truth.

To think about how predictively accurate *T* is, you need to consider what will happen if you repeatedly fit *T* to old data and then see how accurately the fitted model predicts new data. That is, you're going to consider a series of data-set pairs ($<\mathrm{Old}_1$, $\mathrm{New}_1>$, $<\mathrm{Old}_2$, $\mathrm{New}_2>$, . . . , $<\mathrm{Old}_n$, $\mathrm{New}_n>$). Let each data set contain the result of ten tosses. This procedure can be simplified when you assess *F*'s predictive accuracy. Since *F* contains no adjustable parameters, you don't need to fit it to data; you can just see straightaway how accurately *F* predicts the second data set in each pair. What will probably happen is that *F* will do a better job than *T*. How can that happen? As mentioned, *T* is an infinite disjunction with one true disjunct and the rest false. When you fit *T* to a data set of ten tosses, the data may lead you to glom on to a disjunct in *T* that is far from the truth. In contrast, the false hypothesis *F* prevents you from being misled by misleading data, since *F* has no adjustable parameters. The true model *T* dangles temptation before your eyes. The false statement *F* does not; it keeps you to the straight and narrow (or at least close to it). Since *T*'s estimates will, in expectation, be farther from the true value than the one unvarying value that *F* provides, it makes sense to say that *F* is closer to the truth than *T* is. Notice that "closeness to the truth" is here being used to mean that a proposition is close to some *target* proposition that is true.

It really isn't so crazy that a false proposition can be closer to the target than a true one. Consider the following statements:

- Joe's backpack is in the bedroom of his apartment.
- Joe's backpack is in the kitchen of his apartment.
- Joe's backpack is in Madison, Wisconsin.
- Joe's backpack is on planet Earth.

The last three are true, but some are closer to the true point location of Joe's backpack than others.[68] The first statement is false, but it is close to the truth – much closer than is the last item on the list.

In this section, I have not provided a *general* definition of "closeness to the truth" – one that applies to *all* propositions. Rather, I have argued that once

[68] A Bayesian construal of this idea is possible. Think of the average distance from the set of locations circumscribed by a statement to the true location of the backpack.

closeness to the truth is understood in terms of closeness to a target, the concept makes sense for *some* propositions. If this is right, it should not seem paradoxical that a false proposition can be closer to the target than a true one is.[69]

What the two parsimony paradigms have in common

In spite of the differences I've noted that separate the two parsimony paradigms, they do have some common elements. When parsimony is a surrogate for likelihood, and when parsimony is part of what matters in model selection, parsimony is not a subjective aesthetic frill. It has an objective epistemic status. Another similarity is that both the likelihood framework and that of AIC are *contrastive.* In neither case do we examine a single hypothesis and decide whether that hypothesis should be accepted or rejected; rather, we compare two or more hypotheses and see which are better than which others in some relevant sense of "better." The two frameworks also have in common the fact that they do not discriminate between hypotheses that are empirically equivalent in the sense of making the same predictions for all possible data.[70] The law of likelihood says that evidence favors one hypothesis over another only when the two confer *different* probabilities on the observations. In a similar vein, AIC should be applied to a pair of models only when it is possible for them to make different predictions about new data when fitted to old. Another similarity that unites the two parsimony paradigms is that both view "counting causes" as an imperfect epistemological guide. The likelihood comparison of the common cause and the separate cause hypotheses is not

[69] There are some parallels between Popper's (1963) concept of verisimilitude and the concept of predictive accuracy (Forster and Sober 1994). Popper's idea was that a theory's verisimilitude should reflect both its truthfulness and its content; a tautology scores high on the first but low on the second. Popper wanted a false conjunction that has ninety-nine true conjuncts and one false one to have more verisimilitude than a tautology. In contrast, the predictive accuracy of a theory is not defined by counting the true and false conjuncts in a conjunction.

[70] This point connects with Reichenbach's (1958) distinction between *descriptive* and *inductive* simplicity. Reichenbach says that theories that are empirically equivalent can differ in their descriptive simplicity but cannot differ in inductive simplicity. He also says that inductive simplicity applies only when the hypotheses under consideration are incompatible with each other. If we drop this second idea and allow for inductive simplicity to be relevant to the comparison of nested models for their predictive accuracy, then both parsimony paradigms fall squarely in Reichenbach's category of inductive simplicity.

affected by conjoining the separate causes and calling the conjunction a "common cause." And for model selection criteria like AIC, the difference between one cause and two is exactly the same as the difference between two additive causes and two causes that interact; what matters is the number of adjustable parameters, not the number of causes. Another point of agreement between the two paradigms is that the epistemic relevance of parsimony depends on empirical assumptions; in neither case is the principle of parsimony justified *a priori*.

Each of the two parsimony paradigms identifies circumstances in which parsimony has an objective epistemic relevance, but each does something more: each describes how the importance of parsimony compares with the importance of other epistemically relevant considerations. This is clear in the case of AIC; the criterion says that fit to data and parsimony (as measured by number of adjustable parameters) both matter to estimating predictive accuracy and the criterion tells you how to compare theories that have different pluses and minuses. If T_1 is more parsimonious than T_2, while T_2 fits the data better than T_1, AIC doesn't tell you to throw up your hands and say that it is a matter of subjective preference how much parsimony matters as compared with fitting the data. No, the criterion says that these two considerations are commensurable and tells you how to commensurate them. The likelihood paradigm has the same property when it is situated in a Bayesian framework and the problem at hand provides justified values for likelihoods and prior probabilities. The odds formulation of Bayes's theorem tells you how much weight parsimony-qua-likelihood deserves as compared with prior probabilities. Think about the plagiarism example. In this regard, the two parsimony paradigms do not conform to the picture painted by Thomas Kuhn (1977). According to Kuhn, simplicity matters in theory evaluation, but there is no "algorithm" that tells you how much it matters compared with others relevant considerations.

Concluding comments

In discussing common cause versus separate cause hypotheses, I identified a circumstance in which the common cause hypothesis is not just more parsimonious; it also has the higher likelihood. This seems to provide a likelihood justification of Ockham's razor, but the question may linger as to which is the dog and which is the tail. Is parsimony justified because it mirrors likelihoods, or is likelihood justified because it mirrors parsimony? Cases in which parsimony and likelihood walk hand-in-hand do not throw much light on this question. Rather, one needs to turn to cases in which parsimony and likelihood *conflict*.[71] I described a few of these by tinkering with the seven assumptions that together get parsimony to mirror likelihood. In all these cases of conflict, my conclusion is the same: *parsimony be damned!* Here you see a theme that will recur: I am a *reductionist* about parsimony. If parsimony contributes to the achievement of some more fundamental epistemic goal, I am all for it. If it does not, I am not.[72]

Even if we regard likelihood as more fundamental than parsimony, the question remains as to whether the mirroring result really shows that parsimony is epistemically relevant. Maybe what *really* matters is likelihood, and parsimony is a mere epiphenomenon. If so, maybe parsimony's mirroring likelihood provides Ockham's razor with no justification at all. A similar doubt arises in connection with the role that parsimony plays in model selection. Maybe what *really* matters are good estimates of predictive accuracy; since parsimony is not an end in itself, but is merely a means, maybe model selection theory fails to show that parsimony *really* matters. These remarks about "really" really leave me cold. They are like saying that barometer readings aren't evidence of storms, since what *really* matters to whether a storm will occur is the barometric pressure. *E*'s being evidence for *H* is compatible with there being some third factor that screens-off *E* from *H*. The epistemic relevance of parsimony does not require that parsimony be an end in itself.[73]

It is easy to invent a simple and a complex hypothesis where the simple hypothesis seems sensible and the complex one seems preposterous; it also is easy to invent examples in which good sense is on the side of the complex alternative. Both of these results are inconclusive. I have tried to

[71] Cases of conflict between likelihood and parsimony will occupy our attention in Chapter 3 when we consider a concept of parsimony that is used in evolutionary biology.

[72] For other endorsements of reductionism about parsimony/simplicity, see Sober (1988, 1990b), Boyd (1990), Norton (2003), Fitzpatrick (2006), and Woodward (2013). For similar reductionist sentiments about the epistemic relevance of explanatoriness as it is used in the theory of inference called *inference to the best explanation* (Harman 1965; Lipton 1991), see Roche and Sober (2013).

[73] This point bears on Milne's (2003) argument that Bayesianism can provide no account for why simplicity should be epistemically relevant.

avoid this pitfall in the present chapter by using two strategies. One is to look at pairs of hypotheses that differ in complexity but are otherwise similar. In the Reichenbachian comparison of common and separate cause hypotheses, the hypotheses are *very* similar to each other (in virtue of the cross-model homogeneity assumption 7). In the case of model selection, AIC allows you to evaluate hypotheses that differ in complexity but fit the data about equally well. In both instances, the idea is to mimic the empirical scientist's ideal of a controlled experiment where the goal is have two groups that are similar except for the putative causal factor under investigation. The second strategy is to find a formal framework that explains when and why parsimony is epistemically relevant. With a formal framework in hand, you can look at hypotheses that differ in parsimony and are *not* otherwise similar; this is something that AIC hands you on a plate.

The examples discussed in this chapter, like the ones in the chapters to come, are of two kinds. The separation I have in mind involves the distinction between *token* and *type*. Sometimes what unifies two token events is that they trace back to a common token cause. Darwin inferred the existence of common ancestors from the similarities he observed among extant organisms.[74] Common ancestors (like the parents shared by two siblings) are *token* individuals. But unification is often achieved in a different way, by showing that two bodies of data ought to be understood in terms of the same *type* of cause. In the example of the two fields of corn, the choice was between two models, NULL and DIFF. NULL unifies the data because it says that the two fields of corn have the same mean height. The NULL model does not say that the two fields trace back to a common token ancestor; it says that there is some property (a single mean height) that the two fields each exemplify. And Newton's universal law of gravitation is unifying because it says that there is a single law that applies to all material objects; it isn't about the common ancestry of planets and apples.

The distinction between the razor of silence and the razor of denial surfaced several times in Chapter 1. Both address what you should believe – they are about *rational acceptance*. The first says that you should *not believe* hypotheses that aren't needed to explain; the second says that you should *disbelieve* such hypotheses. Given this, it is perhaps a little surprising that the

[74] Lewis (1973, p. 87) says that counting token causes has no epistemic relevance; Nolan (1997), Barnes (2000), and Baker (2003) disagree.

probabilistic turn that we took in the present chapter has failed to directly address the question of what you should believe. The law of likelihood doesn't tell you which of two hypotheses you should believe; it tells you which of the two is favored by the observations. And model selection criteria like AIC also don't tell you which of the models you should believe; rather, they compare models for their predictive accuracies. As noted, the data at hand will sometimes indicate that a model known to be false will be more predictively accurate than a model known to be true. With parsimony sometimes anchored to the law of likelihood and sometimes anchored to model selection criteria, we seem to have drifted away from the very problem that Ockham's razor was invented to address.

In fact, the issue of rational acceptance has not been abandoned. This is because parsimony can be *relevant* to rational belief even when parsimony does not *suffice* to tell you what to believe. This is obvious if you substitute the Bayesian idea of degree of belief for the dichotomous concept of believing deployed by the two razors. The odds formulation of Bayes's theorem (p. 79) shows that the likelihood ratio is one factor, but not the only one, that affects the ratio of posterior probabilities. When parsimony mirrors likelihood, parsimony is relevant to deciding how confident you should be in one hypothesis as opposed to another. As for model selection criteria such as AIC, parsimony is again epistemically relevant; it helps you form beliefs about how models differ in predictive accuracy, but parsimony, by itself, does not tell you which of two models you should expect to be more predictively accurate.

Although rational acceptance does connect with the story I have told, there is another topic that I really have ignored; it concerns a methodological role that parsimony might play in organizing scientific inquiry. Although Jeffreys and Popper disagreed about a lot, they agreed that scientists can't test all theories at once; scientists need to consider some theories before they turn to others. Jeffreys and Popper both suggest that scientists should start by testing simpler theories; if those simpler theories fail, scientists should move to theories that are more complex. Schulte (1999) and Kelly (2007) also endorse this policy, arguing that it has a desirable sort of *efficiency*. They show that this strategy is optimal with respect to the goal of minimizing the number of times you will need to change your mind.[75] It is important to see that

[75] More precisely, Kelly and Schulte argue that the worst-case scenario if you use a simplicity ordering will involve fewer retractions than the worst-case scenario if

Ockham's razor as a search procedure does not conflict with Ockham's razor as a principle for evaluating the theories at hand. Here's an analogy. It is one thing to give advice to tourists about the order in which they should visit several cities; it is something else to tell them what criteria they should apply to a city they visit to determine whether they should move on to the next city on their list or stay put. For Kelly and Schulte, the data tell you whether to accept or reject the theory at hand and this decision is made without help from Ockham's razor. According to their picture, theories are tested one at a time. This may be okay if the candidate theories you are considering are deductively related to observations, but when the relationship is probabilistic, I am skeptical of epistemologies that are non-contrastive. In that circumstance, I think that testing a theory requires testing it against one or more alternatives. I also have doubts about invoking the dichotomous choice between *accept* and *reject* (Sober 2008b).[76]

Several of the philosophers discussed in the previous chapter believed that Ockham's razor is a sensible principle only if nature is simple. In the present chapter, we have seen a steady retreat from this thesis. Jeffreys began the long march; he weakened the thesis by inserting two words. The commitment of his simplicity postulate is not that nature is simple, but that nature is probably pretty simple. For Jeffreys, there is no upper bound on how complex nature might be. We then encountered the Reichenbachian result that says that common cause models have higher likelihoods than separate cause models. The assumptions that suffice to justify this likelihood inequality do not entail that nature is simple or that it is probably simple. We took an additional step along this path when we considered the Akaike Information Criterion; in this instance, the epistemic relevance of parsimony does not depend on the assumption that nature is simple or that it is probably simple. The probabilistic turn that was taken in the twentieth century turned out to be a very big turn indeed.

you use some other ordering. It is unclear why this fact about worst-case scenarios should determine one's choice of search strategy.

[76] See Fitzpatrick (2013) for discussion of Kelly's views on parsimony.

3 Parsimony in evolutionary biology – phylogenetic inference

Given the precision with which the reflection and refraction of light conform to minimum principles (Figure 1.7), it is striking how often evolution produces structures that make a mockery of minimality. As the ancestors of contemporary mammals evolved a higher body temperature, the testes evolved a new location, from deep in the abdomen to a cooler position inside a new structure, the scrotum. The reason for this migration was that the new core temperature was often too high for making and storing healthy sperm.[1] The change in location of the testes may suggest that the tubing (the *vas deferens*) connecting testes to penis should have shortened, but this is often not what happened. The lineage leading to human beings furnishes a typical example. What happened is depicted on the right side of Figure 3.1. The testes went south by moving dorsal to the ureter; the result was that the *vas deferens* ended up with a long loop. What did not happen is shown on the left side of the figure where the journey is ventral and the tubing shortens. Unlike rays of light, mammalian spermatozoa often fail to take the shortest path.

Lots of traits are like this. The remark of Darwin's that I quoted in Chapter 1 bears repeating: many traits "bear the plain stamp of inutility" (Darwin 1859, p. 480). This might lead you to think that evolutionary biology banished the principle of parsimony, but this is not what happened. The principle was not rejected; it was reformulated. Its new application was to the evolutionary process itself, not to the characteristics of organisms that are the products of that process. Ockham's razor was used, not to predict the perfection of an organism's traits, but to infer the history of how those often imperfect traits came to have their present form.

[1] There is disagreement as to when and how many times the shift in location occurred; see Werdelin and Nilsonne (1999), Kleisner *et al.* (2010), and Lovegrove (2014).

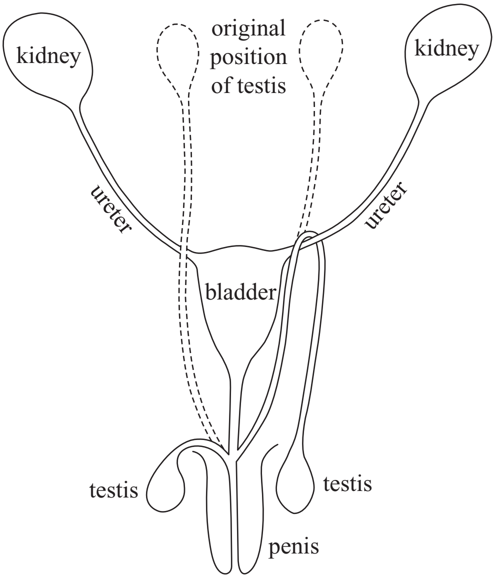


Figure 3.1[2]

Common ancestry

Common ancestry is a fundamental idea in Darwin's theory of evolution and in contemporary evolutionary biology as well.[3] Darwin's idea was that all living things and fossils now on Earth trace back to one or a few original progenitors (Darwin 1859, p. 484, 490). Contemporary biologists are usually less cautious. The standard view now is that there is a single common ancestor shared by all current organisms and fossils here on Earth; this is the hypothesis of *universal* common ancestry. Darwin represented this idea in the one figure he included in *the Origin*; I reproduce it here in Figure 3.2. This picture shows a phylogenetic tree in which extant groups of organisms are at the top; as you move down the page, you move back in time, with modern groups finding their common ancestors and those common ancestors in turn tracing back to more ancient common ancestors. The bottom of Darwin's figure is not the beginning of life on Earth; if you go back further, you'll find more coalescing.

[2] Reprinted by permission of Oxford University Press, USA from George C. Williams's 1992 book *Natural Selection – Domains, Levels, and Challenges*, p. 75, Figure 6.1.

[3] See Sober (2011a, Chapter 1) for discussion of how common ancestry and natural selection (the other main idea in Darwin's theory) are evidentially and causally related to each other.

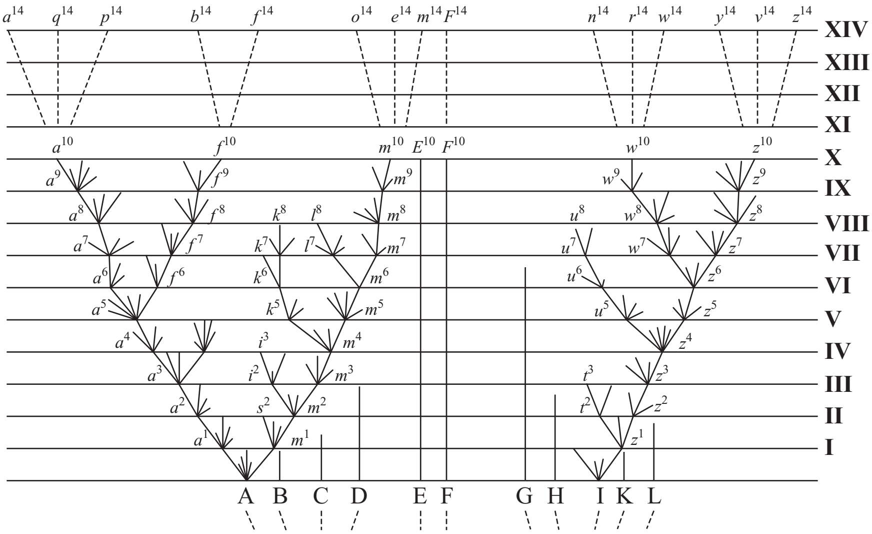

a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} w^{10} z^{10}
X
a^9 f^9 m^9 w^9 z^9
IX
a^8 f^8 k^8 l^8 m^8 u^8 w^8 z^8
VIII
a^7 f^7 k^7 l^7 m^7 u^7 w^7 z^7
VII
a^6 f^6 k^6 m^6 u^6 z^6
VI
a^5 k^5 m^5 u^5 z^5
V
a^4 i^3 m^4 z^4
IV
a^3 i^2 m^3 t^3 z^3
III
a^2 s^2 m^2 t^2 z^2
II
a^1 m^1 z^1
I
A B C D E F G H I K L

Figure 3.2

A principle of parsimony is widely used in the part of contemporary evolutionary biology that reconstructs the genealogies of extant organisms. Given the preceding chapter, with its extended discussion of the idea that common cause explanations are more parsimonious than separate cause explanations, you might expect parsimony to play a large and explicit role in the arguments that biologists now give for the hypothesis of universal common ancestry. This natural expectation isn't quite right. The evidence for thinking that all life traces back to a single common ancestor usually gets described without Ockham's razor being waved in the air. For example, the near-universality of the genetic code is a standard piece of evidence for universal common ancestry (Freeland *et al.* 2000; Theobald 2010). On the assumption that there are many possible genetic codes that would be functional, the near-universality would be very improbable if extant species traced back to multiple start-ups. The near-perfect matching is much more probable under the hypothesis of a single common ancestor. This is a likelihood argument that stands on its own, but notice that universal common ancestry provides a more parsimonious explanation of the shared genetic code than the hypothesis of separate ancestry is able to provide. The matching of genetic codes is like the matching of the student essays discussed in the previous chapter. In both, the more parsimonious hypothesis postulates fewer causes, and parsimony mirrors likelihood.

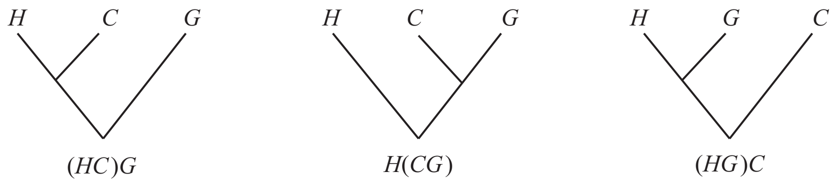


Figure 3.3

Explicit reference to parsimony in evolutionary biology is more common in connection with two other genealogical inference problems. Both are posed under the *assumption* that the species considered have a common ancestor; the species are assumed to occupy the leaves (the tips) of a phylogenetic tree. The problems concern that tree's *topology* and the character states of its *interior nodes*.

To understand what a tree's topology is, consider Figure 3.3. Suppose we observe the characteristics of current human beings (*H*), chimpanzees (*C*), and gorillas (*G*), and we want to know which two of these species are more closely related to each other than either is to the third. The three possible topologies are (*HC*)*G*, *H*(*CG*) and (*HG*)*C*. (*HC*)*G* says that *H* and *C* are more closely related to each other than either is to *G* in the sense that *H* and *C* share an ancestor (depicted by an interior node) that is not an ancestor of *G*, while every ancestor of *H* and *G* is also an ancestor of *C*. These three topologies disagree about what the *monophyletic groups* are. A set of taxa is said to comprise a monophyletic group precisely when the set includes an ancestor and all of its descendants. (*HC*)*G* say that *H* and *C* are in a monophyletic group that does not include *G*; the other two hypotheses reject this claim.

Although the three trees in Figure 3.3 disagree about genealogy, there are several points of agreement. First, each asserts that the three species have a common ancestor. Second, each says that as one moves from root to leaves, branches split but never join. This is a tree in the technical sense that is used in evolutionary biology; there are no *reticulations*.[4] Third, each is incompatible

[4] This means that the leaf species one is considering do not arise by the hybridization of distinct ancestors and that there is little or no horizontal gene transfer that connects branches to each other. This is not always true. Bacteria exhibit massive horizontal gene transfer from other bacteria; other organisms undergo this process, though apparently to a much more modest degree (Doolittle 2000). A very small percentage of the human genome is due to horizontal gene transfer (Hotopp 2011).

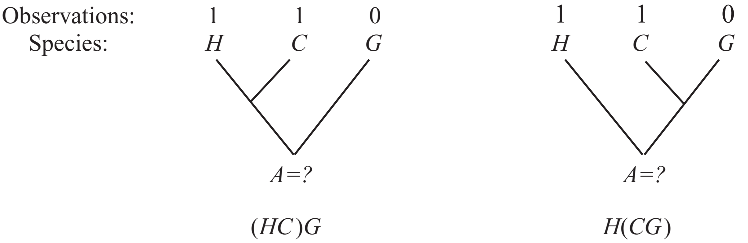


Figure 3.4

with a *star* phylogeny, wherein a single ancestor simultaneously gives birth to three or more descendant lineages with the result that no two leaves are more closely related to each other than either is to a third. A fourth commonality concerns what these tree hypotheses do *not* say. There is no claim as to how long ago the most recent common ancestor of the three species existed, or about what the characteristics are of the ancestors (represented by interior nodes), or about the processes that caused branching and the evolution of different traits in different lineages.

If you know some of the characteristics that humans, chimpanzees, and gorillas each possess, how might those observations be used to evaluate which of the three genealogical hypotheses is best supported? This is where the principle of parsimony comes in. For each tree, you need to determine the minimum number of changes in character state that would allow the tree to produce the observations. The most parsimonious tree is the one that requires the fewest changes. This type of parsimony is called *cladistic* parsimony.[5]

To determine which of the three trees in Figure 3.3 provides the most parsimonious explanation of the observations, you need to start with some view about which character states the most recent common ancestor of the leaf species occupied.[6] This point is depicted in Figure 3.4. We observe that human beings and chimpanzees have a characteristic (coded as 1) that gorillas lack (gorillas therefore are in state 0). If the ancestor *A* was in state 0, then the (*HC*)*G* tree is more parsimonious than *H*(*CG*) as an explanation of the observed 1-1-0 pattern; "1-1-0" encodes the fact that humans and chimpanzees are in state 1 whereas gorillas are in state 0. However, if *A* was in state 1, the two trees are

[5] "Clade" is Greek for branch.

[6] Terminology: color is a character, while, red, green, and blue are character states. Philosophers mark this distinction by talking about determinables and determinates.

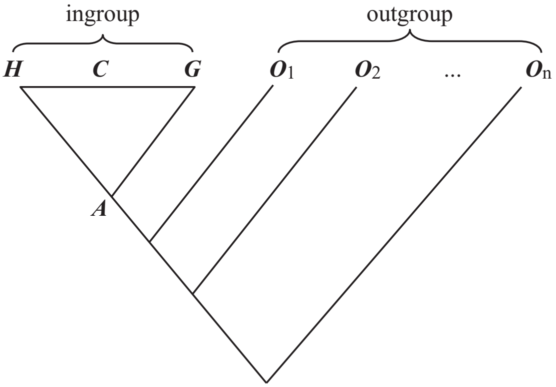


Figure 3.5

equally parsimonious. According to cladistic parsimony, the 1-1-0 pattern is evidence for (*HC*)*G* and against the other two topologies if 0 is the state of the most recent common ancestor, but not if 1 is the state of the ancestor. The question is whether 0 is ancestral (*plesiomorphic*) and 1 is derived (*apomorphic*), or *vice versa*. For cladistic parsimony, it isn't true that any old similarity counts as evidence. If 0 is the state of the ancestor, the 1-1-0 pattern will be evidentially relevant, but the 0-0-1 pattern will not be. It is *synapomorphies* (shared derived characters) that provide evidence; *symplesiomorphies* (shared ancestral characters) do not. This, I hasten to emphasize, is what cladistic parsimony says about the interpretation of similarities. We will examine whether other frameworks for the interpretation of evidence agree.

To infer whether 0 is ancestral and 1 is derived, or *vice versa*, proponents of cladistic parsimony usually use the *method of outgroup comparison*.[7] To polarize a character, you look at species that are outside the *H-C-G* group but are close relatives of those three species. This technique is illustrated in Figure 3.5. If the outgroup species $O_1, O_2, \ldots, O_n$ are all in the same state, you should infer that *A* was also in that state. The justification for this inference is, once again, cladistic parsimony.[8] Notice that inferring the phylogenetic

[7] Two other methods for polarizing characters are used: stratigraphy (wherein the character state seen earliest in the fossil record is taken to be plesiomorphic) and ontogeny (wherein the state seen earliest in development is taken to be plesiomorphic).

[8] If some outgroups are in state 0 and others are in state 1, the problem is more complicated, but, again, it gets solved by using cladistic parsimony.

relationship among humans, chimpanzees, and gorillas requires the assumption that these ingroup species are more closely related to each other than any of them is to the species that count as outgroups. The method of outgroup comparison cannot get off the ground without this genealogical assumption.

If a 1-1-0 character evolves on the (*HC*)*G* tree in Figure 3.4 and 0 is the ancestral state, then the synapomorphy (the shared derived character) that unites humans and chimpanzees may be a *homology*. That is, it is possible that the most recent common ancestor of humans and chimpanzees was in state 1 and this state was retained unchanged all the way up to the two leaves. However, there is another possibility. The most recent common ancestor of humans and chimpanzees may have been in state 0 and then the two branches stemming from that ancestor each evolved from 0 to 1. If this is what happened, then the 1 state found in humans and chimpanzees is a *convergence*. A third possibility is that the most recent common ancestor of humans and chimpanzees was in the 1 state, but then each lineage stemming from that ancestor shifted from 1 to 0 and then back to 1. Here we say that the 1 state experienced two *reversals* and then re-evolved. Both convergence and reversal fall under the heading of *homoplasy*. Homology and homoplasy are opposites.

Given that the 1 state found in humans and chimpanzees *may* be a homology if (*HC*)*G* is the true tree and 0 is the ancestral state, what can we say of a different character distribution? Suppose a character evolves on that same tree with 0 being the ancestral character state, but the result is that the leaves exhibit the 1-0-1 pattern for that character. This matching of humans and gorillas *must* be a homoplasy; it can't be a homology. You can convince yourself that this is so by looking at the (*HC*)*G* tree in Figure 3.4.

It is a mistake to think that a trait is a homology or a homoplasy full stop; rather, it is the presence of a trait in two taxa (or more) that is a homology or a homoplasy. To see this, consider an example I discussed in the previous chapter – the torpedo-shape of many aquatic predators. Gray whales and orcas share this characteristic, but so do gray whales and sharks. The torpedo-shape that gray whales and orcas exhibit is a homology, but the same shape exhibited by gray whales and sharks is a homoplasy. Both of these points are depicted in Figure 3.6, which shows a tree topology to which I have added an indication of the character states of two ancestors. Darwin uses this example to comment on "an apparent paradox":

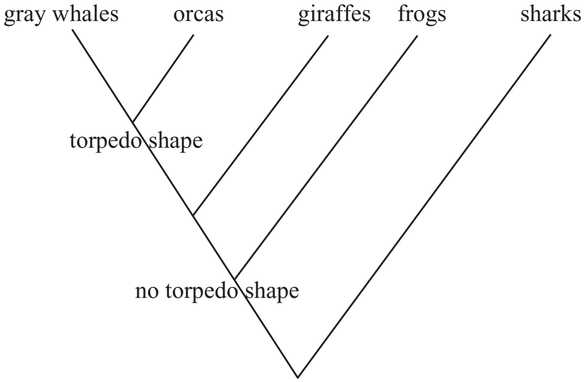


Figure 3.6

> the very same characters are analogical when one class or order is compared with another, but give true affinities when the members of the same class or order are compared one with another: thus the shape of the body and fin-like limbs are only analogical when whales are compared with fishes, being adaptations in both classes for swimming through the water; but the shape of the body and fin-like limbs serve as characters exhibiting true affinity between the several members of the whale family; for these cetaceans agree in so many characters, great and small, that we cannot doubt that they have inherited their general shape of body and structure of limbs from a common ancestor. So it is with fishes. (Darwin 1859, p. 428)

Darwin's "analogy" refers to homoplasies, and his "inheritance from a common ancestor" denotes homologies.

If we score humans, chimpanzees, and gorillas for a dichotomous character (whose two states are 0 and 1), there are eight possible patterns: 1-1-1, 1-1-0, 1-0-1, 0-1-1, 0-0-1, 0-1-0, 1-0-0, and 0-0-0. Regardless of how the character is polarized, cladistic parsimony views 1-1-1 and 0-0-0 as evidentially meaningless; a trait that is shared by all three groups fails to discriminate among the three possible tree topologies. If 0 is the ancestral state of a character, then only three of the six remaining patterns are evidentially meaningful; 1-1-0 favors (*HC*)*G*, 0-1-1 favors *H*(*CG*), and 1-0-1 favors (*HG*)*C*. The 0-0-1, 0-1-0, and 1-0-0 patterns are symplesiomorphies and are therefore judged by cladistic parsimony to be uninformative. Suppose you are considering a hundred dichotomous characters, each of them polarized, and you use "0" to code for the ancestral state and "1" for the derived state of each character. If forty of these synapomorphies have the 1-1-0 pattern while thirty have the 0-1-1 pattern and thirty have the 1-0-1 pattern, what should you conclude if you let cladistic parsimony be your guide? If you assume that each character deserves the same weight, you will conclude that (*HC*)*G* is the best tree overall, since it requires fewer changes to account for the hundred characters than the other trees require.[9] In this example, the most parsimonious tree says that sixty of those hundred synapomorphies are homoplasies; the most parsimonious tree does not say that homoplasies are rare.

Although cladistic parsimony applies only to hypotheses that specify genealogical trees, it applies to problems outside of biology. Cultural evolution resembles biological evolution when both give rise to branching processes in which there are ancestors and descendants.[10] Cladistic parsimony has been used to reconstruct the genealogies of languages (Gray and Jordan 2000; Holden 2002; Atkinson and Gray 2005; Oppenheimer 2006, pp. 290–300, 340–356). And written texts (like the *Bible*, the *Iliad*, and the *Canterbury Tales*) often have a branching history wherein an urtext is copied and then copies are copied, with the result that present versions of the text differ from each other. Cladistic parsimony has been used to reconstruct the genealogy that connects descendant texts to their ancestors (Platnick and Cameron 1977; Robinson and O'Hara 1996; Maas 2010; Barbrook *et al.* 1998). Cladistic parsimony also has been used to infer the genealogies of archaeological artefacts (O'Brien *et al.* 2003).

I mentioned that there are two uses of cladistic parsimony in biology – inferring tree topologies and inferring the character states of ancestors. This second application is illustrated in Figure 3.7. Suppose you use numerous characters to infer that the phylogeny of the leaf species $B, C, \ldots, G$ is the one shown. You then consider a new dichotomous character (one that you have not polarized) and observe which leaf species are in state 1 and which are in state 0, as shown. Parsimony can now be used to infer the characteristics of the five ancestors that this phylogeny postulates. The most parsimonious hypothesis is that A_1 and A_2 were in state 1 and the other three ancestors were

[9] If you think that some characters provide stronger evidence than others, a weighted parsimony procedure can be used. Parsimony does not provide a methodology for weighting characters, but that doesn't show that parsimony is mistaken in its evaluation of the evidence that each character provides about phylogeny. Rather, the proper conclusion is that the method is incomplete.

[10] Darwin (1871, pp. 59–60) makes this point about languages.

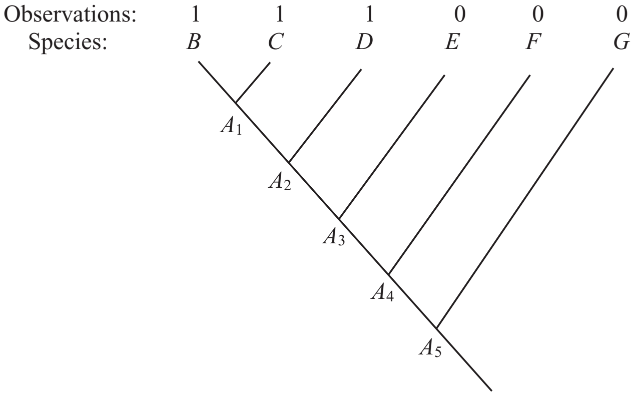


Figure 3.7

in state 0. Alternative assignments of character states to ancestors entail that more changes must have occurred for the tree to produce the observations we have of the leaf species.[11]

There is a third application of cladistic parsimony that I'll discuss at the end of this chapter. It involves inferring the characteristics of a contemporary species from the characteristics found in one (or more) of its contemporary relatives. This application of cladistic parsimony isn't very common in evolutionary biology. Inference problems in that domain usually involve *observing* the characteristics of contemporary species; there is no need to *infer* them. The (controversial) application I have in mind occurs in psychology, where cladistic parsimony has been used to infer the mental characteristics of chimpanzees from the mental characteristics of human beings. Here cladistic parsimony is put to work to defend a kind of anthropomorphism.

Some history

Implicit appeals to cladistic parsimony in evolutionary biology date back to Darwin. For example, consider this passage from *the Origin*:

> Generally when the same organ appears in several members of the same class, especially if in members having very different habits of life, we may attribute its presence to inheritance from a common ancestor; and its absence in some of the members to its loss through disuse or natural selection. (Darwin 1859, p. 192)

[11] See Sober (2011a, Chapter 1) for discussion of Darwin's use of this pattern of inference to test hypotheses about adaptation.

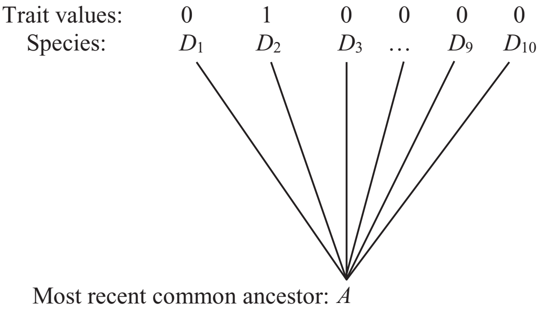

Trait values: 0 1 0 0 0 0
Species: D_1 D_2 D_3 … D_9 D_{10}
Most recent common ancestor: A

Figure 3.8

A simplified example of Darwin's idea is depicted in Figure 3.8. Here we have a star phylogeny in which an ancestor simultaneously gives rise to ten lineages, which result in ten extant descendants.[12] If all but one of those descendants are in character state 0, what is the best estimate of the character state found in their most recent common ancestor *A*? Parsimony says that the answer is 0. This hypothesis requires just one change to have occurred in the tree's interior. The opposite answer, that the ancestor was in state 1, requires nine changes. Darwin does not use the word "parsimony," but it is no great stretch to say that he is thinking along those lines.[13] Darwin isn't saying that the ancestor *must* be in the state found in the vast majority of descendants. Later in the same paragraph he points out that natural selection can lead lineages to converge independently on the same solution to an adaptive problem "in nearly the same way as two men have sometimes independently hit on the very same invention (pp. 193–194)." Was Darwin here alluding to himself and Alfred Russel Wallace (1823–1913), the two co-discoverers of the theory of

[12] Here and in what follows I consider star phylogenies because they provide simple examples in which parsimony applies to the problem of inferring the character state of a common ancestor. I do this even though star phylogenies are rarely considered when the problem is to evaluate competing tree topologies.

[13] Darwin (1862, pp. 306–307) comes closer to connecting phylogenetic inference to a principle of parsimony in his book on orchids: "Can we, in truth, feel satisfied by saying that each Orchid was created, exactly as we now see it, on a certain 'ideal type;' that the Omnipotent Creator, having fixed on one plan for the whole Order, did not please to depart from this plan; that He, therefore, made the same organ to perform diverse functions – often of trifling importance compared with their proper function – converted other organs into mere purposeless rudiments, and arranged all as if they had to stand separate, and then made them cohere. Is it not a more simple and intelligible view that all Orchids owe what they have in common to descent from some monocotyledonous plant?"

evolution by natural selection? The origin of species and the origin of ideas both exhibit homoplasies.[14]

Explicit reference to parsimony in the context of genealogical inference problems is more recent. I won't attempt to date its first appearance, but it is clear that current discussion of parsimony in phylogenetic biology traces back to two separate sources. The first is the work of Willi Hennig (1913–1976), a German entomologist who is the founding father of cladistics; the second is the work of A. W. F. Edwards and Luigi Cavalli-Sforza, two population geneticists.

The enormous influence of Hennig's work began when the English translation of his 1950 book appeared in 1966. Hennig defended an "auxiliary principle" according to which synapomorphies are to be viewed as evidence of phylogenetic relatedness. Hennig's justification of this methodological principle is succinct:

> that the presence of apomorphous characters in different species "is always reason for suspecting kinship [i.e., that the species belong to a monophyletic group], and that their origin by convergence should not be assumed a priori" . . . This was based on the conviction that "phylogenetic systematics would lose all the ground on which it stands" if the presence of apomorphous characters in different species were considered first of all as convergences (or parallelisms), with proof to the contrary required in each case. Rather, the burden of proof must be placed on the contention that "in individual cases the possession of common apomorphous characters may be based only on convergence (or parallelism)." (Hennig 1966, pp. 121–122; square-bracketed material and quotation marks are his)

Hennig's prose is reminiscent of the views of Hume and Kant discussed in Chapter 1. Hennig is saying that phylogenetic reconstruction would be impossible unless we regard synapomorphies as evidence of genealogical relatedness, just as Hume said that inductive inference would be impossible unless we embrace the principle of the uniformity of nature and Kant said that science would be impossible unless we assume that nature is a unity. All three thinkers are saying that a given intellectual activity could not be carried out

[14] In fact the two "independent" discoveries had a common cause; both Darwin and Wallace read Thomas Malthus's 1797 *Essay on the Principle of Population* and both report that this opened their eyes to the idea of natural selection. See Wallace (1905) and Darwin (1958).

unless we make this or that assumption. It is curious that Hennig does not say in this passage that synapomorphies are the *only* evidence of relatedness; he does not say that symplesiomorphies are evidentially uninformative, though this was his intent.

My comments in Chapter 1 on Kant's thesis also apply to Hennig's defense of his auxiliary principle. Maybe cladistic parsimony is the *best* way to infer phylogenies, but it simply isn't true that it is the *only* game in town. For example, you could treat *all* similarities – ancestral as well as derived – as evidence of genealogical relatedness. In addition, biologists now often use statistical methods for reconstructing phylogenies that have no commitment to cladistic parsimony. And there is always the skeptical option – perhaps reliable inference of genealogy *is* impossible. Hennig's point that we can't infer genealogies unless we assume that synapomorphies are evidence of relatedness does not address that worry.

Hennig's followers tried to do better. Some used Popperian ideas about falsifiability and argued that the most parsimonious tree is the one that is least falsified by the data (Eldredge and Cracraft 1980; Wiley 1981). The problem with this approach is that the data that go into phylogenetic inference never "falsify" hypotheses about tree topologies in the strict Popperian sense. In our example of humans, chimpanzees, and gorillas, you can't deduce that (*HC*)*G* is false from a character that has the 1-0-1 distribution (where 0 is the ancestral state). Perhaps 1-0-1 counts as evidence against (*HC*)*G*, but that evidential claim requires a concept of evidence that differs from Popper's concept of falsifiability. Similarly, the fact that a 0-0-1 symplesiomorphy fails to falsify any tree topology does not show that it is evidentially meaningless. Perhaps it is, but that too requires a non-Popperian argument. The simple point here is that tree topologies and data are not related deductively, whereas Popperian falsifiability is all about deductive relations.[15]

[15] As noted in Chapter 2, Popper (1959, p. 197) recognized that probability statements (like the hypothesis that a coin is fair) are not falsifiable in the deductive sense of that term, so he relaxed his definition. He suggested that a hypothesis is falsified if it says that the data are sufficiently improbable. He offers no fixed cut-off here, but says that facts about the precision of one's measurement devices can help one decide what it takes to reject a probabilistic hypothesis. For criticisms of Fisherian significance tests (which also apply to what Popper is here embracing), see Sober (2009b). Popper's relaxed criterion does not settle whether observing one or several 0-1-1 characters suffices to reject (*HC*)*G*.

Farris (1983) provides a non-Popperian defense of cladistic parsimony. His key idea is that we should prefer the genealogy that has the greatest explanatory power and that "the explanatory power of a genealogy is . . . measured by the degree to which it can avoid postulating homoplasies" (p. 17). In terms of our example concerning humans, chimpanzees, and gorillas, Farris's point is that (*HC*)*G* has more explanatory power than *H*(*CG*) does, relative to a synapomorphy that has the 1-1-0 pattern, and that the two have equal explanatory power, relative to a symplesiomorphy of the 0-0-1 pattern. Why are these two claims true? With respect to the first, I agree that the (*HC*)*G* hypothesis, if true, would be part of the explanation of the 1-1-0 pattern. However, the same is true of the *H*(*CG*) hypothesis. Neither topology provides a *complete* explanation of 1-1-0, but each can *help* to explain it. This leaves it unclear why hypotheses that require fewer homoplasies should be regarded as having greater explanatory power. To assess Farris's idea about explanatory power, it would be useful to have a quantitative measure of explanatoriness. One possibility is that explanatory power goes by likelihoods. Understood in this way, Farris's point is that (*HC*)*G* explains a 1-1-0 synapomorphy more than *H*(*CG*) does because $\Pr[1\text{-}1\text{-}0 \mid (HC)G] > \Pr[1\text{-}1\text{-}0 \mid H(CG)]$ is attractive, for reasons I'll explain.

The second intellectual source for contemporary discussion of cladistic parsimony that I mentioned is the work of Luigi Cavalli-Sforza and Anthony Edwards. Both were students of R. A. Fisher who absorbed from their teacher the idea I discussed in the previous chapter under the heading of the law of likelihood. In attempting to apply this law to the evaluation of tree topologies, they encountered some technical difficulties, which they attempted to finesse by proposing a *principle of minimum evolution*. Instead of calculating likelihoods, you can calculate the number of changes that a tree requires. This is how they state their principle:

> The most plausible estimate of the evolutionary tree is that which invokes the minimum net amount of evolution. (Edwards and Cavalli-Sforza 1963)

It may sound as if the authors are saying that the best tree is the one that asserts that evolutionary change has been minimal, but this isn't how they understand the principle. Consider the following statement from a later publication:

> The assumptions underlying this method are not too clear; it may go some way toward handling . . . [situations in which parameter values needed for

> explicitly statistical estimation are unknown], but its success is probably due to the closeness of the solution it gives to the projection of the "maximum-likelihood" tree. The extent of this similarity merits further investigation . . . It certainly cannot be justified on the grounds that evolution proceeds according to some minimum principle. (Cavalli-Sforza and Edwards 1967, p. 555; square brackets mine)

For Cavalli-Sforza and Edwards, cladistic parsimony is justified to the extent that it mirrors likelihoods. They did not demonstrate that this mirroring relation always obtains or even that it obtains in some specified set of circumstances. Instead, their tentative suggestion is that cladistic parsimony reflects likelihoods often enough for it to be a worthwhile method. In contrast with this tentativeness, there is one point on which Cavalli-Sforza and Edwards are more assertive. They are certain that evolution is not a parsimonious process; they have no time for the suggestion that if a lineage evolves from point *A* to point *B* that it does so by following the shortest path (think of Leibniz on light, discussed in Chapter 1).[16] In saying this, Cavalli-Sforza and Edwards were responding to the confident assertion to the contrary made by two other prominent biologists, Joseph Camin and Robert Sokal (1965, pp. 323–324) who say that parsimony "depends on the assumption that nature is indeed parsimonious." Camin and Sokal provide no more argument for their assertion than Cavalli-Sforza and Edwards offer for their contrary claim. So we are left with a puzzle: if evolution is not a parsimonious process, how can the method of cladistic parsimony make sense? How can the most parsimonious tree have the highest likelihood if evolution doesn't follow the shortest path?

Let us start with a more general question: if a method says that the best hypothesis (within a set of competitors) is the one that minimizes *X*, does it follow that the method assumes that *X* has been minimal out there in nature? Farris (1983) addresses this question by distinguishing between minimizing assumptions and assuming minimality. His idea is that cladistic parsimony minimizes assumptions of homoplasy; it does not assume that homoplasies have been rare. We've already seen that the most parsimonious tree for a data set sometimes entails that homoplasies are abundant. However, this is a point about the *hypotheses* that cladistic parsimony sometimes favors; it doesn't settle what the *method* of cladistic parsimony assumes about nature. Methods often involve assumptions that don't surface in the hypotheses those

[16] Edwards (2007) is skeptical about the idea that evolutionary processes obey a minimization (or a maximization) principle.

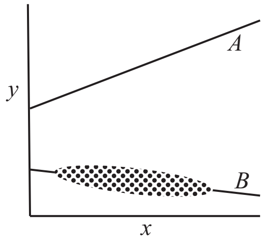


Figure 3.9

methods tell us are best. For example, suppose you want to estimate the mean height in a population. You assume that your data were drawn by random sampling from a single bell-shaped distribution. The maximum likelihood estimate of the population mean is the mean height in the sample. Based on this, you embrace the hypothesis that the mean height in the population is about 5.5 feet. This hypothesis says nothing about random sampling or normal distributions. But that does not show that the method you used to arrive at this hypothesis makes no such assumptions (Sober 1988, p. 137). Methods are one thing, hypotheses another.

Farris (1983) defends his thesis – that minimizing assumptions is not the same as assuming minimality – by discussing an example from outside phylogenetics. In linear regression analysis, you find the straight line that best fits the data. This line minimizes the residual variance (the sum of squared distances between the data points and the line). Does linear regression depend on the assumption that out there in nature there is very little variance around the true regression line? This problem is depicted in Figure 3.9. What's the rationale for preferring line *B* to line *A*, given the cloud of data points represented? The law of likelihood provides a simple justification:

$$\Pr(\text{data} \mid \text{line } B) > \Pr(\text{data} \mid \text{line } A).$$

It suffices to derive this inequality that each line postulates a bell-shaped error distribution whose variance can be estimated from the data. You don't need to know in advance how much variance there is around the true regression line. Farris is right that the justification for preferring the line that minimizes residual variance does not depend on the assumption that the variance around the true regression line is in fact minimal. However, it isn't clear whether this rationale for linear regression carries over to cladistic parsimony. In terms of our example of the hundred synapomorphies that characterize humans, chimpanzees, and gorillas, the question is whether

$$\Pr[\text{data} \mid (HC)G] > \Pr[\text{data} \mid H(CG)], \Pr[\text{data} \mid (HG)C].$$

As Farris realizes, his point about linear regression leaves open whether this inequality is true. We'll get to that soon.

Cavalli-Sforza and Edwards look with favor on cladistic parsimony, and so does Farris. They also agree that the method does not presuppose that homoplasies are rare, though it isn't clear, so far, whether they are right in thinking this. Disagreement arises over the question of how cladistic parsimony is to be justified. Cavalli-Sforza and Edwards want to ground parsimony on likelihood, whereas Farris and other cladists often reject this line of defense. Not that cladists are obliged to maintain that parsimony is rock bottom; as noted, they have tried to justify it by connecting it with concepts like falsifiability and explanatory power. However, Farris (1983, p. 17) says that "the statistical approach to phylogenetic inference was wrong from the start" and many cladists agree. The reason they reject the idea that parsimony derives its authority from likelihood considerations is that likelihood arguments inevitably depend on substantive assumptions about the evolutionary process.[17] Cladists often maintain that parsimony rests on assumptions that are far more modest (Wiley 1975, p. 234; Gaffney 1979, p. 86; Eldredge and Cracraft 1980, p. 4).

Ockham meets Markov

Let's address the question of whether cladistic parsimony assumes that evolution proceeds parsimoniously by considering a simple example. Consider a lineage that extends from an ancestor *A* to a descendant *D*. You observe that the descendent *D* is in state 0 of a dichotomous character. What should you conclude about the state of the ancestor? As noted earlier in connection with the star phylogeny in Figure 3.8, the most parsimonious estimate is that *A* was also in state 0. Does this conclusion depend on the assumption that evolution follows the shortest path? Obviously, that assumption *suffices* to justify the parsimonious conclusion. The question is whether it is *necessary*.

[17] You saw a glimmer of this point in the previous chapter when I provided a justification for the claim that common cause explanations have higher likelihoods than separate cause explanations. Substantive assumptions were needed!

My tools for investigating this question are the law of likelihood and a framework for modeling the evolutionary process that is widely used by statistically minded phylogeneticists.[18] Markov models of evolution treat lineages as objects that have probabilities of changing state in very small instants of time.[19] Since these instants are very brief, the chance of a lineage's changing in an instant is very small. If the character in question is dichotomous, there are two instantaneous probabilities of change:

$$u = \Pr(\text{lineage is in state 1 at time t} + 1 \mid \text{lineage is in state 0 at time t})$$

$$v = \Pr(\text{lineage is in state 0 at time t} + 1 \mid \text{lineage is in state 1 at time t})$$

I'll call u and v the *instantaneous* probabilities of change. These two probabilities may or may not be equal. Markov models of evolution compute the probability that a lineage will end in state i, given that it starts in state j (where i and j each can take values of 0 or 1) and there are t units of time between start and finish. There are four such probabilities to consider:

$$\Pr_{\mathrm{t}}(\text{end in state 1} \mid \text{start in state 0}) = \frac{u}{u+v} - \frac{u}{u+v}(1-u-v)^t.$$

$$\Pr_{\mathrm{t}}(\text{end in state 0} \mid \text{start in state 0}) = \frac{v}{u+v} + \frac{u}{u+v}(1-u-v)^t$$

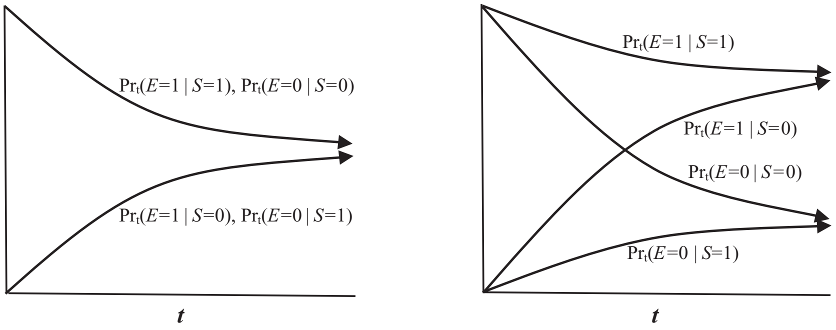
$$\Pr_{\mathrm{t}}(\text{end in state 0} \mid \text{start in state 1}) = \frac{v}{u+v} - \frac{v}{u+v}(1-u-v)^t$$

$$\Pr_{\mathrm{t}}(\text{end in state 1} \mid \text{start in state 1}) = \frac{u}{u+v} + \frac{v}{u+v}(1-u-v)^t.$$

I'll call these four probabilities *branch* transition probabilities. Notice that the first two probabilities sum to one, as do the third and fourth. In each equation, the amount of time t between the lineage's start and finish goes unmentioned in the first addend. The second addend brings this up, and this addend shrinks towards zero as t increases. This means that the first addend describes an *equilibrium* probability – the probability that obtains when there is an infinite amount of time in the lineage. When time is short, the values of these transition probabilities are mainly determined by the lineage's initial state; if the lineage begins in a given state, it will almost certainly end in that same state. For example, if $t = 0$, the first and third probabilities displayed above have values of 0 and the second and fourth have values of 1. As the

[18] In this paragraph and the next, I borrow some prose from Sober (2011a, pp. 156–157).

[19] Andrey Markov (1856–1922) was a Russian mathematician who worked on stochastic processes.



(drift)

(selection for state 1)

Figure 3.10

duration of the lineage is increased, the process plays a progressively larger role in determining the probability of the final state and the initial condition of the lineage is steadily forgotten. The process has the Markov property (which Descartes supplied with a theological explanation that I discussed in Chapter 1); as a lineage evolves from a state *D* in the distant past to a state *R* in the recent past to its state *P* in the present, *R* screens-off *D* from *P*. Each of the four conditional probabilities described above sums over all the possible flip-flops that may occur in the lineage between start and finish. For example, the second and fourth cover all the even numbers and zero as well. It would be inaccurate to say that these two probabilities represent the probability of *stasis* (that there were zero changes in the branch).

The Markov framework can be used to describe the difference between selection and drift. Selection for state 1 is represented by the inequality $u > v$; when there is selection for state 1, the probability of changing from 0 to 1 is greater than the probability of changing from 1 to 0. Pure drift (no selection) means that $u = v$; the two changes have the same probability. This difference between the two processes is reflected in the different branch transition probabilities that are depicted in Figure 3.10. I abbreviate the branch transition probabilities in that figure by using $\mathrm{Pr}_t(E = i \mid S = j)$ to represent the probability that the lineage will end in state *i*, given that it starts in state *j* and there are *t* units of time in between.

The Markov framework entails a "backwards inequality" when the evolving trait is dichotomous (Sober 1988). For any values of *i* and *j* (where $i \neq j$) and *t*, when *u* and *v* are each less than ½,

(BI) $\mathrm{Pr}_t(\text{end in state } i \mid \text{start in state } i) > \mathrm{Pr}_t(\text{end in state } i \mid \text{start in state } j)$.

Compare the first and fourth equations above (and also the second and third). The backwards inequality means that if a descendant is in state i, the hypothesis that its ancestor was in state i has a higher likelihood than the hypothesis that its ancestor was in state j.[20] Don't confuse the backwards inequality with the following "forwards inequality":

(FI) $\Pr_t(\text{end in state } i \mid \text{start in state } i) > \Pr_t(\text{end in state } j \mid \text{start in state } i)$.

The Markov framework leaves open whether a lineage has a higher probability of ending in the same state in which it began than of ending in a different state. Figure 3.10 shows that the backwards inequality is true for both selection and drift, no matter what the lineage's duration is, whereas the forwards inequality is true for drift but false for selection when the amount of time between start and finish is sufficiently large. If a lineage starts in state 0 and there is selection for state 1, then the lineage will probably end in state 1 if there is enough time.

Now back to the problem of inferring the character state of an ancestor A, given that its descendant D was in state 0. As mentioned, the most parsimonious hypothesis about the ancestor is that A was also in state 0. This is also the hypothesis of maximum likelihood, thanks to the Markov model's entailing the backwards inequality BI. What is more, the BI does not depend on the assumption that evolution proceeds parsimoniously. The hypothesis that $A = 0$ has a higher likelihood than the hypothesis that $A = 1$ even when there probably were multiple changes in state between start and finish. If there is drift, the expected number of changes in a lineage that has a duration of t units of time is ut. This quantity can be big or small and the backwards inequality is still true. The same point holds if there is selection. What the expected number of changes in a lineage influences is *how much* the observation that $D = 0$ favors $A = 0$ over $A = 1$. The likelihood ratio

$$\frac{\Pr_t(D = 0 \mid A = 0)}{\Pr_t(D = 0 \mid A = 1)}$$

shrinks as t increases, but it never drops below unity.

You may be thinking that a single descendant that is in state 0 provides scant evidence that its ancestor was also in state 0. Actually, this depends on

[20] In this argument I have assumed that u and v are constant in the lineage from start to finish. This isn't essential. If you segment the lineage into n time periods and assign a different pair of u and v values to each, the backwards inequality still holds.

the value of t, the number of time units between ancestor and descendant. If t is small, the evidence may not be scant. But even when a single descendant's character state provides scant evidence about the character state of its ancestor, the evidence may be substantial if multiple descendants are all in that same state. The Markov framework goes beyond the ideas depicted in Figure 3.10, which shows how the instantaneous transition probabilities u and v are related to branch transition probabilities. The framework additionally assumes that the characteristics of two descendants in a phylogenetic tree are screened-off from each other by the characteristics of their most recent common ancestor.[21] This screening-off assumption, you will recall, played a central role in the Reichenbachian ideas about common causes discussed in the previous chapter. If we use the screening-off idea in the context of a star phylogeny whose n descendant leaves are all in state 0, the likelihood ratio for the two competing hypotheses concerning the state of the ancestor at the root of the tree is

$$\frac{\Pr_t(D_1 = 0 \mid A = 0)}{\Pr_t(D_1 = 0 \mid A = 1)} \times \frac{\Pr_t(D_2 = 0 \mid A = 0)}{\Pr_t(D_2 = 0 \mid A = 1)} \times \cdots \times \frac{\Pr_t(D_n = 0 \mid A = 0)}{\Pr_t(D_n = 0 \mid A = 1)}.$$

This product may be a lot bigger than 1. The same point holds for a bifurcating tree.[22]

The Markov framework thus provides some good news for cladistic parsimony: when all the descendants in a phylogenetic tree are in the same state of a dichotomous character, the most parsimonious hypothesis about the state of the ancestor is also the hypothesis with the highest likelihood. The Markov framework provides a second piece of good news, but the backwards inequality and screening-off are not enough. We saw earlier that Hennig says that phylogenetics would "lose all the ground on which it stands" without the

[21] We saw in Chapter 2 that common causes sometimes fail to screen-off their joint effects from each other. This was illustrated by the example of Mom's genotype failing to screen-off one of her offspring's genotype from the other's. Screening-off fails here because there is a second common cause (Dad's genotype). In the present phylogenetic context, we can imagine something similar. Suppose that the character states of two descendants are influenced both by the state of their most recent common ancestor *and* by some environmental common cause.

[22] The stronger conclusion that an ancestor was probably in a given state can be defended if defensible prior probabilities are assigned to the state of the ancestor. Information about the evolutionary process at work in the lineage can provide such priors, as I'll discuss later.

assumption that synapomorphies are evidence of phylogenetic relatedness. Translated into a likelihood framework, this claim about synapomorphies means that if you observe three taxa *X*,*Y*, and *Z*, and find that *X* and *Y* are in character state 1 and *Z* is in character state 0, where 0 is the character state of the taxa's most recent common ancestor, then this observation (which I code as "1-1-0") favors the (*XY*)*Z* topology over its alternatives, *X*(*YZ*) and (*XZ*)*Y*. Focusing just on (*XY*)*Z* and *X*(*YZ*), the thesis to consider is this:

(SYN) $\Pr[\text{1-1-0} \mid (XY)Z] > \Pr[\text{1-1-0} \mid X(YZ)]$ if trait evolution is Markovian and 1 is apomorphic.

I call this principle "SYN" for synapomorphy. There is a simple model of the evolutionary process in which this principle is true (Sober 1988). Suppose a branching process begins with a single ur-species, S_1. This root species has two offspring (S_2 and S_3) and then S_1 dies. Then S_2 and S_3 each have two offspring, after which S_2 and S_3 die. After n generations, there are 2^n leaf species. You then sample three leaf species at random; call them *X*,*Y*, and *Z*. Consider the different degrees of relatedness that two of these leaf species (*X* and *Y*) might bear to each other. One possibility is that their most recent common ancestor existed one generation ago; in this case *X* and *Y* are siblings. Let's represent that fact by saying that *X* and *Y* are 1-related, writing $R_1(X, Y)$. Another possibility is that the most recent common ancestor of *X* and *Y* existed two generations back; in that case, *X* and *Y* are first cousins, so we write $R_2(X, Y)$. If *X* and *Y* have the most distant degree of relatedness, they are *n*-related. In this branching process, the closer the degree of relatedness, the lower its prior probability. A given leaf has one sibling, two first cousins, four second cousins, and so on.

Now suppose you score *X*,*Y*, and *Z* for a dichotomous character and obtain the 1-1-0 pattern where 0 is the state of the ur-species S_1 at the root of the tree. Does this observation favor (*XY*)*Z* over each of the two other alternatives? Notice that the (*XY*)*Z* hypothesis encompasses many possible degrees of relatedness. For example, (*XY*)*Z* might be true because *X* and *Y* are 2-related and *Y* and Z are 4-related; another possibility is that *X* and *Y* are 3-related and *Y* and *Z* are 9-related. The likelihood of (*XY*)*Z* sums over all these possibilities:

$$\begin{aligned}&\Pr[\text{1-1-0} \mid (XY)Z]\\&\quad = \sum_{i,j} \Pr[\text{1-1-0} \mid R_i(X,Y) \& R_j(Y,Z)] \Pr[R_i(X,Y) \& R_j(Y,Z) \mid (XY)Z].\end{aligned}$$

It turns out that the 1-1-0 observation favors (*XY*)*Z* over *X*(*YZ*) if the Markov framework is supplemented by the following constraint on trait evolution:

(*WTASB*) *Within-trait across simultaneous branches*: The transition probabilities that characterize a trait's evolution on a given branch also characterize that trait's evolution on all simultaneous branches.

The *WTASB* constraint says nothing about branches that exist at different times; a trait may follow different rules of evolution in different time slices. The constraint also leaves open whether it is selection or drift that governs the trait's evolution in a given time slice. Appendix 3.1 at the end of this chapter provides a proof of the fact that *WTASB* entails *SYN*.[23]

This result does not provide a likelihood justification of parsimony for cases in which the WTASB assumption is false. The question is left open. However, the result does lay to rest the idea that cladistic parsimony's interpretation of synapomorphies depends on the assumption that homoplasies are rare (or that branches have low probabilities of changing state). In the *n*-generation branching process described, there is a probability in each generation that a descendant's character state will differ from the state of its immediate ancestor. There is no assumption that this one-generation change in state is improbable, though the Markov framework does say that the probability must be less than ½ if drift is the process at work. If *p* is the average probability of a branch's changing state in one generation, then the expected number of changes in a branch that stretches *n* generations from the root to a leaf is *pn.* Whether *pn* is large or small, the ratio

$$\frac{\Pr[1\text{-}1\text{-}0 \mid (XY)Z\,]}{\Pr[1\text{-}1\text{-}0 \mid X(YZ)]}$$

is greater than 1. What the probability of homoplasy influences is *how much* 1-1-0 favors (*XY*)*Z* over *X*(*YZ*), not *whether* 1-1-0 favors (*XY*)*Z* over *X*(*YZ*). I hope you see how this point connects with what I said before concerning the problem of inferring an ancestor's character state from the character state of its descendant.

[23] The assumption of a regular branching process with random sampling from the leaves guarantees that Pr[(*XY*)*Z*] = Pr[*X*(*YZ*)] = Pr[(*XZ*)*Y*]. The ordering of likelihoods and the ordering of posterior probabilities are therefore the same.

The arguments presented in this section concern two simple problems. The first involves reconstructing the character state of an ancestor in a star phylogeny when all leaves are in the same state of a dichotomous character. The second concerns interpreting the evidential significance of a synapomorphy in the context of an evolutionary process that obeys the *WTASB* assumption. There are more problems to consider, and not all of them provide good news for cladistic parsimony. But there is some more good news, to which I now turn.

Two models that entail mirroring

Felsenstein (1973) proved the following result: for any data set D, and for any two trees T_1 and T_2 where T_1 is more cladistically parsimonious than T_2 with respect to D, $\Pr_M(D \mid T_1) > \Pr_M(D \mid T_2)$ if the model M says that each character on each branch of the two phylogenetic trees has a very low probability of changing state. The model isn't just saying that the instantaneous probabilities u and v are small (of course they are); the assumption is that branch transition probabilities of the form $\Pr_t$ (end in state i | start in state j) for $i \neq j$ are small for the values of t that are relevant to the tree being considered.

Felsenstein's result was widely misinterpreted. The word on the street was that Felsenstein showed that cladistic parsimony *requires* the assumption that change is improbable. What is true is that this assumption *suffices* to justify parsimony (if we use the law of likelihood as our gold standard for evaluating evidence). Although Felsenstein (1973, p. 243) is careful to say that his result provides a "sufficient condition," he goes on to make a comment that may have encouraged the misapprehension. He says that "if we relax the assumption that the probability of change is small, there is no necessary connection between likelihood and parsimony" (Felsenstein 1973, p. 245). This is ambiguous; it may mean that when you relax the assumption, the parsimony ordering of trees *may* differ from the ordering in terms of likelihood, or it could mean that the two orderings *must* fail to match.

The ambiguity was resolved when Tuffley and Steel (1997) described a very different evolutionary model that also suffices to make more parsimonious trees more likely. They do not require that the probability of change is very low, but they do require that each trait evolves by drift. Their model is one of "no common mechanism," meaning that each possible change for each trait on each branch will obey its own rules of neutral evolution.[24]

Felsenstein (1973, p. 244) says that there is a "fly in the ointment" in connection with his result. If change is extremely improbable, then most of the traits in one's data should fail to vary across leaf taxa, a very small number of characters should involve just one change, and an even smaller number should involve two changes. Real data sets rarely look like this. As in our hypothetical example of the hundred synapomorphies, it often turns out that all the candidate trees entail that homoplasies are common. Tuffley and Steel's (1997) model does not suffer from this deficiency, though adaptationists may complain that the model assumes that the traits that serve as data in phylogenetic inference all evolve by neutral evolution.[25]

Setting aside these questions concerning the realism of each model, we can learn something interesting about what parsimony does *not* presuppose. For the most parsimonious tree to be the tree of maximum likelihood, it is not required that change be very improbable, nor is it required that drift be the process governing trait evolution (Sober 2009b). Neither of these assumptions is *necessary* for the most parsimonious tree to have the highest likelihood, though each is sufficient. Cavalli-Sforza and Edwards, and Farris, were on the right track.

So what *does* cladistic parsimony presuppose? Felsenstein and Tuffley and Steel were looking for assumptions about the evolutionary process that get parsimony and likelihood to *agree*. This makes sense if the goal is to find a likelihood justification for parsimony. But if the goal is to find out what parsimony presupposes, we must shift gears. To uncover parsimony's

[24] Tuffley and Steel (1997) use a characteristically frequentist strategy for calculating the likelihood of a tree topology. As Figure 3.10 shows, it isn't so straightforward to say what the value is of $\Pr_t$(the branch ends in state i | the branch starts in state j); the value depends on the values of u, v, and t. Tuffley and Steel assign to a branch's probability of changing state the maximum value that is permitted by their assumed model. If a branch ends in a state that differs from the one in which it began, that maximum value will be ½ if the trait is dichotomous. Using language from the previous chapter's discussion of model selection, we can say that Tuffley and Steel (1997) calculate Pr[data | $L[(XY)Z]$, not Pr[data | $(XY)Z$]; the latter is an average, while the former is a best-case assignment of values to branch transition probabilities.

[25] This problem will not arise if the characters in the data are chosen because they are known to evolve by drift.

presuppositions, we need to find cases in which parsimony and likelihood *dis*agree. If H_1 is a more parsimonious explanation of the data than H_2 is, and empirical assumption A entails that H_1 has the lower likelihood, we can conclude that cladistic parsimony assumes that A is false.

Symplesiomorphies

Cladistic parsimony claims that synapomorphies are evidence of relatedness *and* that symplesiomorphies are not. This point was illustrated in Figure 3.4. I discussed the first of these claims and showed that it is true in the context of a model that makes the WTASB assumption – that the probabilities that govern a trait's evolution on one branch of a phylogenetic tree also govern its evolution on all simultaneous branches. I now want to explore what this model of the evolutionary process says about symplesiomorphies. If you observe three species (X,Y,Z) and observe that X and Y are in state 0 while Z is in state 1, and you know that 0 is the state of their most recent common ancestor, then the three trees (XY)Z, X(YZ), and (XZ)Y are equally parsimonious; each requires a single change from 0 to 1 to explain the observations. Does this judgment, that symplesiomorphies are evidentially meaningless, have a likelihood justification? That is, is the following principle true?

(SYMP) $\Pr[0\text{-}0\text{-}1 \mid (XY)Z] = \Pr[0\text{-}0\text{-}1 \mid X(YZ)]$ if 0 is plesiomorphic.

It turns out that SYMP is false in the WTASB model; the " = " in SYMP should be changed to ">" (see Appendix 3.1). In that model, both synapomorphies *and* symplesiomorphies are evidence of genealogical relatedness. For cladistic parsimony, the WTASB model is a glass half full (= half empty).

How does this finding comport with the results derived by Felsenstein (1973) and Tuffley and Steel (1997)? The first point is that the Tuffley and Steel model does not obey the WTASB constraint. Consider what their model says about the transition probabilities that attach to branches in an X(YZ) tree when the observation is a 0-0-1 symplesiomorphy and M is the most recent common ancestor of Y and Z. The no-common-mechanisms model says that the probability of a change in state in the branch leading from M to Z is ½ and that all other branches have a probability of 0 of exhibiting a change in state. The second point is that the (XY)Z tree also confers on the 0-0-1 observation a probability of ½. The model entails that symplesiomorphies are uninformative, which is what cladistic parsimony says.

The relationship of Felsenstein's (1973) model to SYMP is more subtle. I began my discussion of Felsenstein's paper by saying that it establishes a result concerning trees that differ in parsimony: if T_1 is more parsimonious than T_2, then T_1 will have the higher likelihood when the probability of change is small enough (though non-zero). However, there is more to getting likelihood to mirror parsimony than this. If T_1 and T_2 are *equally* parsimonious explanations of the data, what does Felsenstein's model say about their likelihoods? He proves that as the probability of change approaches zero, the ratio of the likelihoods of two equally parsimonious trees approaches a positive finite constant. This contrasts with the case in which T_1 is more parsimonious than T_2; then the likelihood ratio approaches positive infinity as the probability of change approaches zero. In our simple example of the single 0-0-1 symplesiomorphy, Felsenstein's model says that the ratio of the likelihoods of (*XY*)*Z* and *X*(*YZ*) approaches a finite value greater than one as the probability of change approaches zero. The model therefore entails that SYMP is false. Felsenstein's model does a better job of getting likelihood to mirror parsimony when trees differ in parsimony than it does when trees are equally parsimonious. In contrast, Tuffley and Steel hit the nail on the head for both problems.

Does the falsehood of SYMP in a WTASB model represent a clash between likelihood and parsimony? I prefer to say that the conflict is among three ideas, not between two. These are cladistic parsimony (which is committed to the claim that symplesiomorphies are evidentially uninformative), the law of likelihood, and the evolutionary model I have described. Friends of the law of likelihood will conclude that the conflict is between the first item and the third. They will say that cladistic parsimony is committed to the model's being false. That may not seem like a terribly daring commitment, in that the WTASB assumption is widely thought to be false. Whether the process is selection or pure drift, the idea that a given trait has precisely the same probabilities of evolving in different simultaneous branches, is, to put it politely, an "idealization." Even so, it is interesting that symplesiomorphies are evidentially meaningful in a WTASB model; this result shows that we must abandon the idea that cladistic parsimony makes no assumptions about the evolutionary process.[26] Another interpretative option is to reject the law

[26] We already took a small step in this direction when we saw that cladistic parsimony assumes that genealogies are tree-like. Now we are seeing something more.

of likelihood. Why should cladistic parsimony be answerable to this idea from statistics? Rather than frowning on parsimony when it departs from the dictates of the law of likelihood, perhaps we should frown on the law when it departs from the dictates of parsimony. Here I ask a question that I posed in the previous chapter: which is the dog and which is the tail?

The Smith/Quackdoodle theorem

If synapomorphies and symplesiomorphies are both evidence of genealogical relatedness in a WTASB model, which provides the stronger evidence? If one trait is 1-1-0 [and therefore favors (*XY*)*Z* over *X*(*YZ*)] and a second trait is 1-0-0 character (and therefore has the opposite evidential significance), does one of those observations "trump" the other? Is the synapomorphy stronger evidence than the symplesiomorphy? I proved a result that bears on this question (Sober 1988); I called it the Smith/Quackdoodle theorem. If you meet two people named "Smith" you might take that similarity to be evidence that they are related; if you meet two people named "Quackdoodle," you might draw the same conclusion. Who do you think are more closely related – the two Smiths or the two Quackdoodles? I proved that sharing a rare name is evidence for closer relatedness than sharing a common name in the setting of a simple model about how names are inherited. This result about the two pairs of observations suggests the following conjecture about a triplet of leaf species:

(SvS) With WTASB and WBAT, $\dfrac{\Pr[\text{1-1-0} \mid (XY)Z]}{\Pr[\text{1-1-0} \mid X(YZ)]} > \dfrac{\Pr[\text{1-0-0} \mid X(YZ)]}{\Pr[\text{1-0-0} \mid (XY)Z]}$ if and only if $p < q$.

"SvS" means *synapomorphy versus symplesiomorphy*. Here p is the probability that each leaf has of being in state 1 and q is the probability that each leaf has of being state 0 ($p + q = 1$). It also is true that p is the expected frequency of character state 1 at the leaves and q is the expected frequency of character state 0. The WBAT constraint means the following:

> *Within-branch across traits (WBAT):* The transition probabilities that characterize the 1-1-0 trait's evolution on a given branch also characterize the 1-0-0 trait's evolution on that same branch.

As before, we assume that the root of the tree is in state 0 for both characters. A proof of SvS is given in Appendix 3.2.

If a tree is in character state 0 at the root and then experiences drift, the expectation is that 1 will be the minority trait at the leaves, no matter how many generations there are that separate root from leaves. On the other hand, if the tree starts in state 0 and there is selection for state 1, the expectation is that 1 will eventually become the majority trait (see Figure 3.4). This means that if a 1-1-0 trait and a 1-0-0 trait evolve on a tree in accordance with the WTASB and WBAT constraints, then cladistic parsimony is right to conclude that the synapomorphy trumps the symplesiomorphy if the two traits evolve by drift. However, the precedence given to synapomorphies is misplaced if there is selection for the derived state and there is lots of time between root and leaves. The SvS result reveals (in the context of the simple model assumed) that synapomorphies don't always provide more genealogical information than symplesiomorphies. When they do so, the reason is that a leaf has a low probability of exhibiting an apomorphy and a high probability of exhibiting a plesiomorphy.

Estimating character states of ancestors – two examples in which mirroring fails

A few pages back, I discussed the problem of inferring an ancestor's character state from the observed character states of descendants in a star phylogeny when all of the descendants are in the same state of a dichotomous character. I concluded that parsimony and likelihood offer identical solutions to this problem. Regardless of whether there is selection or drift, if the descendants are all in state 0, then the most parsimonious estimate of the state of their most recent common ancestor is also the estimate of maximum likelihood; parsimony and likelihood agree that the observations favor the conclusion that the ancestor was in state 0. This result does not depend on assuming WTASB or WBAT; all that is needed is the backwards inequality, which is entailed by the Markov framework.

Does this mean that parsimony and likelihood agree about the correct solution to *all* problems of estimating ancestral character states in a known tree? The answer is *no*. The first problem case that I want to discuss involves a "tie." Consider a star phylogeny in which half the descendants are in state 0 and half are in state 1. The two hypotheses about the character state of the most recent common ancestor are equally parsimonious. Are they equal in likelihood? Let's investigate this question by adopting the WTASB assumption – that the transition probabilities for a single trait are the same across simultaneous lineages. In the context of this assumption, the problem of half-and-half can be analyzed, without loss of generality, by considering a star phylogeny that has just two leaves, *X* and *Y*, whose most recent common ancestor is *M*. If *X* is in state 1 and *Y* is in state 0, the problem is to determine whether

$$\Pr(X{=}1 \,\&\, Y{=}0 \mid M{=}1) = \Pr(X{=}1 \,\&\, Y{=}0 \mid M{=}0)\,.$$

Assuming that *M* screens-off *X* from *Y*, this becomes

$$\Pr(X{=}1 \mid M{=}1)\Pr(Y{=}0 \mid M{=}1) = \Pr(X{=}1 \mid M{=}0)\ \Pr(Y{=}0 \mid M{=}0).$$

With the WTASB assumption, this entails that

$$\Pr(X{=}1 \mid M{=}1)\Pr(X{=}0 \mid M{=}1) = \Pr(X{=}1 \mid M{=}0)\Pr(X{=}0 \mid M{=}0)\,.$$

Notice that this last equality has the form $a(1-a) = (1-b)b$. This equality holds exactly when $a = b$, which is to say when $\Pr(X{=}1 \mid M{=}1) = \Pr(X{=}0 \mid M{=}0)$. This is true if the two lineages undergo drift, but not if there is selection for state 1 (or for state 0) in both lineages; see Figure 3.10. Once again, parsimony mirrors likelihood under some assumptions about the evolutionary process, but not under others.

In Chapter 2 I discussed an epistemological idea that I called Darwin's Principle. The principle says that when two species have trait *T*, the similarity provides weak evidence for common ancestry if *T* is adaptive, whereas the similarity provides strong evidence if *T* is neutral or deleterious. Darwin's principle is worth revisiting in the context of the present problem. If *X* is in state 1 and *Y* is in state 0 and if selection for state 1 took place in both branches, then parsimony is wrong in saying that $M{=}0$ and $M{=}1$ are equally well supported by the observations. But which hypothesis about the state of the common ancestor is more likely? The observations $X{=}1$ and $Y{=}0$ together favor $M{=}0$ over $M{=}1$ precisely when $a(1-a) < (1-b)b$. This simplifies to $(a - b) < (a-b)(a+b)$. If there is selection for state 1 in both lineages, then $a > b$, so the criterion for $M{=}0$ to have the higher likelihood is that $(1-a) < b$, which is an instance of the backwards inequality. So the answer is that $M{=}0$ has the higher likelihood when there is selection for state 1 in both lineages.[27] The

[27] A look at Figure 3.10 clarifies why $M{=}0$ has a higher likelihood than $M{=}1$ when there is selection for character state 1 and the observations are $X{=}1$ and $Y{=}0$. Notice

character state of *X* is adaptive, and the character state of *Y* is not; it is the latter that carries more weight in this inference problem. The question of which type of similarity (adaptive or non-adaptive) provides stronger evidence for the common ancestry of *X* and *Y* differs from the question of which descendant's character state (*X*'s or *Y*'s) provides stronger evidence about the state of their most recent common ancestor, assuming that there is one. Yet, the answers to the two questions are on the same page.[28]

There is one more problem of inferring the state of an ancestor that I want to consider. Once again, it is a case in which parsimony and likelihood part ways. Consider an ordered *n*-state character whose states are 1, 2, . . . , *n*. By "ordered," I mean that a lineage can pass from one state directly to another only if the two states are adjacent. For example, if a lineage is to pass from state 2 to state 4, it must go through state 3. Suppose we observe that a lineage is now in some intermediate state; for example, suppose it now is in state 8 and $n = 20$. What is the best estimate of the ancestral state of the lineage, say, a hundred time intervals ago? Cladistic parsimony says that the best estimate is 8. What does likelihood say? If the lineage is experiencing pure drift, likelihood agrees with parsimony (Tuffley and Steel 1997, Theorem 6). But suppose there is directional selection, so that at each time step the lineage has a higher probability of having its trait value move towards *n* than away from it. Given this, the maximum likelihood estimate of the ancestral condition will be less than 8. How much less will depend on the amount of time separating ancestor and descendant and on the strength of selection. If this claim about the maximum likelihood estimate puzzles you, consider an analogy. You observe a log floating in a river; it is on the right bank at a certain location. Suppose that the log got to this point on the right bank by starting on the left. What is the maximum likelihood estimate of where on the left bank the log began? If there is a current in the river, the estimate is *not* that the log began directly across from where it is now. The estimate is that the log began upstream; how much upstream will depend on the strength of the current and the width of the river (Sober 2009b, p. 209; Gascuel and Steel 2010).

that at all times, a is greater than both b and $1 - b$, and each of these is greater than $1 - a$. This is why $a(1 - a) < b(1 - b)$.

[28] For further discussion of how the type of evolutionary process affects inferences about the past state of a lineage based on an observation of its present state, see Sober and Steel (2014).

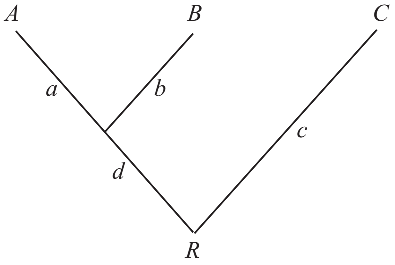


Figure 3.11

Another criterion – statistical consistency

I have been using the law of likelihood to judge when cladistic parsimony does a good job in phylogenetic inference problems. However, there is another criterion that has impressed many phylogenetic biologists – the criterion of statistical consistency. I'll start by giving an informal characterization of that concept, using the example of estimating a coin's bias. Suppose you toss the coin a hundred times and get 51 percent heads. The maximum likelihood estimate of the coin's bias, its probability of landing heads when tossed, is $p = 0.51$; this estimate of the bias confers a probability on the observed frequency that is greater than the probability conferred by any other estimate. Suppose you toss the coin another 100 times and obtain 53 heads. Using your total evidence of 104 heads in 200 tosses, you replace your old estimate of the coin's bias with a new one, that $p = 0.52$. This is the new maximum likelihood estimate. If you gather more and more data and revise your estimate again and again, you should expect your estimate to converge on the true value of p, whatever that true value happens to be.[29] In this example, maximum likelihood estimation is statistically consistent.

What would it take for cladistic parsimony to be statistically consistent? The "parameter" you are estimating is a tree's topology, and you base this estimate on the observations you make of leaf taxa. Consider the simple example represented in Figure 3.11, where *a*, *b*, *c*, and *d* label the branches of the (*AB*)*C* tree. Suppose this tree is true, that all the characteristics you are going to observe in the leaves are dichotomous (0 and 1 are the two states), and

[29] This should remind you of the law of large numbers described at the start of the previous chapter. Convergence does not require a lockstep monotonic approach to the coin's true value as the number of tosses increases.

that 0 is the ancestral condition of each character. Cladistic parsimony interprets 1-1-0 characters as evidence for (*AB*)*C* and 1-0-1 characters as evidence for (*AC*)*B*. Parsimony therefore will mislead you if 1-0-1 characters outnumber 1-1-0 characters in your data set. And if

(*) Pr [1-0-1 | (AB)C] $>$ Pr [1-1-0 | (AB)C] for each character,

then the probability that you will be misled increases as you gather more and more evidence. If the (*) inequality is true, parsimony will be statistically inconsistent as an estimator of the genealogical relationship of *A*, *B*, and *C*.

It is important to see that asking whether cladistic parsimony is statistically consistent and asking whether it mirrors likelihood are distinct questions. Consider the accompanying 2×2 table, each cell of which represents a probability of the form Pr(data | genealogy). To investigate whether parsimony mirrors likelihoods, you consider "horizontal" comparisons – whether $w > x$ *and* $z > y$. To investigate whether parsimony is statistically consistent, you look at "vertical" comparisons – whether $w > y$ and $z > x$.

	(*AB*)*C*	(*AC*)*B*
1-1-0	*w*	*x*
1-0-1	*y*	*z*

To determine whether parsimony is statistically consistent, Felsenstein (1978) constructs a simple model of the evolutionary process. A slightly simplified rendition of Felsenstein's argument can be given by using the (*AB*)*C* tree in Figure 3.11. Here is his model:

(i) WBAT: all characters on the same branch evolve by following the same rules.

(ii) No reversals: $\Pr(1 \to 0) = 0$ for each trait on each branch.

(iii) Branches are paired up: $\Pr_a(0 \to 1) = \Pr_c(0 \to 1) = p$ and $\Pr_b(0 \to 1) = \Pr_d(0 \to 1) = q$.

Given these assumptions, Felsenstein shows that the (*) inequality will be true if p is substantially bigger than q. There is a region (labeled "FZ") in the unit square shown in Figure 3.12 where parsimony is statistically inconsistent.

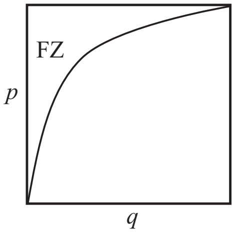


Figure 3.12

This region has come to be called "the Felsenstein zone." This is a place, like *The Twilight Zone* on 1960s television, where you are apt to be misled.[30]

Farris (1983) thinks that Felsenstein's result leaves parsimony unblemished. His reason is that Felsenstein's model of the evolutionary process is extremely unrealistic. Felsenstein (1978) agrees that his model is unrealistic, but notes that it would be naïve to assume that parsimony will be consistent in more realistic models of the evolutionary process.[31] Farris is a friend of parsimony and Felsenstein is a foe, but there is a point on which they agree; they agree that it is reasonable to use parsimony to estimate phylogenetic relationships only if you have reason to think that it will be statistically consistent. If this assumption is correct, Felsenstein's argument establishes that parsimony makes an assumption about the evolutionary process: parsimony assumes that the conjunction of (i), (ii), (iii), and $p \gg q$ is false.[32] Farris (1983) notes that rejecting this conjunction (which does not require rejecting every conjunct!) is a very safe commitment. As before, the assumptions of parsimony, as judged by some criterion, become visible when parsimony fails to satisfy the criterion, not when it succeeds.

[30] Biologists use the phrase "long branch attraction" to describe what happens in Felsenstein's example. In Figure 3.11, branches *a* and *c* are "long," not in terms of their durations, but in terms of the expected amount of change that will occur on them.

[31] The question of parsimony's consistency in other models of evolution has not been much explored. Here's a small and unsurprising result: given the WTASB assumption (that a trait's rules of evolution are the same across simultaneous branches) and the assumption that evolution is Markovian, $\Pr[1\text{-}1\text{-}0 \mid (AB)C] > \Pr[1\text{-}0\text{-}1 \mid (AB)C]$, $\Pr[0\text{-}1\text{-}1 \mid (AB)C]$ for each character. See Appendix 3.1.

[32] I question whether consistency is a necessary property of an acceptable estimator in Sober (1988, pp. 172–183).

It would be nice if there were a method of phylogenetic inference that guarantees that you will converge on the truth as you gather more data, regardless of what that truth happens to be. You then could set parsimony to one side and step onto firmer ground. Felsenstein (1978, p. 408) seems to suggest that there is a method that has this virtue; he says that "it can be shown quite generally that the maximum likelihood estimation procedure has the property of consistency." The contrast Felsenstein draws between the method of maximum likelihood and the method of maximum parsimony needs to be understood carefully. The description of statistical consistency given at the start of this section fails to bring out one of its important features. The missing nuance can be understood via the example of coin tossing. If the tosses of a coin *are* independent and identically distributed (i.i.d.), and you *assume* that the tosses are i.i.d. and use that assumption to construct a maximum likelihood estimate of the one parameter in that model (which represents the coin's unwavering probability of landing heads), then maximum likelihood estimation will converge on the truth as you gather more and more data. Notice the two mentions of the i.i.d. model in the previous sentence; the sentence describes what will happen *if* the model is true *and* is used in the estimation procedure. Here is a more careful definition of statistical consistency for the coin-tossing problem that brings out the double role played by the i.i.d. model:

> $\mathrm{MLE}_{\text{i.i.d.}}(-)$ is a consistent estimator of the coin's bias precisely when, for any $\varepsilon > 0$,
>
> $\Pr[s$ is within ε of $t \mid \mathrm{MLE}_{\text{i.i.d.}}(O_1 \& O_2 \& \ldots \& O_n) = s$ & the i.i.d. model is true & t is the true bias of the coin$] \to 1$ as $n \to \infty$.

Here s is the maximum likelihood estimate of the coin's bias obtained from n observations and ε (epsilon) is any margin of error you please. What is defined is not the consistency of MLE (full stop) but of MLE in the context of an assumed model – hence the "i.i.d." subscript. Maximum likelihood estimation of the value of a parameter is always estimation of the value of a parameter in some model or other.[33] This pattern transfers to using maximum likelihood in phylogenetic inference:

[33] Recall the comment in Chapter 2 that parameters are always "creatures" of a model.

> $\text{MLE}_{\text{M}}(-)$ is a consistent estimator of the tree topology of a set of leaf taxa precisely when
>
> $\Pr[s = t \mid \text{MLE}_M(O_1 \& O_2 \& \ldots \& O_n) = s \& M \text{ is true } \& t \text{ is the true topology}] \rightarrow 1$ as $n \rightarrow \infty$.[34]

If you estimate the topology by assuming that M is true, and M is false, then you have no assurance that your maximum likelihood procedure will converge on the true topology as you gather more and more data (Gaut and Lewis 1995; Steel and Penny 2000). In this case, statisticians say that the model has been *misspecified*. This shows that maximum likelihood estimation of trees and maximum parsimony estimation of trees have something in common: in both cases, your estimate can fail to converge on the truth if nature does not cooperate. This point also applies to the mundane example of coin tossing; if the coin is not i.i.d. but you assume that it is and set about estimating the value of p, your estimates won't converge on the true value of p since there is no such thing.

What if you use a model of the evolutionary process to obtain a maximum likelihood estimate of the tree topology for a set of leaf taxa and the model is true? Surprisingly, statistical consistency can fail even here. Tuffley and Steel's (1997) no-common-mechanism model is an example. Convergence requires there to be a fixed set of model parameters that you gain more and more information about as you gather more data. In Tuffley and Steel's model, each new character brings with it a new suite of parameters to estimate. In this model, parsimony mirrors likelihood, and both are statistically inconsistent.

Still another criterion – when is parsimony more reliable than guessing?

In this chapter, I first considered likelihood as the gold standard for evaluating cladistic parsimony and then turned to statistical consistency as another possible criterion. Now it is time for a third. If the lineage leading to the most recent common ancestor of the leaves of a tree is subject to the same process that occurs in the tree's interior, and that lineage has been around for a good

[34] The business about s's being within ε of t that I used to explain what statistical consistency means in the coin-tossing example isn't needed in this case since tree topology is a discrete parameter whereas the bias of a coin is a continuous quantity.

long time, you can use a model of that process to obtain prior probabilities for the state of the most recent common ancestor. For example, when the process in question involves the evolution of a dichotomous character where the instantaneous probabilities for change are u and v, the prior probabilities for the two possible states of the most recent common ancestor are $\frac{u}{(u+v)}$ and $\frac{v}{(u+v)}$. When drift is the process at work (so $u = v$), the priors each have a value of ½. We now can ask when parsimony is more reliable in its reconstruction of the most recent common ancestor's character state than the method of ignoring the states of the leaves and using the prior probability as one's guess. In the present context, this means asking when parsimony will correctly infer the state of the root with a probability that is greater than ½. This is a Bayesian question, but one that is not to be sneezed at. It can be as well-grounded as the question I asked in Chapter 2 concerning Susan's tuberculosis test. Prior probabilities and likelihoods are justified by an empirical theory, and posterior probabilities therefore are too. If you think the empirical theory is plausible, you should be willing to ask (and happy to answer!) the Bayesian question.

Whether parsimony does better than guessing depends on the relationship of two processes. One of them is the process of character change that occurs in branches; it is described by the instantaneous probabilities u and v as just noted, but it is also described by the substitution rate m, which can range from zero changes in a million years, to one change in a million years, to a million changes in one year, and beyond. The other process is the branching process itself. Suppose the rate at which nodes give birth to daughter nodes is λ; it also ranges from 0 to positive infinity. Gascuel and Steel (2010) discovered, for dichotomous traits evolving by drift, that when $\lambda/m < 6$, parsimony does no better than guessing the root state (as $t \rightarrow \infty$). However, when $\lambda/m > 6$, the probability that parsimony will get the root state right (as $t \rightarrow \infty$) is strictly greater than ½ and converges to 1 as $\lambda/m \rightarrow \infty$. In other words, the reliability of parsimony in this inference problem requires the rate of branching to be appreciably greater than the rate of character substitution.

Forwards and backwards

In all the problems that I have posed thus far for cladistic parsimony to address, the task is to infer past from present. This is true whether you want to infer a tree's topology or the character states of ancestors in a known tree. Even though this is what the usual applications of cladistic parsimony look like, the method, taken at its word, does not limit itself to these backwards-directed tasks. Instead of trying to infer an ancestor's character state from the observed character state of its descendant, you might want to infer the character state of a descendant from the observed character state of its ancestor. If the ancestor is in state 0, cladistic parsimony says that the best estimate of the descendant's character state is also state 0. Markov models of the evolutionary process are in agreement if the process is one of drift, but not if there is selection for state 1 and the lineage is of sufficiently long duration. As noted earlier, the forwards inequality (FI) isn't always true. Notice that FI is a claim about the *probability* of the descendant's state, not the *likelihood* of that state. This brief comment on forward-directed inference, coupled with the earlier discussion of backwards-directed inferences, leaves one more problem to address – the use of cladistic parsimony to infer the state of one present day species from the state of another. This synchronic problem is the final topic of the chapter.

Estimating the character state of a leaf – anthropomorphism and comparative psychology

Morgan's canon, an inference principle that I discussed in Chapter 1, continues to exert considerable influence on psychology in general and on comparative psychology in particular.[35] Morgan, recall, puts the canon like this:

> In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale. (Morgan 1894, p. 53)

Although most friends of the canon have thought that it is a version of Ockham's razor, this was not how Morgan saw things. On the contrary, for him the point of the canon was not that we should seek *simple* theories but that we should reign in our inclination to *over*-simplify (Morgan 1894, p. 54).

Following Morgan's canon reduces the risk of making a mistake – the mistake of *anthropomorphism*. We know that human beings have higher mental faculties. For example, humans often form beliefs about the mental states of other human beings; we have beliefs about what other people see, want, and believe. This is what cognitive psychologists call *mind-reading*. Do non-human

[35] In this section I use some material from Sober (2012a).

organisms have the same ability? Anthropomorphism encourages us to think that the answer is *yes.* Morgan's canon aims to prevent us from falling into the error of mistakenly attributing such mental states. If we *can* explain the behaviors of non-human organisms without viewing them as mind-readers, the canon tells us that we *ought* to do so.

		possible states of the world	
		O does not have *M*	*O* has *M*
possible actions	believe that *O* does not have *M*		type-2 error
	believe that *O* has *M*	type-1 error	

It is true that obeying Morgan's canon reduces the risk of one kind of error, but there is another kind of error that the canon does nothing to prevent. The two types of error are depicted in the accompanying table. The rows represent possible actions – you can believe that a non-human organism *O* has the higher mental state *M* or believe that it does not. The columns represent possible states of the world – either the organism has *M* or it does not. Each of the four cells represents the outcome of performing an action given a possible state of the world. Type-1 error is the error of mistaken anthropomorphism. Type-2 error is also an error, but it does not have a standard name. The primatologist Frans de Waal (1999) coined one; he calls it the error of mistaken *anthropodenial.* Is one type of error worse than the other? If you engage in anthropomorphism, people may call you tender-minded; if you engage in anthropodenial, people may call you tough-minded (James 1907). Surely an exclusive preoccupation with avoiding just one type of error is a mistake.

Morgan thought that the simplest explanation of the behavior of non-human organisms is anthropomorphic. He was right, if we understand simplicity in terms of cladistic parsimony (not that he thought of simplicity in that way). To see why, consider the problem represented in Figure 3.13. We observe that humans and chimpanzees both perform behavior *B*, and we know that humans do so by using the proximate mechanism *M*. We assume that humans and chimpanzees have a common ancestor. The box in the figure represents their most recent common ancestor (MRCA). What inference should be drawn concerning the proximate mechanism that chimpanzees deploy? To make the problem simple, suppose that there are exactly two proximate mechanisms (*M* and *N*) that could produce the behavior *B* and that each proximate mechanism suffices for *B* to occur.

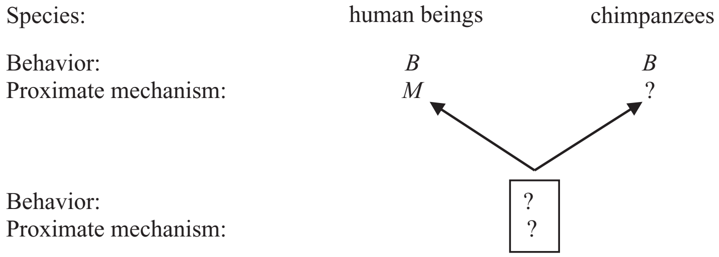


Figure 3.13

Since humans and chimpanzees perform behavior *B*, the most parsimonious hypothesis about the state of the MRCA is that it performed *B* as well. This assignment is most parsimonious because it requires no change in character state between MRCA and its two descendants. The parsimonious assignment of *B* to MRCA means that the MRCA used either *M* or *N*. Which of these two proximate mechanisms should be assigned to the MRCA, and which to chimpanzees? The most parsimonious hypothesis is that both use *M*. Cladistic parsimony endorses anthropomorphism in this simple inference problem.

Notice that the logic of this cladistic inference applies to any pair of related species that share a phenotype where the proximate mechanism that causes this phenotype is known for one species but unknown for the other. For example, if species *X* and *Y* have a common ancestor and both transport oxygen to their tissues, with *X* doing this by using the hemoglobin molecule, then the most parsimonious hypothesis is that *Y* also uses hemoglobin to transport oxygen. The cladistic argument has nothing essentially to do with *us* or with the fact that the proximate mechanism is *psychological.*

De Waal (1991) notes that cladistic parsimony endorses anthropomorphism:

> By far the simplest assumption regarding the social behavior of the chimpanzee, for example, is that if this species' behavior resembles that of ourselves then the underlying psychological and mental processes *must* be similar too. To propose otherwise requires that we assume the evolution of divergent processes for the production of similar behavior. (p. 298)

> The most parsimonious assumption concerning non-human primates is that if their behavior resembles human behavior the psychological and mental processes involved are *probably* similar too. (p. 316)

De Waal isn't just *describing* what parsimony considerations entail; he *embraces* the anthropomorphic conclusions (see also de Waal 1999, 2009).

How strong an inference can one draw about the characteristics of chimpanzees, based just on the observation of a single, albeit closely related, species? De Waal says "must" and "probably." Surely the conclusion that chimpanzees *must* be like us is too strong. And perhaps the conclusion that they *probably* are like us is too strong as well (Karin-D'Arcy 2005, pp. 185–186). But even if De Waal's formulations go too far, there remains a more modest anthropomorphic thesis with which to reckon. This is the claim that if human beings and chimpanzees both exhibit behavior B and if human beings produce B by using mental state M, then this is *evidence* that M is also the proximate mechanism that chimpanzees deploy in producing B. It is useful to understand this thesis by using the Bayesian concept of confirmation discussed in Chapter 2: given background assumptions A, observation O confirms hypothesis H if and only if $\Pr_A(H \mid O) > \Pr_A(H)$. Applied to the case at hand, the thesis is that

$$\text{(ANTHROPO-CONF)} \qquad \Pr_A(\text{chimpanzees have } M \mid \text{humans have } M) > \Pr_A(\text{chimpanzees have } M).$$

The background assumptions A in ANTHROPO-CONF include the fact that humans and chimpanzees both have the behavioral trait B and that M is a proximate mechanism for producing B. I want to explore whether this inequality is justified by the fact that humans and chimpanzees have a common ancestor. Notice ANTHROPO-CONF does not entail that $\Pr_A(\text{chimpanzees have } M \mid \text{humans have } M) > 1/2$, which is why this evidential thesis about anthropomorphism is more modest than the two formulations of De Waal's that I quoted.

As noted earlier in this chapter, cladistic parsimony now is controversial in evolutionary biology. Many evolutionary biologists appeal to parsimony when they make phylogenetic inferences, but many others prefer to use statistical inference procedures that explicitly assume a particular model of the evolutionary process and make no mention of parsimony. Skeptics about parsimony may be inclined to reject De Waal's parsimony argument for anthropomorphism and to draw the same negative verdict about ANTHROPO-CONF. This, I suggest, is throwing out the baby with the bathwater. ANTHROPO-CONF can be defended on grounds that have nothing to do with parsimony.

The argument I have in mind for ANTHROPO-CONF is a consequence of an idea from Chapter 2. It is due to Hans Reichenbach (1956), who wasn't thinking about anthropomorphism or evolution when he proved his theorem. Reichenbach's more general and abstract project was to describe a probability model in which two effects trace back to a common cause. Reichenbach's model applies to the problem at hand if we think of human beings and chimpanzees as effects and their most recent common ancestor as the common cause. Here is Reichenbach's theorem, applied to the case at hand: if human beings and chimpanzees have a most recent common ancestor and if the relationship of the two effects to their common cause satisfies three assumptions, then the traits of human beings and chimpanzees will be correlated. The three assumptions, stated more carefully in Chapter 2, are these: (i) all the probabilities in the common cause model are strictly between 0 and 1; (ii) the common cause screens-off the two effects from each other; and (iii) there is a positive correlation between the common cause and each of the two effects. Reichenbach showed that these three assumptions entail that the two effects will be positively correlated:

$$\begin{aligned}&\Pr(\text{chimpanzees have } M \,\&\, \text{humans have M})\\&\quad > \Pr(\text{chimpanzees have } M)\Pr(\text{humans have } M).\end{aligned}$$

This correlation claim entails ANTHROPO-CONF. To see why, think about the "definition" of conditional probability discussed in Chapter 2. The fact that humans have trait *M* raises the probability that chimpanzees do too. Reichenbach's three assumptions are routine in Markov models of the evolutionary process; these assumptions are neutral on whether evolution is driven by drift or by selection and also on whether evolution obeys the WTASB or the WBAT assumptions discussed earlier in this chapter.

ANTHROPO-CONF says that observing that humans have *M* is evidence that chimpanzees do too. But how strong is the evidence thus provided? Here we need a measure of the degree to which observation *O* confirms hypothesis *H*. As mentioned in Chapter 2, several measures of Bayesian *degree of confirmation* have been proposed. The measure I want to consider here is the likelihood ratio:

$$\frac{\Pr(\text{humans have } M \mid \text{chimpanzees have } M)}{\Pr(\text{humans have } M \mid \text{chimpanzees lack } M)}.$$

We know that human beings have *M*. To what degree does that observation favor the hypothesis that chimpanzees have *M* over the hypothesis that they do not? Reichenbach's assumptions entail that this ratio is larger than one. But how much larger is it? The bigger the likelihood ratio, the stronger the favoring.

To make this likelihood ratio big, there are two steps:

(1) First make Pr(humans have *M* | MRCA has *M*) big and make Pr(humans have *M* | MRCA does not have *M*) small.
(2) Then make Pr(chimpanzees have *M* | MRCA has *M*) big and make Pr(chimpanzees have *M*) small.[36]

There are two steps here because if you don't do the first one, the second won't help; see Appendix 3.3 of this chapter for an explanation of why this is so. The three conditional probabilities in these two steps all describe the probability of a descendant's having *M* given the state of an ancestor. As discussed earlier in the present chapter, Markov models of evolution say that there are two factors that affect how big these probabilities are – the amount of time between ancestor and descendant and the probability that a lineage will change state in a brief moment of time. Making both of these small has the effect of getting the three conditional probabilities to do the right thing. However, the mere fact that the MRCA of humans and chimpanzees is a mere six million years ago isn't sufficient, since it says nothing about the probability of change per unit time. What about the unconditional probability mentioned in the second step? If the behavior *B* is rare, *M* must also be rare since *M* entails *B*; this might make it reasonable to estimate that Pr(chimpanzees have *M*) is small.

De Waal (1991) says that 6 million years is too short a time for humans and chimpanzees to have different proximate mechanisms for producing a shared behavioral trait *B*. As just noted, this point about time isn't sufficient to get the likelihood ratio to be big. But, in addition, it should raise eyebrows when viewed against the backdrop of the growing number of documented cases of rapid evolution. Consider, for example, the evolution of adult lactose

[36] Although I here discuss what it takes to make the likelihood ratio big, the same considerations pertain when other standard Bayesian measures of degree of confirmation are used, a result that Branden Fitelson and William Roche (in personal communications) independently confirmed by simulations.

tolerance, a trait that some 25 percent of present day human beings possess. This phenotype evolved since agriculture began about 10,000 years ago, and the trait evolved at least twice, with different gene complexes evolving in different human populations (Tishkoff *et al.* 2007). This example does not show that 6 million years suffices for humans and chimpanzees to evolve different mental mechanisms (*M* and *N*) for producing behavior *B*. But it does show that the claim that "6 million years is not enough" requires a detailed argument that focuses on the specifics of the traits involved. Cladistic parsimony does not supply that argument, nor does the fact that 6 million years is a mere flicker in the 3.9 billion years since life on Earth began.

Biologists are now exploring data and theory concerning how an ancestor can have a phenotype that it transmits to two of its descendant species and yet the underlying genetic mechanisms in the two descendants are different. True and Haag (2001) call the process that produces this outcome "developmental systems drift" (DSD), but they do not mean to restrict the process to neutral evolution. For example, suppose that an ancestor has behavior B_1 (owing to the presence of gene G_1) and transmits both the phenotype and the gene to two descendant species. Both descendants initially lack behavior B_2, but then a mutation occurs in the second species; the new gene (G_2) causes both B_1 and B_2. Because B_1 and B_2 are both favored by selection, G_2 replaces G_1 in the second species. A G_2 mutation may never occur in the first species, but even if it does, perhaps B_2 will be disadvantageous in that species; in either case, the first species sticks with G_1. The result is that the two species have a behavioral similarity (B_1) but for different genetic reasons. Selection has produced an "unparsimonious" arrangement in which similar effects do not have similar causes. True and Haag (p. 101) say that "at the short end of the time scale, there is a great deal of evidence of DSD between recently diverged species."

When humans and chimpanzees both produce behavior *B*, is the fact that humans produce this behavior by being in mental state *M strong* evidence that chimpanzees also use *M* to produce *B*? I have argued that the answer is *no.* But drop the word "strong" and a weaker thesis appears: the fact that humans have *M* is *evidence* that chimpanzees do too. This modest anthropomorphism is the kernel of truth that remains after the chaff has been discarded. Cladistic parsimony has a likelihood justification in this inference problem and the Reichenbachian assumptions that provide that justification are often satisfied. However, if the link between parsimony and likelihood is not to be shattered, the cladistic argument must not be overstated. It does not show that chimpanzees *must* have mental mechanism *M* nor does it show that they *probably* do.

Parsimony old and new

At the same time that cladistic parsimony faced criticism in phylogenetic biology, another kind of parsimony attracted attention. Biologists started using model selection criteria like AIC in preference to older forms of likelihood inference that require the investigator to commit to a single model of the evolutionary process and then decide, using that assumed model, which tree topology is best (Kishino and Hasegawa 1990; Posada and Crandall 2001; Posada and Buckley 2004). The limitation of the old likelihood approach is that biologists are often uncertain as to which processes have governed the evolution of the traits in their data. In a model selection framework, it is possible to use multiple process models and to consider different tree topologies in the light of each. Models are penalized for their complexity, but the meaning of the penalty has changed. In cladistic parsimony, you count the changes in character state that a tree topology requires if it is to account for the data; in model selection, you count the adjustable parameters in a model of the evolutionary processes. Whereas cladistic parsimony goes straight from a calculation of how parsimonious a tree is to a judgment about the tree's plausibility, model selection takes parsimony as one of two considerations, the other being fit to data. For further discussion, see Sober (2008b, pp. 332–342).

Concluding comments

In this chapter I have examined the relation of cladistic parsimony and likelihood in three phylogenetic inference problems: inferring a tree's topology, inferring the character states of ancestors in a tree that is assumed to be true, and inferring the character state of one leaf species from the known character state of another. The most general conclusion I have drawn is that cladistic parsimony makes assumptions about the evolutionary process. However, these assumptions are not what you might have guessed.

Thanks to Tuffley and Steel (1997), we know that cladistic parsimony does not assume that homoplasies are rare. Thanks to Felsenstein (1973), we know that the method does not assume that it is drift, rather than selection, that drives trait evolution. In addition to these negative claims concerning what cladistic parsimony does not assume, I defended some positive claims about what it does assume. Using the law of likelihood, I argued that cladistic parsimony must reject the WTASB model, since that model entails that symplesiomorphies provide evidence of genealogical relatedness. This conclusion holds regardless of whether it is selection or drift that the WTASB model describes.

I then turned to the use of cladistic parsimony in inferring the character states of ancestors. When all leaf species are in the same state of a dichotomous character, parsimony mirrors likelihood. This conclusion does not depend on special assumptions about trait evolution, but follows just from the idea that trait evolution is a Markov process. However, when half the leaves in a star phylogeny are in state 1 and half are in state 0, the problem of inferring the character state of their most recent common ancestor has a different solution. If we adopt the WTASB model, parsimony and likelihood agree if drift is the process, but not if there is selection. And when an ordered n-state character ($n > 2$) evolves in a single lineage, whether parsimony and likelihood draw the same conclusion about an ancestor's character state depends on whether drift or selection is the process at work.

Finally, I considered an inference that has been proposed in comparative psychology; it draws conclusions about the mental states of chimpanzees from premises concerning the mental states of human beings. Cladistic parsimony has a likelihood rationale in this instance (and there is no need to decide about WTASB and WBAT or about drift versus selection), as long as the cladistic inference is not over-stated. Cladistic parsimony correctly describes how evidence concerning the state of one leaf species discriminates between hypotheses concerning the unobserved character state of another. However, it is left open whether the evidence thus provided is very strong, very weak, or somewhere in between.

Not only do the assumptions of cladistic parsimony diverge from what many biologists have thought; the assumptions vary from problem to problem. For example, the WTASB model leads parsimony and likelihood to coincide in what they say about synapomorphies, but to disagree in what they say about symplesiomorphies. And when it comes to inferring the character state of a common ancestor from the observed character states of its descendants, sometimes it matters whether the process is one of selection or drift, but at other times it does not. Maybe there is no such thing as "the" assumptions that cladistic parsimony makes in all inference problems.

In the previous chapter, I derived a Reichenbachian likelihood inequality concerning common cause and separate cause explanations. In the present chapter, I argued that the WTASB Markov model entails that synapomorphies and symplesiomorphy both discriminate between competing phylogenetic trees. These two problems, at first glance, look very different. The Reichenbachian problem involves two (or more) objects, and the question is whether they have a common cause. In the phylogenetic problem, the problem involves three (or more) objects that are *assumed* to have a common ancestor, and the question is which phylogenetic tree connects these objects to each other. Yet, the assumptions that entail the two results are strikingly similar. In both, common causes screen-off their effects from each other, the effect that changing an ancestor's character state has on one of its descendants is in the same direction as its effect on the other's, and all probabilities are intermediate.

Another link between Chapter 2 and this one concerns the issue of which types of observation provide stronger evidence and which provide evidence that is more modest. Just as Darwin's principle makes a claim about which similarities between two objects provides stronger evidence for the hypothesis that they have a common ancestor, so the SvS result describes when a synapomorphy provides stronger evidence for a tree than a symplesiomorphy does.[37] These two ideas are connected by the thought that traits that have lower probabilities of occurring in the present are often evidentially better than traits that have higher probabilities. If one state of a dichotomous character is adaptive and the other is maladaptive, the maladaptive state has the lower probability of occurring in the leaves of a tree; put that point side by side with the fact that people now have a lower probability of being named "Quackdoodle" than they have of being named "Smith." And in the plagiarism example, the fact that both essays are divided into paragraphs provides scant evidence of plagiarism, whereas the fact that the two essays misspell the same words in the same ways provides evidence that is more substantial. Paragraphed essays are common, but those curious misspellings are rare.

In the previous chapter, I espoused a "reductionistic" attitude towards Ockham's razor. Parsimony is never an epistemic end in itself; rather, it is sometimes a means to a more ultimate end. In this chapter I have examined a

[37] For further discussion of the question of how one chooses the characters that will serve as data in a phylogenetic inference problem, see Richards (2003) and Winther (2009).

particular formulation of Ockham's razor – the parsimony concept explored here is *cladistic* parsimony. Is reductionism the right view to take of this concept? Biologists tend to think so, whether they are friends or foes of cladistic parsimony. Farris (1983) defends cladistic parsimony by arguing that more parsimonious genealogies have greater explanatory power. Felsenstein (1978) criticizes cladistic parsimony by arguing that it can fail to be statistically consistent. And Felsenstein (1973) and Tuffley and Steel (1997) use the law of likelihood to identify circumstances in which parsimony is a legitimate inferential tool. Is it vulgar scientism to follow the scientists here? Scientists are people, not omniscient gods; they can be wrong, even about matters scientific. In this instance, I do not think the scientists are wrong.

Appendix 3.1: A sufficient condition for a 1-1-0 synapomorphy to favor (*XY*)*Z* over *X*(*YZ*), for a 0-0-1 symplesiomorphy to do the same, and for parsimony to be statistically consistent

In a balanced binary tree, there will be 2^n leaves if there are n generations between root and leaves. Three species (call them X,Y, and Z) are sampled at random from the leaves and you observe that X and Y are in state 1 while Z is in state 0. Here I'll prove

> *Proposition 1*: Pr [1-1-0 | $(XY)Z$] > Pr [1-1-0 | $X(YZ)$] if evolution is Markovian, 0 is the ancestral state, and the WTASB assumption holds.[38]

The WTASB ("within traits across simultaneous branches") assumption applies as follows to the tree in Figure 3.14. I divide the time between root and leaves into three stages. WTASB says that simultaneous lineages have the same probabilities of changing state; this leaves open whether the probabilities of change vary from one temporal stage to another. Let e_k = Pr(end in state 1 | start in state 0) for any branch in stage k ($k = 1, 2, 3$) and r_k = Pr(end in state 0 | start in state 1) for any branch in stage k. For mnemonic purposes you can think of "e" as representing evolution and "r" as representing reversal. The figure represents the specific case in which X and Y are i-related and Y and Z are j-related, where $i < j$. The hypothesis $(XY)Z$ is a disjunction over all i, j pairs where $i < j$.

[38] This simplifies (and sometimes corrects) the arguments given in Sober (1988, pp. 199–212).

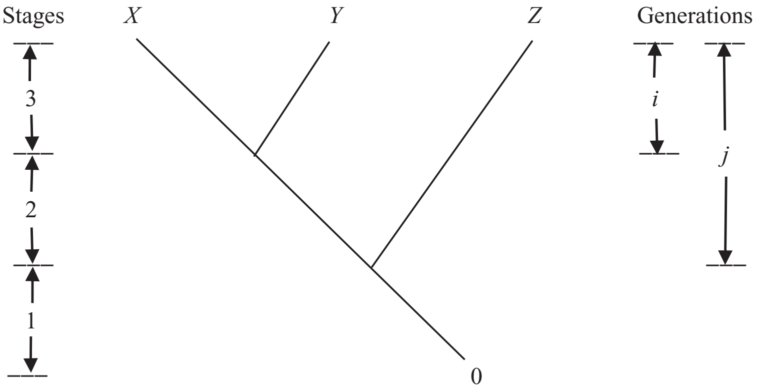


Figure 3.14

I want to show that

$$\Pr[\text{1-1-0} \mid R_i(X,Y) \text{ and } R_j(Y,Z)]$$
$$> \Pr[\text{1-1-0} \mid R_j(X,Y) \text{ and } R_i(Y,Z)], \text{ for each } i < j.$$

Given the WTASB assumption, this is equivalent to

(*) $\Pr[\text{1-1-0} \mid R_i(X,Y) \text{ and } R_j(Y,Z)] > \Pr[\text{1-0-1} \mid R_i(X,Y) \text{ and } R_j(Y,Z)]$, for each $i < j$.

This inequality can be expressed as follows:

$$\begin{aligned} & e_1[r_2(1-e_3)+(1-r_2)r_3][r_2e_3{}^2+(1-r_2)(1-r_3)^2] \\ & \quad +(1-e_1)[e_2r_3+(1-e_2)(1-e_3)][e_2(1-r_3)^2+(1-e_2)e_3{}^2] \\ & \quad > e_1[r_2e_3+(1-r_2)(1-r_3)][r_2e_3(1-e_3)+(1-r_2)r_3(1-r_3)] \\ & \quad +(1-e_1)[e_2(1-r_3)+(1-e_2)e_3][e_2(1-r_3)r_3+(1-e_2)e_3(1-e_3)], \end{aligned}$$

which simplifies to

$$(1-r_3-e_3)^2\,[e_1r_2\,(1-r_2)+(1-e_1)\,e_2\,(1-e_2)] > 0.$$

This last inequality is true as long as $(1 - r_3 - e_3) \neq 0$. The backwards inequality guarantees that $(1 - r_3 - e_3)$ is positive. So (*) is true. Proposition (1) follows, since $\Pr[R_i(X,Y) \,\&\, R_j(Y,Z)] = \Pr[R_j(X,Y) \,\&\, R_i(Y,Z)]$ for each i, j.

A similar result holds for symplesiomorphies:

Proposition 2: $\Pr[\text{0-0-1} \mid (XY)Z] > \Pr[\text{0-0-1} \mid X(YZ)]$ if evolution is Markovian, 0 is the ancestral state, and the WTASB assumption holds.

Merely replace e_3 with $(1 - e_3)$ and r_3 with $(1 - r_3)$ in the above argument.

The argument for Proposition 1 also provides a sufficient condition for parsimony to be statistically consistent:

> *Proposition 3*: Pr[1-1-0 | (*XY*)*Z*] > Pr[0-1-1 | (*XY*)*Z*], Pr[1-0-1 | (*XY*)*Z*] if evolution is Markovian, 0 is the ancestral state, and the WTASB assumption holds for each trait.

If (*XY*)*Z* is true, the probability approaches 1 that 1-1-0 synapomorphies will outnumber each of 0-1-1 and 1-0-1 as the number of characters sampled approaches infinity.

Appendix 3.2: When do synapomorphies provide stronger evidence for relatedness than symplesiomorphies?

Suppose three species *X*,*Y*, and *Z* are scored for two characters.[39] The first character says that *X*=1, *Y*=1, and *Z*=0; the second says that *X*=1, *Y*=0, and *Z*=0. In each case, we assume that the most recent common ancestor of the three leaves was in state 0. If each trait obeys the WTASB constraint (that a trait's rules of evolution on a branch are the same across all simultaneous branches), Propositions 1 and 2 in Appendix 3.1 entail that the first character favors (*XY*)*Z* over *X*(*YZ*) and that the second character favors *X*(*YZ*) over (*XY*)*Z*. Given this disagreement between the two characters, what do the two characters taken together say about the two topologies? This question is answered by the following theorem about synapomorphies versus symplesiomorphies:

(SvS) With WTASB and WBAT,

$$\frac{\Pr[X{=}1 \,\&\, Y{=}1 \,\&\, Z{=}0 \mid (XY)Z]}{\Pr[X{=}1 \,\&\, Y{=}1 \,\&\, Z{=}0 \mid X(YZ)]} > \frac{\Pr[X{=}1 \,\&\, Y{=}0 \,\&\, Z{=}0 \mid X(YZ)]}{\Pr[X{=}1 \,\&\, Y{=}0 \,\&\, Z{=}0 \mid (XY)Z]}$$

if and only if $p < q$.

Here p is the probability that each leaf has of being in state 1 and q is the probability that each has of being in state 0 ($p + q = 1$); notice that p and q are also the expected frequencies of the two traits across the leaves. WBAT means that on any branch, the two characters follow the same rules of evolution.

[39] My thanks to Mike Steel for helping me derive this proof, which improves on what I said about this problem in Sober (1988).

The (SvS) proposition is equivalent to the following:

With WTASB and WBAT,

$$\frac{\Pr[X{=}1 \,\&\, Y{=}1 \,\&\, Z{=}0 \mid (XY)Z]}{\Pr[X{=}1 \,\&\, Y{=}1 \,\&\, Z{=}0 \mid X(YZ)]} > \frac{\Pr[X{=}0 \,\&\, Y{=}0 \,\&\, Z{=}1 \mid (XY)Z]}{\Pr[X{=}0 \,\&\, Y{=}0 \,\&\, Z{=}1 \mid X(YZ)]} \text{ if and only if } p < q. \tag{1}$$

In the rest of this Appendix, I'll assume that WTASB and WBAT are true, so my numbered propositions will not repeat the four words that start proposition (1).

As noted before, the likelihood of a topology like (XY)Z is an average over all the specific degrees of relatedness that are compatible with that topology. I will show that (1) is true by demonstrating a stronger result:[40]

For each $i < j$,

$$\frac{\Pr\left[X{=}1 \,\&\, Y{=}1 \,\&\, Z{=}0 \mid R_i(X,Y) \,\&\, R_j(Y,Z)\right]}{\Pr\left[X{=}1 \,\&\, Y{=}1 \,\&\, Z{=}0 \mid R_j(X,Y) \,\&\, R_i(Y,Z)\right]} > \frac{\Pr\left[X{=}0 \,\&\, Y{=}0 \,\&\, Z{=}1 \mid R_i(X,Y) \,\&\, R_j(Y,Z)\right]}{\Pr\left[X{=}0 \,\&\, Y{=}0 \,\&\, Z{=}1 \mid R_j(X,Y) \,\&\, R_i(Y,Z)\right]} \text{ if and only if } p < q. \tag{2}$$

The WTASB and WBAT assumptions permit us to rewrite (2) as follows:

For each $i < j$,

$$\frac{\Pr\left[X{=}1 \,\&\, Y{=}1 \,\&\, Z{=}0 \mid R_i(X,Y) \,\&\, R_j(Y,Z)\right]}{\Pr\left[X{=}1 \,\&\, Y{=}0 \,\&\, Z{=}1 \mid R_i(X,Y) \,\&\, R_j(Y,Z)\right]} > \frac{\Pr\left[X{=}0 \,\&\, Y{=}0 \,\&\, Z{=}1 \mid R_i(X,Y) \,\&\, R_j(Y,Z)\right]}{\Pr\left[X{=}0 \,\&\, Y{=}1 \,\&\, Z{=}0 \mid R_i(X,Y) \,\&\, R_j(Y,Z)\right]} \text{ if and only if } p < q. \tag{3}$$

Notice that all four likelihoods in (3) conditionalize on the same hypothesis, which henceforth I will refer to as "*H*." We can rewrite proposition (3) as:

For each $i < j$,

$$\frac{\Pr[X{=}1 \,\&\, Y{=}1 \mid H]\,q}{\Pr[X{=}1 \,\&\, Y{=}0 \mid H]\,p} > \frac{\Pr[X{=}0 \,\&\, Y{=}0 \mid H]\,p}{\Pr[X{=}0 \,\&\, Y{=}1 \mid H]\,q} \text{ if and only if } p < q,$$

[40] Proposition (2) entails (1) if Simpson's paradox doesn't get a grip. It does not, thanks to the fact that the branching process I described earlier in this chapter involves the following equality among prior probabilities: for each i, j, $\Pr[R_i(X,Y) \,\&\, R_j(Y,Z)] = \Pr[R_j(X,Y) \,\&\, R_i(Y,Z)]$.

which simplifies to

$$\text{For each } i < j,\quad \frac{\Pr[X{=}1 \,\&\, Y{=}1 \mid H]}{p^2} > \frac{\Pr[X{=}0 \,\&\, Y{=}0 \mid H]}{q^2} \text{ if and only if } p < q. \tag{4}$$

Since $p = \Pr(X{=}1 \,\&\, Y{=}1 \mid H) + \Pr(X{=}1 \,\&\, Y{=}0 \mid H)$ and $q = \Pr(X{=}0 \,\&\, Y{=}0 \mid H) + \Pr(X{=}1 \,\&\, Y{=}0 \mid H)$, we can rewrite (4) as

$$\text{For each } i < j,\quad \frac{p - \Pr[X{=}1 \,\&\, Y{=}0 \mid H]}{p^2} > \frac{q - \Pr[X{=}1 \,\&\, Y{=}0 \mid H]}{q^2} \text{ if and only if } p < q. \tag{5}$$

This simplifies to

$$\text{For each } i < j,\quad pq\,(q - p) > \Pr[X{=}1 \,\&\, Y{=}0 \mid H](q^2 - p^2) \text{ if and only if } p < q. \tag{6}$$

This biconditional is obviously true if $p = q$, since then both sides of (6) are false. If $p < q$, the left-hand side of the biconditional in (6) is equivalent to

$$\text{For each } i < j,\ pq > \Pr[X{=}1 \,\&\, Y{=}0 \mid H]. \tag{7}$$

If $p > q$, the strict reverse of the left-hand side of the biconditional in (6) is equivalent to the condition:

$$\text{For each } i < j,\ pq\,(q - p) < \Pr[X{=}1 \,\&\, Y{=}0 \mid H](q^2 - p^2),$$

which can be rewritten as

$$\text{For each } i < j,\ pq\,(p - q) > \Pr[X{=}1 \,\&\, Y{=}0 \mid H](p^2 - q^2),$$

and this also simplifies to (7).

It remains to justify (7) for the case when $p \neq q$. Reichenbach's theorem from Chapter 2 does this for us. Notice that the most recent common ancestor of X and Y (i) has a non-zero probability of being in state 0 and a non-zero probability of being in state 1, (ii) screens-off X from Y, and (iii) has the same nonzero correlation with each of X and Y. It follows that $\Pr[X{=}1 \,\&\, Y{=}1 \mid H] > p^2$ and $\Pr[X{=}0 \,\&\, Y{=}0 \mid H] > q^2$, and these two inequalities entail that $\Pr[X{=}1 \,\&\, Y{=}0 \mid H] < pq$.

Appendix 3.3: What makes $\frac{\text{Pr(humans have } M \mid \text{chimpanzees have } M)}{\text{Pr(humans have } M \mid \text{chimpanzees lack } M)}$ a lot bigger than 1?

In discussing De Waal's use of cladistic parsimony to defend anthropomorphism, I noted that Reichenbachian assumptions suffice to ensure that

$$\frac{\Pr(\text{humans have } M \mid \text{chimpanzees have } M)}{\Pr(\text{humans have } M \mid \text{chimpanzees lack } M)} > 1.$$

I now will investigate what assumptions get this ratio to be a lot bigger than 1.[41]

First some notation. I will discuss three taxa, humans (*H*), chimpanzees (*C*), and their most recent common ancestor (*A*). I will write "*C*=1" to represent the proposition that chimpanzees have mental trait *M* and "*C*=0" to represent the proposition that they have *N*. That chimpanzees have either *M* or *N* is justified by the fact that they exhibit behavior *B* and the assumption that *M* and *N* are the only possible proximate mechanisms that can produce that behavior. Ditto for the other taxa and their possible character states. The above likelihood ratio can be expanded as follows:

$$\frac{\Pr(H{=}1 \mid C{=}1)}{\Pr(H{=}1 \mid C{=}0)} = \frac{\Pr(H{=}1 \mid A{=}1)\Pr(A{=}1 \mid C{=}1) + \Pr(H{=}1 \mid A{=}0)\Pr(A{=}0 \mid C{=}1)}{\Pr(H{=}1 \mid A{=}1)\Pr(A{=}1 \mid C{=}0) + \Pr(H{=}1 \mid A{=}0)\Pr(A{=}0 \mid C{=}0)}. \tag{1}$$

The right-hand side of this equality has the form

$$\frac{ax + b(1-x)}{ay + b(1-y)}. \tag{2}$$

Notice that the numerator and denominator in (2) each take a weighted average of *a* and *b*. As for *a* and *b* themselves, it is standard to assume that $a > b$; this is the backwards inequality discussed earlier in this chapter. If *a* isn't much bigger than *b*, then there is no way for the ratio displayed in (2) to be much greater than unity. For example, if $a = 0.7$ and $b = 0.6$, the largest value the ratio could have is $0.7/0.6 = 1.2$. It is interesting that this very basic consideration about the value of the ratio shown in (2) concerns human evolution alone; chimpanzees have nothing to do with it.

[41] Although there is no precise cut-off separating strong evidence from weak, Royall (1997) recommends using a likelihood ratio of at least 8 to define what strong evidence is.

The chimpanzee side becomes relevant if a and b are very different. For example, if $a = \Pr(H = 1 \mid A = 1) = 0.9$ and $b = \Pr(H = 1 \mid A = 0) = 0.01$, the values of x and y are worth considering. Ratio (2) can be rewritten as

$$\frac{x\,(a-b)+\ b}{y\,(a-b)+\ b}.$$

Since $a > b$, this ratio is made larger by increasing x and decreasing y. Bayes's theorem allows each of those terms to be expanded as follows:

$$x = \Pr(A = 1 \mid C\ = 1) = \frac{\Pr(C\ = 1 \mid A = 1)\Pr(A = 1)}{\Pr(C\ = 1)}$$

$$y = \Pr(A = 1 \mid C\ = 0) = \frac{\Pr(C\ = 0 \mid A = 1)\Pr(A = 1)}{\Pr(C\ = 0)}.$$

Given the common term $\Pr(A = 1)$, the way to make x large and y small is to make $\Pr(C = 1 \mid A = 1)$ large and $\Pr(C = 1)$ small.

4 Parsimony in psychology – chimpanzee mind-reading

Are chimpanzees *mind-readers*? That is, do they form mental representations of the mental states of others? Or are they just *behavior-readers*, forming mental representations of the behaviors of others?[1] I considered these questions in the previous chapter as a problem of phylogenetic inference. Given that the most recent common ancestor of chimpanzees and human beings existed about 6 million years ago, what can you conclude about whether chimpanzees are mind-readers from the fact that human beings are mind-readers? The answer I defended in the last chapter was: *only a little*. Modest Reichenbachian assumptions show that there is an evidential connection between us and them, but those modest assumptions do not show that the evidential connection is strong. Can psychological experiments on chimpanzees build a stronger case for the hypothesis that chimpanzees are mind-readers or for the hypothesis that they are not? And how, if at all, is Ockham's razor relevant to the interpretation of those experiments? My goal in this chapter is to answer these questions.

For experimental psychologists, the problem of chimpanzee mind-reading is a problem of *blackbox inference* (Sober 1998). You observe the environments that chimpanzees occupy and the behaviors they produce. Your task is to figure out what the psychological mechanisms are that mediate the connection between stimulus and response. These internal mechanisms are *intervening*

[1] Lurz (2011, pp. 25–26) notes that behavior-reading can involve sophisticated beliefs and inferences. For example, if chimpanzees predict the behavior of another individual by asking "What would I do if I were in such a situation?" and obtain an answer just by remembering their own environment/behavior pairings, this does not involve their having beliefs about the mental states of others. In what follows, I'll use the phrase "behavior reading" to cover all mental representations that don't involve mind-reading; beliefs about the color of a wall will count as behavior-reading.

variables. You don't observe internal processes directly; rather, you need to infer what they are like from what you do observe.

Corresponding to the two inference problems, phylogenetic and blackbox, there are two types of parsimony. In assessing hypotheses for their cladistic parsimony, you see which hypothesis requires the smallest number of evolutionary changes in character state to explain the data at hand. If human beings and chimpanzees both produce a given behavior, and human beings do this by mind-reading, then the most parsimonious hypothesis (in the sense of *cladistic* parsimony) is that chimpanzees do the same. The hypotheses to which cladistic parsimony applies get represented by phylogenetic trees. Blackbox parsimony doesn't address this evolutionary issue; it sets aside the genealogical relationship of human beings and chimpanzees and evaluates hypotheses about chimpanzee minds just on the basis of what chimpanzees do. Blackbox parsimony applies to this inference problem by evaluating hypotheses that are represented by flow charts that connect stimulus conditions to intervening variables, which in turn are connected to behavioral responses. Although cladistic parsimony sanctions the hypothesis that chimpanzees are mind-readers, it does not follow that blackbox parsimony will do the same.

Psychologists think that blackbox parsimony is relevant to the question of whether chimpanzees are mind-readers, but they disagree about what that relevance is. Tomasello and Call (2006, p. 371) argue that the mind-reading hypothesis provides a single unifying explanation that covers a range of very different experiments, whereas the behavior-reading hypothesis requires that each of these experiments be explained differently. If unifying theories are more parsimonious (an idea I discussed in Chapter 2), then Tomasello and Call's point about unification entails that mind-reading hypotheses are more parsimonious. Povinelli and Vonk (2004, p. 9) think this is backwards; they think that the mind-reading hypothesis cannot be more parsimonious, since mind-reading hypotheses postulate *more* mental representations, not *fewer*. How should the conflict between these two claims be resolved?

Besides evaluating these parsimony arguments, I'll address a wider philosophical issue. Both sides of this psychological debate think it is clear that chimpanzees have mental states. The debate concerns *which* mental representations chimpanzees have, not *whether* they have any at all. The days of methodological behaviorism – the thesis that one ought not to attribute inner mental states to explain behavior – are now long past. Even so, the philosophical question of what's wrong with behaviorism is worth considering; it is joined at the hip with the more specific questions that psychologists now discuss. If parsimony tells against attributing mind-reading to chimpanzees, why doesn't it also tell against attributing any beliefs at all to them? And if this is true when it comes to explaining chimpanzee behavior, why isn't it also true when it comes to explaining human behavior? It might be suggested that we human beings know by introspection that we have mental states, but this use of "we" is misplaced. The most that *I* know by introspection is that *I* have mental states. The question is what this tells me about my fellow human beings. Am I falling prey to naïve Soberomorphism if I assume that other humans are just like me? What is it about their behavior that justifies my attributing mental states to them? The problem about nonhuman organisms begins at home; its traditional name is *the problem of other minds*.

Experiments

Here's an experiment that *failed* to provide evidence for chimpanzee mind-reading.[2] Povinelli *et al.* (1990) gave their chimpanzee subjects a choice. Each chimpanzee faced two human beings; the human beings had food in front of them, but the food was not visible to the chimpanzee. The chimpanzee could either beg for food from an experimenter who could see the food or from an experimenter who was wearing a blindfold or had a bucket over his or her head. Begging from the former caused the experimenter to hand over the food, whereas begging from the latter did not. Experimenters who could see had blindfolds around their necks or carried buckets on their shoulders. Even after repeated trials, the chimpanzees begged just as frequently from experimenters who could not see as they did from experimenters who could.

Not much time passed before psychologists started to wonder whether this experiment was the right one to run. Hare *et al.* (2000) argued that the experiment lacked *ecological validity*. The name of the game for chimpanzees in the wild isn't to seek cooperation from a human experimenter. Rather, interactions are with other chimpanzees, and those interactions are more often competitive than cooperative when it comes to obtaining food.

[2] I am indebted to Fitzpatrick (2009) and Clatterbuck (forthcoming) for their analyses of these experiments.

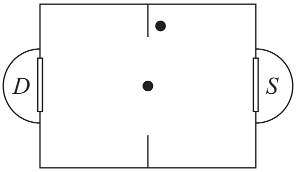


Figure 4.1

Hare *et al.* (2000) put this idea to work by designing experiments in which chimpanzee–chimpanzee interactions replace chimpanzee–human, and competition replaces cooperation.[3]

In their first study. Hare *et al.* (2000) placed a dominant and a subordinate chimpanzee in cages that were on opposite sides of a room containing two food items. One food item was visible to both chimpanzees while the other was placed behind an opaque barrier so that it was visible only to the subordinate, as shown in Figure 4.1. The doors of the cages were slightly ajar so that the chimpanzees could see each other and the layout of the room, but the doors were sufficiently closed that the chimpanzees were confined to their cages. The subordinate's door then opened, and the dominant's door opened a few seconds later. Out in the wild, dominant chimpanzees typically take all the food that is within their reach and punish subordinates who challenge them. Hare *et al.* predicted that if subordinates form beliefs about what dominants can and cannot see, then subordinates should scoop up the food behind the barrier and avoid the food that is out in the open. This is precisely what happened.

Hare *et al.* introduced variations on this experiment. In one, the dominant's door opened only after the subordinate had entered the room and started to go for one of the food items. Subordinates still preferentially targeted the hidden food, and this could not have been due to their merely reacting to the trajectory the dominant adopted upon entering the room. In another variation, the opaque barrier was replaced with a transparent one. In this

[3] Flombaum and Santos (2005) and Santos *et al.* (2006) followed up on Povinelli *et al.* (1990) by conducting experiments with rhesus monkeys. Whereas Povinelli *et al.* gave their chimpanzee subjects the opportunity to *beg* for food, Santos and colleagues gave their subjects the chance to *steal.* Given the choice between stealing food placed in front of a blindfolded experimenter and stealing food placed in front of an experimenter with unobstructed vision, the monkeys consistently chose to steal from the blindfolded experimenter.

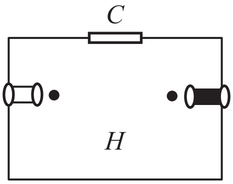


Figure 4.2

new setting, subordinates did not show a preference, thus undermining the hypothesis that subordinates merely prefer food behind barriers. In a third variation, the roles of subordinate and dominant were reversed, with the dominant seeing both food items and the subordinate seeing just one; in this experiment, the dominant first scooped up the food that was out in the open and then took the food that was concealed behind the barrier.

Hare *et al.* (2001) did a follow-up study in which the room contains two opaque barriers and the experimenter places a single food item on the subordinate's side of one of them. In one variation, both the subordinate and the dominant see the experimenter place the food; in another, only the subordinate sees the experimenter do this. Subordinates went for the food more often when the dominant did *not* observe the experimenter place the food. In a third variation, the dominant chimpanzee who had witnessed the placement of the food was replaced, just before the doors opened, by a dominant who had not. Subordinates went for the food more often when their competitors had not witnessed the food placement.

Melis *et al.* (2006) devised an experiment in which chimpanzees compete against a human experimenter. The experimenter is inside a booth that has a window on the front, but the walls of the booth are opaque; the chimpanzee is on the outside, as shown in Figure 4.2. The left and right sides of the booth have holes, which allow chimpanzees outside to reach into the booth. The holes open onto short tunnels, one of which is transparent while the other is opaque. Chimpanzees can see into the booth through the window; the experimenter inside the booth can see individuals outside only if they are in front. In a test condition, the experimenter places food items at the mouths of the two tunnels and then stares straight ahead. Chimpanzees who use the opaque tunnel succeed in getting the food, but when chimpanzees use the transparent tunnel, the experimenter snatches the food away before they can take it. In a control condition, the experimenter places food next to the mouths of the tunnels and then exits the booth. The upshot was that subjects use the opaque tunnel to take food in the test condition more often than they do in the control condition. Melis *et al.* (2006) did a variation on this experiment in which hearing replaces vision as the relevant sense modality. Chimpanzees now choose between using a silent trapdoor and a noisy trapdoor when they try to take the food. If a human competitor is inside the booth, chimpanzees succeed in getting the food if they use the silent trapdoor, but the competitor prevents them from succeeding if they use the noisy trapdoor. The chimpanzee subjects used the quiet trapdoor more often when a human competitor was in the booth than they did when no competitor was there.

Conflicting interpretations

The authors of the studies I just summarized think that their data strongly support the conclusion that non-human primates are, at least sometimes and in some respects, mind-readers. These psychologists are not saying that chimpanzees have all the mind-reading abilities that adult human beings have. On the face of it, the argument they make for this conclusion has nothing to do with parsimony; it seems a straightforward application of the law of likelihood. The authors seem to be saying that the outcomes of these experiments are what you'd expect if the mind-reading hypothesis were true, whereas the outcomes are improbable under the behavior-reading hypothesis.

Povinelli and Vonk (2004) strongly dissent from this conclusion. Not only do they think that these studies fail to provide any such evidence; they think that no similar study could do any better. Their view is that all such experiments are beset by a "logical problem":

> The general difficulty is that the design of these tests necessarily presupposes that the subjects notice, attend to, and/or represent, precisely those observable aspects of the other agent that are being experimentally manipulated. Once this is properly understood, however, it must be conceded that the subject's predictions about the other agent's future behavior could be made either on the basis of a single step from knowledge about the contingent relationships between the relevant invariant features of the agent and the agent's subsequent behavior, or on the basis of multiple steps from the invariant features, to the mental state, to the predicted behavior. (Povinelli and Vonk 2004, pp. 8–9)

The picture on offer here is based on the idea that mind-reading requires behavior-reading, but not conversely.[4] For example, if mind-reading occurs in the experiments that involve competition between two chimpanzees, there is a causal chain that goes from the behavior of a dominant chimpanzee (*D*) to the behavior of a subordinate chimpanzee (*S*) where that chain passes through two intermediate links:

(*M*) *D*'s behavior → *S*'s beliefs about *D*'s behavior →
S's beliefs about *D*'s mind → *S*'s behavior

If all we observe are *D*'s behavior and *S*'s behavior, do these observations discriminate between M and the following more austere causal chain?

(*B*) *D*'s behavior → *S*'s beliefs about *D*'s behavior → *S*'s behavior

The B chain is obtained from the M chain by snipping out a link. Povinelli and Vonk aren't saying that the data favor *B* over *M*; rather, they think the data are neutral with respect to *M* and *B*. The authors are using the razor of silence, not the razor of denial.[5] In posing their logical problem, Povinelli and Vonk are speaking in the voice of evidentialism; recall the discussion of Mill in Chapter 1.[6]

[4] Recall that in Chapter 1, I considered an interpretation of Morgan's canon according to which "higher" entails "lower," but not conversely.

[5] Penn *et al.* (2008) take a different stance; here the authors tend more towards the razor of denial.

[6] Povinelli and Vonk (2003, p. 160) say they are open to the possibility that a new type of experiment might be able to yield data that discriminate between a pure behavior-reading hypothesis and a hypothesis that postulates both behavior-reading and mind-reading. Their skepticism is limited to the kinds of experiments that I have described. Inspired by an idea from Heyes (1998), they suggest the following experiment:

> imagine that we let a chimpanzee interact with two buckets, one red, one blue. When the red one is placed over her head total darkness is experienced; when the blue one is similarly placed, she can still see. Now have her, for the first time, confront others (in this case the experimenters) with these buckets over their heads. If she selectively gestures to the person wearing the blue bucket we could be highly confident that the nature of her coding was, in part, mentalistic – that is, that she represented the other as 'seeing' her.

As Andrews (2005), Fitzpatrick (2009), and Lurz (2011) note, it is unclear how this experiment would escape the logical problem that Povinelli and Vonk (2004) formulate. A mind-reading explanation for this experiment would presumably postulate

Tomasello and Call respond to this criticism by conceding that the observations can be explained by a set of purely behavior-reading hypotheses, but that this explanation is inferior to a hypothesis that invokes mind-reading:

> The results of each experiment may be explained by postulating some behavioral rule that individuals have learned that does not involve an understanding of seeing. But the postulated rule must be different in each case, and most of these do not explain more than one experiment. This patchiness of coverage gives this kind of explanation a very *ad hoc* feeling, especially since there is rarely any concrete evidence that animals have had the requisite experiences to learn the behavioral rule – there is just a theoretical possibility. It is thus more plausible to hypothesize that apes really do know what others do and do not see in many circumstances. (Tomasello and Call 2006, p. 371)

Tomasello and Call do not mention parsimony in this passage, but they do assert that the mind-reading hypothesis provides a *unifying* explanation of the data from different experiments, whereas the behavior-reading hypothesis is disunifying. Given the connection between unification and parsimony discussed in Chapter 2, Tomasello and Call's point about unification entails that the mind-reading hypothesis is more parsimonious.

The idea that parsimony is on the side of mind-reading is something that Povinelli and Vonk (2004, p. 9) had earlier rejected. They say that Hare *et al.* (2000, 2001) "seem to imply that parsimony should push us toward assuming that [chimpanzees] . . . represent mental states." Povinelli and Vonk think this has things backwards:

> Reasoning about mental states . . . proceeds by observing behavior (in all its subtleties) and, on the basis of those noticed observable features, generating inferences about unobserved mental states. Thus, possession of a theory of

two causal chains:

E_1 is wearing a blue bucket → C believes that E_1 is wearing a blue bucket → C believes that E_1 can see → C gestures to E_1

E_2 is wearing a red bucket → C believes that E_2 is wearing a red bucket → C believes that E_2 cannot see → C does not gesture to E_2

Here E_1 and E_2 are experimenters and C is the chimpanzee. Just as one can snip away at the M chain to obtain B, so one can snip away at the two causal chains just described.

> mind does not somehow relieve the burden of representing the massive nuances of behavior or the statistical invariances that sort them into more and less related groups.
>
> There is no sense in which a system that makes inferences about behavioral concepts alone provides a *less* parsimonious account of behavior than a system that must make all of those same inferences *plus* generate inferences about mental states. (p. 11, their emphasis)

With respect to the *M* and *B* causal chains that I described, Povinelli and Vonk's point is that *M* can't be more parsimonious than *B*, since *B* is obtained from *M* by snipping (aka razoring). Penn and Povinelli (2007, p. 731) underscore this criticism of experimental arguments for mind-reading when they say that "comparative researchers have never specified a 'unique causal work' that representations about mental states do above and beyond the work that can be done by representations of the observable features of other agents' past and occurrent behaviors."[7]

A question that seemed to be cleanly answered by data turned out to be more complicated. Povinelli and Vonk's pessimism has been infectious; for example, Melis *et al.* (2006, p. 161) cite Povinelli and Vonk (2003, 2004) and say that "because a system that reasons about mental states uses information about behavior to generate inferences about the role of psychological states in producing behaviors . . . every mentalistic interpretation can be substituted by a behavioral account . . . The problem with this debate is that it is nearly impossible to resolve empirically." If observation is impotent, can parsimony be a tie breaker? If it can be, which hypothesis is more parsimonious?

Whiten's arrows

The suggestion that a mind-reading hypothesis can be more parsimonious than a (pure) behavior-reading hypothesis has a history. Drawing on ideas from Miller (1959) and Hinde (1970), Whiten (1996) describes "an earlier phase of animal psychology" in which researchers noticed that an observed set of

[7] It may seem that the behavior-reading hypothesis has a higher prior probability than the mind-reading hypothesis, simply because mind-reading entails behavior-reading but not conversely. This is a mistake. The behavior-reading hypothesis involves the *denial* of mind-reading and there is no general reason why a conjunction of the form *B&notM* should be more probable than *B&M*, a point I discussed in Chapters 1 and 2.

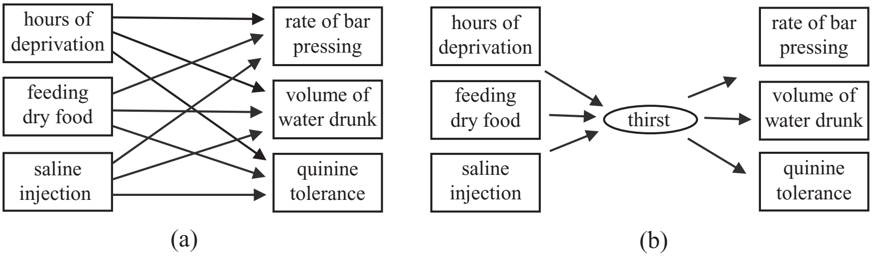
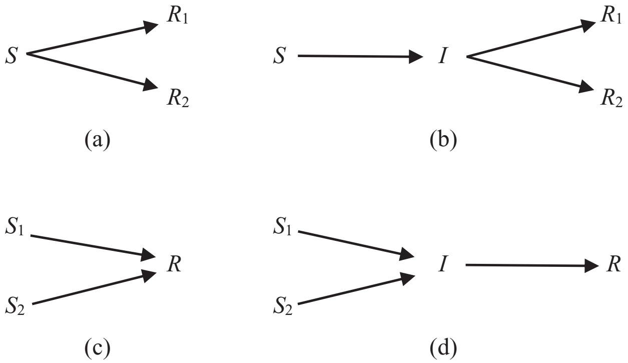


Figure 4.3

stimulus/response relationships can justify the postulation of an intervening variable. He begins with the example depicted in Figure 4.3. Experiments on rats reveal that each of the stimuli described in the left column of Figure 4.3(a) promotes each of the responses described on the right. Notice that there are nine causal arrows in this diagram. If an intervening variable is introduced, as shown in Figure 4.3(b), the number of arrows drops to six. Whiten (p. 284) says that this second figure is “more economic of representational resources.” He also notes that if there were more than three stimuli and three responses, the introduction of an intervening variable would provide an even greater economy. Whiten thinks this point generalizes beyond the example of thirst; he argues that the same logic applies to the introduction of a variable representing an organism’s beliefs about the desires and knowledge states of others. As in Figure 4.3(b), he uses “bottleneck” arrow diagrams in which multiple stimuli feed into a single intervening variable, which in turn issues in multiple behaviors (pp. 286–287).

Whiten is right that a model with an intervening variable has fewer arrows than a model without if the two have the structures shown in Figure 4.3. But consider Figure 4.4. Starting with 4.4(a), the introduction of an intervening variable between stimulus and response yields 4.4(b), which has more arrows, not fewer. The same point holds if you start with 4.4(c) and introduce an intervening variable so as to obtain 4.4(d). In these examples, should you prefer the model that declines to postulate an intervening variable? Or do these examples throw doubt on arrow counting as a way of deciding which models are better?

To answer these questions, we need to get clear on whether counting arrows is epistemically relevant. Whiten writes of “economy,” but what exactly does that mean? Saving the scientist’s ink and mental effort isn’t epistemically relevant. However, sometimes Whiten talks about arrangements



(a) (b) (c) (d)

Figure 4.4

that are economical *for chimpanzees* in the sense that minds with more arrow-saving intervening variables are better adapted than minds with fewer. This leads him to consider a "paradox" (p. 287): if intervening psychological variables that give the organism the ability to mind-read are so adaptive, why are there so few organisms that form representations about the mental states of others? One answer is that it would be adaptive for zebras to have machine guns to use in repelling lion attacks, but there is no paradox in the fact that zebras lack this adaptation. In any event, our subject here is not reducing the scientist's mental effort or identifying structures that would be adaptive for organisms if only they had them. Rather, we want to make inferences that are based on observed associations between stimulus and response. Why think that counting arrows is relevant to assessing models in the context of blackbox inference?

The two parsimony paradigms

In Chapter 2, I described two contexts in which parsimony is epistemically relevant. In the first, more parsimonious hypotheses have higher likelihoods; in the second, more parsimonious models have fewer adjustable parameters.[8] Do either of these ideas help explain whether introducing an intervening variable in the context of blackbox inference is a good idea?

[8] I described a third parsimony paradigm in Chapter 2, which arises when parsimony reflects non-first priors. Fitzpatrick's (2009) analysis of parsimony arguments concerning chimpanzee mind-reading falls in this category.

If arrows represent probabilistic parameters, then counting arrows might be a good idea in the context of model selection. I once tried to apply this idea to experiments on mind-reading (Sober 2009c), but I ran into a problem. The problem is that using model selection criteria like AIC requires that you obtain unique maximum likelihood estimates of all the parameters in each of the models considered. Unfortunately, the probabilistic parameters associated with arrows leading into and out of a postulated intervening variable, by their very nature, cannot be estimated in that way. For example, if a model says that $\pm C$ causes $\pm I$, which in turn causes $\pm E$, where $\pm C$ and $\pm E$ are observable but $\pm I$ is "hidden," you can't observe how often C events are followed by I events or how often I events are followed by E events. What you can observe is how often C events are followed by E events, and that observation permits you to construct a maximum likelihood estimate of $\Pr(E \mid C)$. However, that estimate doesn't give you unique estimates of the "component" probabilities. This is because $\Pr(E \mid C)$ is an average:

$$\Pr(E \mid C) = \Pr(E \mid I\&C)\Pr(I \mid C) + \Pr(E \mid notI\&C)\Pr(notI \mid C).$$

If $\pm I$ screens-off C from E in this causal chain, the equation simplifies to

$$\Pr(E \mid C) = \Pr(E \mid I)\Pr(I \mid C) + \Pr(E \mid notI)\Pr(notI \mid C),$$

but the problem remains. Models that contain adjustable parameters whose values can't be estimated are not *identifiable* (a technical term from statistics).[9] As noted in Chapter 2, AIC and related model selection criteria do not apply to them. Later in this chapter I'll return to the question of how model selection criteria can be applied to models that postulate intervening variables. Right now I am making a conditional claim: *if* Whiten's arrows represent adjustable parameters that need to be estimated to apply model selection criteria, then model selection criteria are inapplicable.

What about the other parsimony paradigm – the idea that parsimony sometimes mirrors likelihoods? Do Whiten's two models (Figure 4.3) confer

[9] Here's a simpler example. In the familiar *i.i.d.* (independent and identically distributed) model of coin tossing, there is a single adjustable parameter, p, which is the coin's probability of landing heads on a toss. You can estimate p by tossing the coin repeatedly. The observed frequency of heads is the maximum likelihood estimate of p. But suppose you decide that p is the sum of two other quantities so that $p = x + y$. Your frequency data do not permit you to estimate x or y, though of course they do allow you to estimate their sum.

different probabilities on the observed association of stimulus and response conditions? The answer is *yes*, but this answer has nothing to do with parsimony as measured by arrow counting. To see why, consider the point made in Chapter 2 that counting causes is sometimes a crude method for assessing parsimony. If smoking cigarettes and asbestos exposure cause lung cancer, it doesn't matter whether you describe this as two separate causes or as a single (composite) cause. If the variables are dichotomous, you can say that there are two dichotomous causes or a single four-state cause. Applying this lesson to Whiten's diagrams in Figure 4.3 shows that it shouldn't matter whether you talk about three causes (hours of deprivation, feeding dry food, saline injection) or a single composite cause that reflects the values of all three. In Whiten's diagram, a 3*S*-to-3*R* model competes with a 3*S*-to-1*I*-to-3*S* model; here "*S*" means stimulus, "*I*" means intervening variable, and "*R*" means response. The comparison of the two models should not be affected if you choose to describe the first as a 1*S*-to-3*R* model and the second as a 1*S*-to-1*I*-to-3*R* model. But notice that this recoding affects which model has fewer arrows. As mentioned, the 3*S*-to-3*R* formulation has nine arrows, and the 3*S*-to-1*I*-to-3*R* formulation has six. However, the 1*S*-to-3*R* formulation has three arrows, whereas the 1*S*-to-1*I*-to-3*R* formulation has four. The result of comparing models by counting arrows depends on how you diagram the models. However, the goal is to assess what the models *say*, which does not depend on how you say them. In just the same way, your assessment of competing models should not depend on whether you write them in English or Chinese.

Although arrow counting is problematic, the two Whiten models in Figure 4.3 *do* make different predictions if each is understood in a way that is standard in the literature on causal modeling (Spirtes *et al.* 2001; Pearl 2009). I propose to interpret model 4.3(a) as saying that the stimulus conditions together screen-off each response variable from each of the others, and I interpret his bottleneck model 4.3(b) as saying that the stimulus conditions together do not screen-off the response variables from each other.

A simpler example makes it easy to grasp this interpretation of Whiten's two arrow diagrams. Consider the (non-metaphorical) blackbox shown in Figure 4.5; it has a button (*B*) on one side and two lights (L_1 and L_2) on the other. You push the button once every ten seconds and watch what happens to the lights. You observe that if the button is pushed, each light goes on 80 percent of the time, but then you notice something more: when the button is

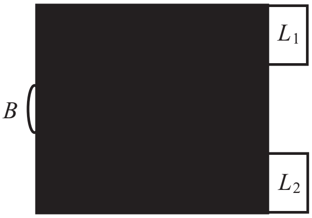


Figure 4.5

pushed, both lights are on 75 percent of the time. Notice that 75% > 80% × 80%. Given these observations, you consider two models about what is going on inside the box; each is named for the configuration of the wires that are said to connect the button to the lights. The *V* model says that there is a wire from *B* to the first light and an entirely distinct wire from *B* to the second light. The *Y* model says that there is a wire that runs from *B* to an intervening variable *I*, a wire that runs from *I* to L_1, and a third wire that runs from *I* to $L_{2.}$ For the purpose of my example, assume that a wire running from one point to another establishes a probabilistic connection, not a deterministic one, between the two points. The two models agree that *B* is a common cause of L_1 and L_2. However, the *Y* model postulates a more recent common cause (*I*); the *V* model does not. The *V* model says that the two lights are conditionally independent of each other. The *Y* model denies this. The observations I described favor the *Y* model.[10] What matters in this test of *V* against *Y* is not how often each light goes on when the button is pushed. It isn't the absolute values of freq(L_1 | *B* is pushed) or of freq(L_2 | *B* is pushed) that are telling. Even if the lights are almost always in the same state (on together or off together), that isn't the critical point, either. What matters is an inequality; we observe that freq(L_1&L_2 | B) > freq(L_1 | B) × freq(L_2 | B). This observed inequality has a higher probability under *Y* than it has under *V*.[11]

Is the *Y* model more parsimonious than the *V* model? Arguably not, since the *Y* model postulates an intervening variable while the *V* model does not. Similarly, *Y* says there are three pieces of wire inside the box, whereas *V* says

[10] I didn't say what the observations are when the button is *not* pushed. If you like, suppose that, when the button is not pushed, that each light goes on 10 percent of the time, but that the lights are on together 9 percent of the time. Again we have an inequality, since 9% > 10% × 10%.

[11] I will refine this likelihood assessment later.

there are only two. But from the point of view of the likelihood comparison, these facts about more and less do not matter. What matters is just the difference in likelihoods.

My denying that parsimony mirrors likelihood in the blackbox inference problem depicted in Figure 4.5 may sound strange. How can I be so negative, given that I advanced a Reichenbachian argument in Chapter 2 for thinking that parsimony mirrors likelihood? The answer is that the problem has changed. In Chapter 2, the competition was between common cause and separate cause explanations. In the present competition between V and Y, both models postulate a common cause, namely the state of the button. The question is whether an *additional* common cause should be postulated. If the data are as I described, likelihood is on the side of Y even though Y is intuitively less parsimonious. In this example, parsimony and likelihood clash. I say here what I said about other conflict situations in the previous chapter: *so much the worse for parsimony*.

Lessons from the blackbox

One reason the experiment on the blackbox in Figure 4.5 can provide evidence as to whether there is an intervening variable is that there is more than one observable effect. The fact that there happens to be a single observable cause doesn't matter; if there were two buttons on the left side of the box and each affects both lights, that would be fine. As noted earlier, two dichotomous causes are conceptually just like a single four-state cause, and the question will still be whether the two dichotomous causes (or equivalently, the single four-state cause) screen-off one effect from the other.

But having more than one effect is not enough. You need to run the right experiment. Suppose you did two separate experiments on the blackbox. In the first, you push the button repeatedly and record whether the top light (L_1) goes on. In the second, you do the same manipulation and record whether the bottom light (L_2) goes on. The results from the first experiment permit you to say what the frequency is of L_1's going on, and the results from the second allow you to say what the frequency is of L_2's going on, but what you cannot do is compute how often L_1 and L_2 *both* go on. What you should do is a *single* experiment in which each push of the button has *two* outcomes associated with it, one for L_1, the other for L_2. Now you have the data you need to test for screening-off. This experiment allows you to enter data (a simple *yes* or *no*) in each cell of the following table. The design of this experiment induces a pairing of observations of L_1 with observations of L_2. I will say that this experiment has *n* trials and two wings.

	t_1	t_2	...	t_n
L_1 is on?				
L_2 is on?				

These blackbox ideas suggest a new experiment that tests for chimpanzee mind-reading. The experiment has two wings and *n* trials. It allows you to test whether there is an intervening variable by determining whether two behaviors are positively associated. I'll describe a mind-reading hypothesis that predicts that the behaviors should be associated and a behavior-reading hypothesis that predicts that there should be no association. The raw materials for this new experiment can be found in two of the experiments that Melis *et al.* (2006) ran, even though the authors weren't trying to test for screening-off and the experiments they ran would not have permitted them to do so even if that had been their goal.

Melis *et al.* (2006), you'll recall, ran two experiments on the same chimpanzees; the two experiments were a year apart. In the first, there was a sequence of trials in which the human competitor was sometimes present and sometimes absent and chimpanzees had to choose between an opaque and a transparent tunnel. In the second, the human competitor was sometimes present and sometimes absent and chimpanzees had to choose between a silent and a noisy trapdoor. The main finding was that the chimpanzees had a higher probability of choosing the opaque tunnel when the human competitor was present than when the competitor was absent; ditto for the chimpanzees' probability of choosing the silent trapdoor. Notice that each of these results concerns what happened *within* a single experiment, not how the two experiments are related to each other. Melis *et al.*'s two experiments resemble the two sequential experiments I just described on the blackbox. In the first you get data about the first light; in the second you get data about the second. That experiment can't get at the question of whether the two lights are correlated, conditional on the button's being pushed. Similarly the two separate experiments from Melis *et al.* don't get at the question of whether the two chimpanzee behaviors – choosing the opaque tunnel and choosing the silent trapdoor – are correlated, conditional on the human competitor's being present.

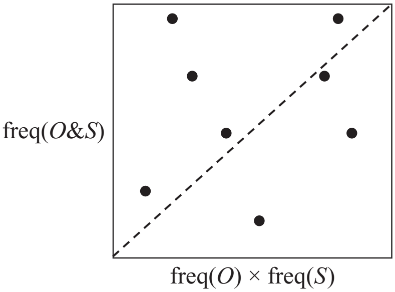


Figure 4.6

The two sequential experiments that Melis *et al.* ran need to be turned into a single experiment in which tunnel and trapdoor problems arise simultaneously. Here's the idea: each chimpanzee confronts a sequence of pairs of tasks. At time t_1 each chooses between the opaque and the transparent tunnel and immediately thereafter each chooses between the silent and the noisy trapdoor. Then, at time t_2, the same pair of tasks is presented again, with the order of the tunnel and trapdoor problems randomly varied. And so on, for n trials. Just as the two-wing n-trial experiment on the blackbox involves pushing the button each time, so the two-wing n-trial experiment on the chimpanzees has the human competitor present each time. The competitor, so to speak, is pushing the buttons of the chimpanzees. The results can be recorded by entering "yes" or "no" in each cell of the accompanying table.

	t_1	t_2	. . .	t_n
Chooses the opaque tunnel?				
Chooses the silent trapdoor?				

It may turn out that the performance of some chimpanzees is strongly at odds with screening-off, whereas the performance of others conforms to what that hypothesis asserts. Figure 4.6 depicts a hypothetical data set in which chimpanzees differ in how closely they conform to the predictions of the screening-off hypothesis. O means choosing the opaque tunnel; S means choosing the silent trapdoor. The dashed line represents what the screening-off hypothesis leads you to expect. The triangular space above the broken line is the region in which there is a positive correlation between O and S. Each dot represents the data from a single chimpanzee.

I now want to describe a mind-reading hypothesis (*MRH*) and a behavior-reading hypothesis (*BRH*) that disagree about screening-off. These are shown

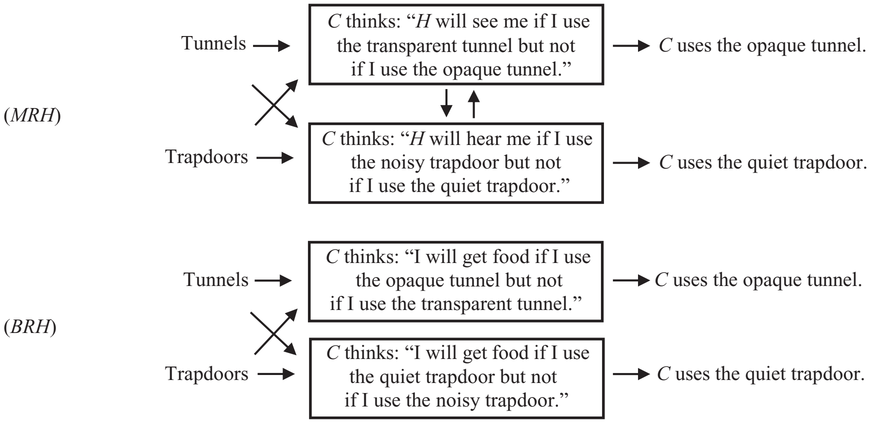

(MRH)
Tunnels
Trapdoors
C thinks: "H will see me if I use the transparent tunnel but not if I use the opaque tunnel."
C thinks: "H will hear me if I use the noisy trapdoor but not if I use the quiet trapdoor."
C uses the opaque tunnel.
C uses the quiet trapdoor.
(BRH)
Tunnels
Trapdoors
C thinks: "I will get food if I use the opaque tunnel but not if I use the transparent tunnel."
C thinks: "I will get food if I use the quiet trapdoor but not if I use the noisy trapdoor."
C uses the opaque tunnel.
C uses the quiet trapdoor.

Figure 4.7

in Figure 4.7. The items in boxes are the intervening variables postulated by *MRH* and *BRH*. The unboxed items represent observable stimuli and observable behaviors. The arrows represent causality. Causes raise the probabilities of their effects; they do not necessitate them. The diagonal arrows represent the possibility that experiencing the tunnel problem may influence what a chimpanzee believes when confronted with the trapdoor problem, and *vice versa*. These diagonal arrows could be erased without affecting my argument. It is the vertical arrows in Figure 4.7 that matter; they have the consequence that the two hypotheses disagree about something we can observe. *MRH* denies that the conjunction Tunnels&Trapdoors screens-off the two behaviors from each other; in particular, *MRH* entails that the two behaviors should be positively correlated. The *BRH* asserts that there is screening-off (= zero correlation, conditional on the stimuli).

Why are there vertical causal arrows in *MRH* but not in *BRH*? One of those arrows represents the following thought: if a mind-reading chimpanzee has a belief about what the human experimenter can and cannot see in the tunnel task that takes place at a given time, this should raise the probability that the chimpanzee will have a belief about what the human experimenter can and cannot hear in the trapdoor task that occurs immediately thereafter. The other vertical arrow represents the reciprocal possibility, that having a belief about what the human experimenter can hear in the trapdoor wing of the experiment should influence what the chimpanzee will believe about what the experimenter can see in the tunnel wing of the experiment that immediately follows. A mind-reader has the resources to apprehend these connections, which is not to say that the connections will always or even usually be drawn; all that is needed here is that the one state of believing would raise the probability of the other. In contrast, if a chimpanzee who is not a mind-reader forms the belief that he or she will get food by using the opaque tunnel rather than the transparent tunnel at a given time, this should not raise the probability of that individual's believing that he or she will get food by using the silent trapdoor rather than the noisy one immediately thereafter. A purely behavior-reading chimpanzee will lack the resources for drawing this connection. Figure 4.7 represents the idea that mind-reading and purely behavior-reading chimpanzees don't just differ in the contents of what they believe; they also differ in their abilities to apprehend a connection between the two wings of the experiment. The vertical arrows in *MRH*, and their absence in *BRH*, allow the two models to have different entailments about screening-off.

The phrases "mind-reading ability" and "behavior-reading ability" do not occur in the diagrams in Figure 4.7. The diagrams explicitly mention only stimuli, responses, and beliefs. Nonetheless, the abilities are there; they are in the arrows, not in the prose. Arrows represent causal relations; since these are probabilistic, they come in degrees. According to *BRH*, chimpanzees with great powers of behavior-reading will, with high probability, form the beliefs about how to get food that are described in the diagram. According to *MRH*, chimpanzees with great powers of mind-reading will probably figure out what the human competitor can see in the tunnel wing and what the competitor can hear in the trapdoor wing. But neither model commits to the idea that chimpanzees are great.

Notice that Povinelli and Vonk's (2004) point that mind-reading requires behavior-reading (but not conversely) does not upset the logic of the analysis I am proposing. If you like, take out your pencil and add some boxes to the *MRH* in Figure 4.7 that represent the behavior-reading beliefs that cause the mind-reading beliefs that are already in the diagram. Adding these boxes would not affect the point that *MRH* and *BRH* have different entailments about screening-off.[12]

It is a consequence of associating behavior-reading with a screening-off model and mind-reading with a model that denies screening-off that a

[12] Both *MRH* and *BRH* can be supplemented in other ways. This is fine, as long as the two models continue to disagree about screening-off.

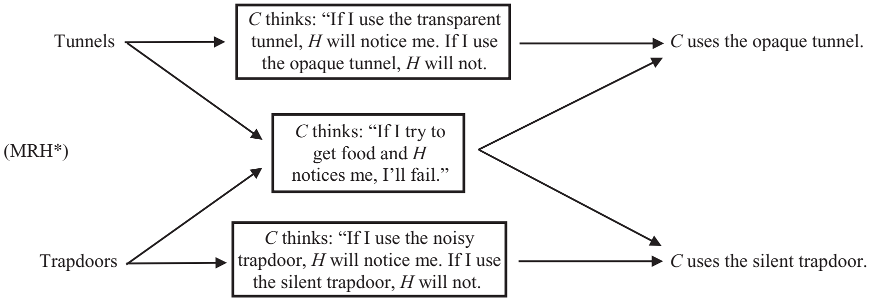

Tunnels
(MRH*)
Trapdoors
C thinks: "If I use the transparent tunnel, H will notice me. If I use the opaque tunnel, H will not.
C thinks: "If I try to get food and H notices me, I'll fail."
C thinks: "If I use the noisy trapdoor, H will notice me. If I use the silent trapdoor, H will not.
C uses the opaque tunnel.
C uses the silent trapdoor.

Figure 4.8

chimpanzee who gets perfect scores on both wings of the experiment provides no evidence for mind-reading.[13] This may seem odd, but I think it is correct. The mind-reading hypothesis says that chimpanzees draw a connection between the first and second wings. Getting perfect scores on each offers no evidence that a connection has been drawn. In just the same way, if I infallibly distinguish the singing of the Beatles from the singing of the Rolling Stones, and I also flawlessly distinguish the taste of red wine from the taste of gin, this provides no evidence that there is a common skill that I deploy in the two discrimination tasks.

A different mind-reading model that also entails no screening-off

The *MRH* described in Figure 4.7 postulates two intervening variables that causally interact, where each involves a representation that is about one experiment but not the other. There is another way to formulate the mind-reading hypothesis in which a postulated representation is about both experiments; an example is the *MRH** model shown in Figure 4.8. *MRH** uses the single concept of *noticing* whereas *MRH* uses two concepts, *seeing* and *hearing*. *MRH** says that a chimpanzee's choice in each wing of the experiment depends on the states of two variables that are logically independent of each other. Like *MRH*, *MRH** predicts that the stimulus conditions described in Tunnels&Trapdoors will fail to screen-off the two behaviors from each other. It is interesting that *MRH** resembles Whiten's bottleneck arrow diagram in

[13] As noted in the last chapter in the discussion of two species that match on each of eight dichotomous characters, there is no frequency association when there are two perfect scores, since 100% = 100% × 100%.

which thirst is postulated as an intervening variable (Figure 4.3). MRH* also shows how a mind-reading model can unify physically different experiments, an idea that Tomasello and his colleagues have emphasized.

Learning and screening-off

The mind-reading and behavior-reading models depicted in Figure 4.7 differ over whether the stimulus conditions screen-off responses in one wing of the experiment from responses in the other. It is important to see that the question of screening-off is distinct from the question of whether chimpanzees learn anything as they move through the experiment. The distinctness of these two questions can be seen in the accompanying table. There are four cells representing the four possible combinations of ±screening-off and ±learning. Each cell describes a triplet of values for Pr(*O*), Pr(*S*), and Pr(*O&S*) at the start of the experiment and a second triplet of values for those probabilities at the end. As before, *O* means choosing the opaque tunnel (rather than the transparent one) and *S* means choosing the silent trapdoor (rather than the noisy one). The numbers are hypothetical; they serve to illustrate the logical independence of learning and screening-off.[14]

		Learning			No learning		
		Pr(*S*)	Pr(*O*)	Pr(*S&O*)	Pr(*S*)	Pr(*O*)	Pr(*S&O*)
No screening-off	Finish	0.6	0.7	0.6	0.4	0.5	0.4
	Start	0.4	0.5	0.4	0.4	0.5	0.4
Screening-off	Finish	0.6	0.7	0.42	0.4	0.5	0.2
	Start	0.4	0.5	0.2	0.4	0.5	0.2

If screening-off and learning are logically independent, and if it is the former that matters to testing mind-reading against behavior-reading, this

[14] It is consistent with the logical independence of learning and screening-off that learning can induce a failure of screening-off in the following sense. Suppose in the first half of the experiment Pr(*O*) $= p_1$ and Pr(*S*) $= q_1$ and *O* and *S* are probabilistically independent of each other, while in the second half of the experiment Pr(*O*) $= p_2$ and Pr(*S*) $= q_2$ and *O* and *S* are again independent. If there is learning (so that $p_2 > p_1$ and $q_2 > q_1$), then there is a positive correlation between *O* and *S* in the entire experiment, though not in either of the two subparts. This is an instance of Simpson's paradox. A proper test for screening-off should take this complication into account.

has implications concerning another interesting experiment that Povinelli *et al.* (1990) carried out, this one on knowing and guessing. In that experiment, a chimpanzee looks through a window at two trainers in a room. One trainer (the "guesser") exits the room, while the other (the "knower") places food in one of four containers (the chimpanzee can't see which). The guesser then returns. The knower points to the container where the food is hidden, the guesser points to a different container, and the chimpanzee has to choose which of the two containers to examine. If the knower's container is chosen, the chimpanzee finds food and gets to eat it; if the guesser's container is selected, the chimpanzee fails to receive a food reward. After a number of rounds, three of the four chimpanzees in the experiment learned to choose the knower's container. The hypothesis under test is that chimpanzees form the belief that one trainer saw something that the other did not. The chimpanzees then went into a new experiment. Now the knower and the guesser both remain in the room while the food is hidden by a third party; the knower watches the food being hidden, while the guesser wears a bag over her head. The chimpanzees in this second experiment initially acted as if they were learning from scratch; initial success rates were around 50 percent. The chimpanzees then learned how to discriminate knower from guesser in this new set-up, and their mean success rate in the second experiment eventually reached the level they had attained in the first experiment. However, the chimpanzees got there faster; they took fewer trials in the second experiment than they needed in the first to attain the same degree of reliability with respect to choosing knower over guesser. Povinelli (1994) summarizes these findings by saying that there was no "immediate transfer" from the first experiment to the second, but there was "delayed transfer." He adds that immediate transfer, had that occurred, would have been evidence for mind-reading, but that the delayed transfer provides no evidence that the chimpanzees attributed mental states to the trainers. The experiment yielded evidence of learning. However, looking back on this experiment, Povinelli (1994) says that the kind of learning that took place does not favor mind-reading over behavior-reading.

I think that Povinelli's assessment is doubly mistaken. The experiment does not allow one to draw a conclusion about whether either type of transfer occurred. And even when the experiment is redesigned so as to furnish evidence about both types of transfer, it is a mistake to think that immediate transfer is evidence for mind-reading but that delayed transfer is not.

To assess whether experience of the first experiment influenced what the chimpanzees did in the second, a control group is needed. You need to know how well chimpanzees would do in the second experiment if they had no prior exposure to the first. Given a control group, you can tell whether experience of the first experiment caused immediate transfer, delayed transfer, or neither. Povinelli *et al.* (1990) did not have a control group, though perhaps we can take the performance of chimpanzees in the first experiment as providing a reasonable estimate of how chimpanzees would behave in the second experiment if they had no prior exposure to the first. This assumption leads to the conclusion that the delayed transfer in the second experiment was influenced by experience of the first, since the chimpanzees learned faster in the second experiment than they did in the first. In any event, it is hard to see why the plausibility of mind-reading depends on *when* transfer occurs; the ability to mind-read and the ability to catch on right away are distinct (Heyes 1994, p. 243).

If we want to focus on the question of screening-off, the data assembled by Povinelli *et al.* (1990) present the same difficulty as the data that come from Melis *et al.* (2006). In both, chimpanzees first go through a run of trials on one experiment and then make their way through a run of trials on another. Running one experiment after the other is not a good way to address the question of screening-off. What is needed is a revamping of the experimental design of Povinelli *et al.* that parallels the revamping already discussed for the two studies by Melis *et al.* Rather than having two experiments, one following the other, there needs to be a single experiment in which there are two alternating wings and *n* trials.

Associations, correlations, and testing

In my discussion of Reichenbach's ideas on common cause and separate cause explanations in Chapter 2, I emphasized the importance of distinguishing associations from correlations. The former is something you observe in your present data – for example, that $\text{freq}(O\&S) > \text{freq}(O) \times \text{freq}(S)$. This is distinct from a parallel claim about probabilities, that $\Pr(O\&S) > \Pr(O) \times \Pr(S)$. The probability statement is not merely a description of what you observe in your data; it is a claim about the underlying processes that generate today's data and so it may apply to tomorrow's data as well. We want to test claims about probabilities by using data on frequencies. This is why hypotheses about correlations need to be separated from observed facts about association.

In this light, let us consider more carefully how the behavior-reading hypothesis depicted in Figure 4.7 should be tested against the mind-reading hypothesis shown there. By extracting the entailments about screening-off from the full models, we have

(*BRH*) $\Pr(O\&S \mid \text{Tunnels\&Trapdoors}) = \Pr(O \mid \text{Tunnels\&Trapdoors}) \times \Pr(S \mid \text{Tunnels\&Trapdoors})$

(*MRH*) $\Pr(O\&S \mid \text{Tunnels\&Trapdoors}) > \Pr(O \mid \text{Tunnels\&Trapdoors}) \times \Pr(S \mid \text{Tunnels\&Trapdoors})$

These hypotheses can be rewritten in a format that should be familiar from the discussion of model selection in Chapter 2:

(*NULL*) $\Pr(O\&S \mid \text{Tunnels\&Trapdoors}) - \Pr(O \mid \text{Tunnels\&Trapdoors}) \times \Pr(S \mid \text{Tunnels\&Trapdoors}) = 0$

(*DIFF*) $\Pr(O\&S \mid \text{Tunnels\&Trapdoors}) - \Pr(O \mid \text{Tunnels\&Trapdoors}) \times \Pr(S \mid \text{Tunnels\&Trapdoors}) = p,$

where $p > 0$.

Bayesians and frequentists take different approaches to assessing these competing models. Bayesians, using the law of likelihood, need to figure out what the probability is of the observed association under each hypothesis. This, it turns out, depends on the different values that the three probabilities that are mentioned in each hypothesis might have.[15] Pr(data | NULL) is therefore an average and the same is true of Pr(data | DIFF). Frequentists don't consider these averages. A standard frequentist procedure for this problem is to do a chi-squared test of independence. In that test, you focus on the NULL hypothesis and ask whether an association at least as strong as the one you observe would be very improbable if NULL were true. If it is improbable enough, you reject the NULL hypothesis; if it is not, you decline to reject. But how improbable is improbable enough? The answer to this question (the choice of a "level of significance") is widely acknowledged to be a matter of convention. Notice that the chi-squared test does not consider what DIFF

[15] The probability density of an observed degree of association under the *NULL* hypothesis of screening-off depends on the values of Pr(*S*) and Pr(*O*). I'm grateful to Casey Helgeson for pointing this out to me.

predicts about the experimental outcome. DIFF sits passively on the sidelines, waiting to see whether NULL fails.[16]

Parsimony redux

Early in this chapter, I said that if the arrows that run into and out of intervening variables represent probabilistic parameters that need to be estimated for model selection criteria to be applied, then model selection criteria cannot be used to compare pure behavior-reading hypotheses with hypotheses that postulate mind-reading. The subsequent discussion has thrown doubt on the *if* in this statement. In that discussion, I focused just on the screening-off claim made by the behavior-reading hypothesis and the no-screening-off claim made by the mind-reading hypothesis; these two claims concern the relationship of observable stimulus conditions to observable behaviors.

The problem that needs to be addressed now is that NULL, as formulated above, has zero adjustable parameters, but it doesn't by itself say how probable the observed association is. So let us reformulate *NULL* and its competitor as follows:

(*NULL**) Pr(O | Tunnels&Trapdoors) $= a$, Pr(S | Tunnels&Trapdoors) $= b$, and Pr(O&S | Tunnels&Trapdoors) $= ab$.

(*DIFF**) Pr(O | Tunnels&Trapdoors) $= a$, Pr(S | Tunnels&Trapdoors) $= b$, and Pr(O&S | Tunnels&Trapdoors) $= ab + c$ (where $c > 0$).

*NULL** has two adjustable parameters, *DIFF** has three, and the data consist of three frequencies. *DIFF** will be able to fit those frequencies perfectly; *NULL** will almost certainly do worse. To apply AIC and similar model selection

[16] For a useful tutorial on how the chi-squared test of independence works, see http://stattrek.com/chi-square-test/independence.aspx. Here's hypothetical data (from a single chimpanzee) on 100 trials of the 2-winged experiment arranged in a 2×2 contingency table:

		Trapdoor	
		Silent	Noisy
Tunnel	Opaque	40	11
	Transparent	10	39

In this example, the upper-left and lower-right numbers are bigger than you'd expect under the null hypothesis of independence. For example, the expected value of upper-left under the null hypothesis is $\frac{51\times 50}{100} = 25.5$.

criteria, you need to ask whether *DIFF** fits the data *sufficiently* better than *NULL** does to justify the additional parameter.

The mind-reading model *is* more complex, as Povinelli and Vonk (2004) insist, though not because mind-reading requires behavior-reading; rather, the reason is that the mind-reading model has more adjustable parameters than the behavior reading model, at least when the focus is on what each says about screening-off. Yet, it also is true that the mind-reading hypothesis is unifying when it postulates an internal representation that arises in each of several physically different experiments. This point, emphasized by Tomasello and Call (2006), is especially vivid in connection with *MRH** in Figure 4.8. So we here have a case in which unification and parsimony are at odds with each other. This contrasts with the example of the two fields of corn discussed in Chapter 2; there unification and parsimony go hand-in-hand.

What are we to say about such cases of conflict? If parsimony is nice and so is unification, which trumps the other when the two conflict? It is at this point that we must invoke a more fundamental epistemological consideration. Parsimony is not an end in itself and neither is unification. A model selection framework tells us that NULL* is more parsimonious than DIFF* and explains why that difference is epistemically relevant. If the unifying representation postulated by a mind-reading hypothesis entails a violation of screening-off, whereas the disunifying representations postulated by a behavior-reading hypothesis entail that there is screening-off, then parsimony is on the side of behavior-reading. At this point, the data must be permitted to speak. It remains to be seen whether the two-winged experiment I have described will favor mind-reading over behavior-reading.

A cross-chimpanzee comparison that is not about screening-off

The two-winged experiment I have described can be carried out on a single chimpanzee or on several. When it involves several subjects, one can determine whether all, some, or none of them deviate from the prediction that the behavior-reading screening-off model makes (Figure 4.6). However, there is another way to look at data coming from several chimpanzees that does not involve the screening-off question; you can see if chimpanzees who do well on one wing of this experiment also tend to do well on the other. Each chimpanzee obtains two “scores” in the two-winged experiment – one is the

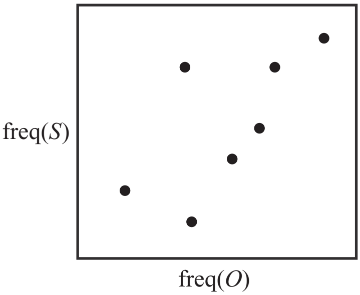


Figure 4.9

frequency of choosing the opaque tunnel (O) rather than the transparent one; the other is the frequency of choosing the silent trapdoor (S) rather than the noisy one. Figure 4.9 provides a hypothetical data set in which each chimpanzee's two scores are represented by a single point. The question is whether the best-fitting line for these data has a positive slope.

Suppose the slope is not just positive, but sufficiently positive for you to feel confident that the O and S scores are positively correlated. What does this result tell you about mind-reading versus behavior-reading? The mind-reading hypothesis says that each chimpanzee has some non-zero ability to mind-read and that this unitary ability is used in both wings of the experiment. The hypothesis does not say that all chimpanzees have the same degree of acuity. Understood in this way, the mind-reading hypothesis predicts that chimpanzees who do well on one wing of the test will tend to do well on the other, and that those who do poorly on one will tend to do poorly on the other. The hypothetical data shown in Figure 4.9 favor this hypothesis over one that says that there are two unrelated abilities that work separately and independently in the two wings of the experiment. However, it isn't clear that these data favor the unitary mind-reading hypothesis over a *unitary* behavior-reading hypothesis, where the latter says that chimpanzees have a unitary ability to behavior-read; this hypothesis says that chimpanzees who are good at behavior-reading in one wing of the experiment will also be good at behavior-reading in the other. Behavior-reading can be a unitary mechanism just as mind-reading can be. If so, this experimental outcome does not favor mind-reading over behavior-reading. Screening-off gets at a question that this cross-subject correlational analysis does not.

As it happens, the hypothetical data depicted in Figure 4.9 can be replaced by real data from the two separate experiments that Melis *et al.* (2006) ran.

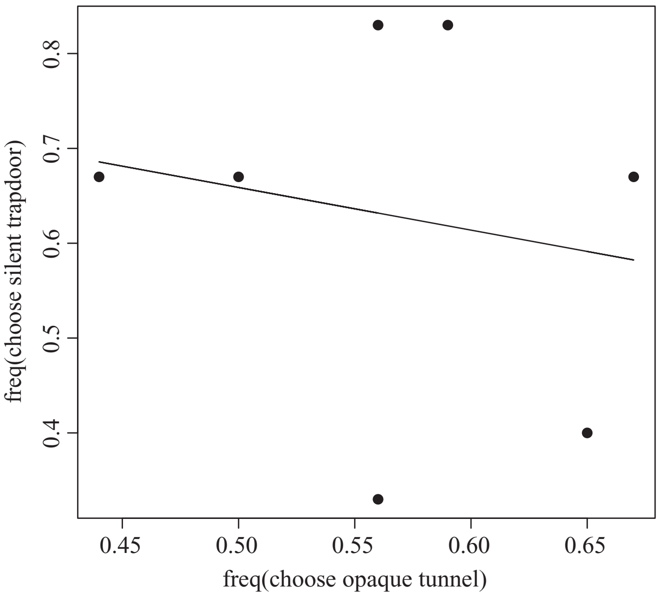


Figure 4.10

These data are shown in Figure 4.10. The best-fitting straight line has a slightly negative slope; chimpanzees who did better than average on the tunnel test tended to do worse than average on the trapdoor test, but only by a little.[17] I mentioned that these two experiments were a year apart. It isn't clear whether the same pattern would emerge if the two experiments were interleaved to form a single experiment with two wings.

Behaviorism versus mentalism

At the start of this chapter, I pointed out that Povinelli and Vonk's (2004) logical problem can be posed about the postulation of any belief state at all. Their assertion that a mind-reading hypothesis and a behavior-reading hypothesis make the same predictions generalizes. Just as a model that postulates a

[17] The deviation of the best-fitting regression line from a slope of zero is not statistically significant, nor is there much of a difference between the AIC scores of the two-parameter linear model (where the slope and the y-intercept are both adjustable parameters) and a one-parameter linear model (where the slope is set at zero and the y-intercept is the sole adjustable parameter). I am grateful to Brian McLoone for these two findings.

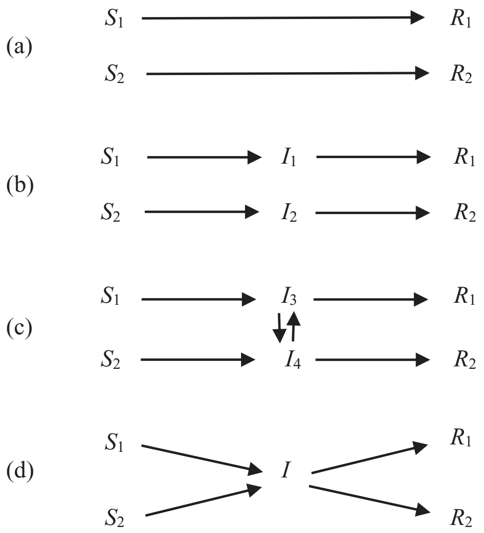


Figure 4.11

causal chain from stimulus to response that passes through both behavior-reading and mind-reading beliefs can be shortened by snipping away the mind-reading, so this shortened model can be shortened even more, by snipping away the behavior-reading. The result is a model that conforms to the dictates of methodological behaviorism; it postulates a causal chain from stimulus to response that fails to invoke any intervening variable at all. This austere hypothesis is silent about the existence of intervening variables; it does not deny that they exist. Those convinced of the folly of methodological behaviorism may see this as a *reductio* of Povinelli and Vonk's argument. And yet their argument is right about something: the observed association of stimulus and response (the start and the finish of all three of these causal chains) does not tell you that any of these chain models is better than any of the others.

The solution to this puzzle is to stop focusing on causal chains and to start considering stimulus conditions that have multiple effects. Figure 4.11 depicts four models of the relation of two stimulus variables S_1 and S_2 to two response variables R_1 and R_2. The first model, 4.11(a), avoids mention of intervening variables altogether, whereas the second, 4.11(b), introduces one on each chain. These two models are statistically indistinguishable if your data consist of observed associations between stimuli and responses.

However, each is distinguishable from the third, 4.11(c) and from the fourth, 4.11(d); these last two both entail no-screening-off.[18]

A now-standard argument against behaviorism and for mentalism traces back to Chomsky's (1959; see especially p. 31) landmark review of Skinner's book *Verbal Behavior.* The argument appeals to "novel" behaviors and is given a pithy presentation by Dennett (1981). Dennett (pp. 65–67) says that the behavior of a chicken that has been conditioned to turn around when a light goes on can be adequately explained without postulating internal mental states, but that matters change when you consider Skinner's example of a man who is robbed for the first time in his life. In Dennett's first-person telling of this story, a robber brandishes a gun and says "your money or your life." Dennett gingerly hands over his wallet. He has never been robbed before, and when he has been threatened, he has tended to apologize, not to give away money. Dennett concludes that a behaviorist explanation is impossible here and that there is every reason to think that a mentalistic explanation is the way to go. I do not dispute the conclusion; it is the argument that I think needs to be scrutinized. When Dennett says that the stimulus and the response in this example are "novel," what does this mean? After all, every event is novel in some respects and quotidian in others. The answer is that the robbery can be said to be "similar" to past events in Dennett's lifetime only if you use mentalistic descriptions. However, this is off-limits for behaviorists; they can't explain Dennett's behavior by saying that Dennett had been conditioned to do what he thinks people want him to do when he believes they have threatened him. On the other hand, if we use a non-mentalistic descriptor, the robbery really *is* novel, and so it can't be explained by saying that it is an instance of a kind of behavior that Dennett has been conditioned to perform in a kind of circumstance that has occurred in the past.

I agree with Dennett's argument, but which part of behaviorism does it undermine? Behaviorism's negative thesis is that one should not postulate intervening variables to explain stimulus/response connections. Behaviorism's positive thesis is that a present behavior should be explained non-mentalistically by showing that the same kind of behavior was conditioned in the same kind of circumstance in the past. The point about novel

[18] A fifth arrow diagram can be obtained from Figure 4.11(d) by adding arrows that go from S_1 to R_1 and from S_2 to R_2. This also entails no screening-off; it corresponds to the *MRH** in Figure 4.8.

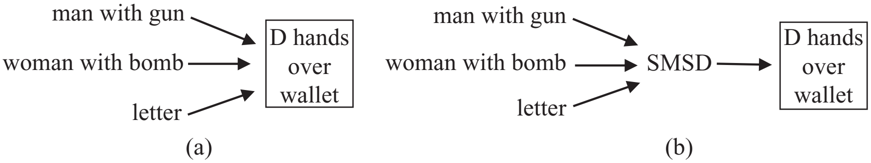


Figure 4.12

behaviors undermines the positive thesis, but how does it impugn the negative one? Dennett rightly criticizes Skinner's wholesale refusal to postulate unobserved internal mechanisms, but an openness to the legitimacy of such postulates leaves open what sorts of observations would favor them.

Earlier in the paper, Dennett (1981, pp. 66–67) says that the problem cases for behaviorism are behaviors that exhibit novelty and generality. Let us consider the latter property. There are lots of stimulus conditions that could make Dennett hand over his wallet. Besides the man with the gun, Dennett mentions a woman with a bomb, and we might add the receipt of a letter containing a certain sequence of words. There are numerous other stimuli that would suffice as well. Dennett asks a few pages later (p. 70) what these different stimuli have in common that allows them all to cause him to hand over his wallet. He says that the behaviorist cannot answer this question, but the mentalist can. The stimuli have in common the fact that Dennett will believe that each of them poses a threat with which he ought to comply. Dennett thinks that this general mentalistic explanation is better than the piecemeal behaviorist argument that treats each stimulus as a law unto itself.[19] I hope the reader senses that we are now in the neighborhood of Whiten's arrow diagrams and of Tomasello and Call's (2006) preference for explanations that are unified rather than patchy. Dennett's generality argument is on to something, but we have already seen that the way forward is not to flesh out the argument in the way represented in Figure 4.12. Perhaps we feel sure that the explanation represented in 4.12(b) is right – that an SMSD (a single mental state of Dennett's) is what knits these stimuli together, allowing each to lead to the same behavior. The problem is that the observed association of each stimulus with the response does not favor 4.12(b) over

[19] Dennett's point about explanation resembles an argument against reductionism due to Putnam (1967, 1975) that says that explanations are better when they are more general; I discuss this argument in the next chapter.

4.12(a). If the observations are to do that epistemic work, we need multiple effects.[20]

Maybe we should assume the following as an *a priori* principle about causation: if *C* causes *E*, where the two are spatially and temporally separated, then there is an intervening variable *I* such that *C* causes *I* and *I* causes *E*. Since pushing the button on the left side of the blackbox in Figure 4.5 causes the lights on the right to go on, we naturally assume that there is something going on in between. Notice that the attractiveness of this assumption does not depend on whether the lights are correlated or on whether we embrace the *Y* model or the *V*. This principle of "no action at a distance" does not entail that the radical behaviorist model 4.11(a) is false, but just that it is incomplete if it is true. Data become relevant if we wish to compare either 4.11(a) or 4.11(b) with either 4.11(c) or 4.11(d). Behaviorism competes with some intervening variable models but not with others.

Suppose you are prepared to assume the following conditional, for some specific stimulus (*S*) and response (*R*): if *S* causes *R*, then there is an intervening mental state of kind *M* that connects the two. Given this assumption, your observing that an instance of *S* causes an instance of *R* will lead you to conclude that an instance of *M* is present. Perhaps this is how you think about Dennett's handing over his wallet when the robber threatens him. But suppose you want to address the prior question of whether the assumed conditional is correct. In this situation, observing that an instance of *S* causes an instance of *R* is useless. This is the dialectical situation in which the debate between behaviorism and mentalism occurs; it is also the situation in which mind-reading and behavior-reading are competing hypotheses. In these circumstances, we must look beyond the simple case of a single *S* and a single *R*. The thing to look at is whether there is screening-off when a stimulus causes multiple responses.

[20] The model in Figure 4.12(b) entails that anything that causes SMSD will cause Dennett to hand over his wallet, but the 4.12(a) model does not have this implication. Does this allow observations to provide evidence that 4.12(b) is true and that 4.12(a) is false? *No*, and for two reasons. First the discovery of new causes of Dennett's handing over his wallet does not disconfirm 4.12(a); it just shows that the model is incomplete, not that it is false. Second, what we can observe is that there are numerous ways to get Dennett to hand over his wallet. This, by itself, isn't evidence that they all pass through the same intervening variable. This point is relevant to Whiten's (2013) use of novel stimuli to test for mind-reading, on which see Heyes (2015).

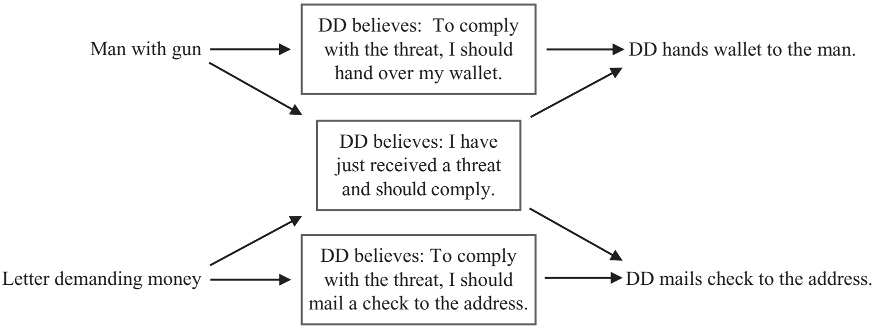

Man with gun
Letter demanding money
DD believes: To comply with the threat, I should hand over my wallet.
DD believes: I have just received a threat and should comply.
DD believes: To comply with the threat, I should mail a check to the address.
DD hands wallet to the man.
DD mails check to the address.

Figure 4.13

It isn't difficult to fit the robbery example into this framework of multiple causes and multiple effects. When the robber flashes his gun and says "your money or your life," Dennett hands over his wallet, but when Dennett receives a threatening note demanding that a check be sent to a certain address, he writes one and puts it in the mail. The stimuli are physically different and so are the two responses. Figure 4.13 provides a mentalistic model that unifies the two responses; it predicts that the two stimuli will fail to screen-off the two responses from each other.[21] I hope the similarity of Figures 4.13 and 4.8 is patent.

This analysis is not yet complete. We need to compare this no-screening-off model with a non-mentalistic model that entails screening-off. The non-mentalistic model can grant that there are intervening variables between stimulus and response; what is key is that it says that the stimuli screen-off the behaviors from each other. And now we need to take one final step – we need to describe a single experiment in which subjects experience both stimuli. A two-winged *n*-trial format will serve.

Concluding comments

In this chapter, I used two experiments from Melis *et al.* (2006) to design a single experiment. I then considered two specific models of what is going on in this experiment – one postulating pure behavior-reading, the other

[21] This strategy for thinking about the postulation of intervening variables applies to psychological explanations of behavior that don't involve propositional attitudes. For an example, see my discussion of Wickens (1938) in Sober (1998).

postulating mind-reading (Figure 4.7). Neither the simple idea that chimpanzees are mind-reading in this experiment nor the simple idea that they are not is concrete enough to make testable predictions. Rather, each has to be fleshed out with specifics. Recall the point made in Chapter 2 about the futility of trying to test "catchall hypotheses." There is no helping the fact that it is specific hypotheses, not catchalls, that are tractable.

The example of the blackbox in Figure 4.5 should make it plain that even if a screening-off test provides evidence for the existence of an intervening variable, this, by itself, does not show that the intervening variable should be characterized mentalistically. And even if it should be, the failure of screening-off does not tell you what representational content that intervening variable has.[22] So why all the fuss about screening-off? I argued, in the discussion of behaviorism versus mentalism, that if you have no evidence for an intervening variable, it is pointless to ask whether the variable in question has this or that mental characteristic. This is why I think the screening-off test is important.

In order to use the screening-off test on the question of mind-reading, I had to construct mind-reading and behavior-reading hypotheses that have different implications, not just about the contents of the beliefs that chimpanzees have, but also about the causal connections between the postulated belief states. This is why, in Figure 4.7, the two models differ over whether there are vertical arrows. The question then arises of why a behavior-reading hypothesis should be committed to there being no vertical arrows. Friends of behavior-reading may contest the reason I gave and insist on inserting a vertical arrow. But what will they say if the experiment I describe yields data that strongly supports screening-off? Will they regard this as evidence against their view that chimpanzees are pure behavior-readers? I suspect that the answer is *no*. Does this mean that friends of behavior-reading think that their position is compatible both with having vertical arrows and with not having them? This open-mindedness is fine, except that it means that we have no model that we can test. It is essential that *both* models have something concrete to say about which belief states cause which others. It would be helpful if a theory

[22] Here we have an analog of a distinction drawn in Chapter 3: it is one thing to infer a phylogenetic tree and something else to infer what the character states are of the ancestors in that tree. Inferring common cause *variables* is one thing; inferring the *states* of those variables is another.

about this, agreeable to both parties, were made explicit.[23] My understanding is that behavior-readers can conceptualize two events as similar only if they share a perceivable property. The tunnel set-up and the trapdoor set-up "look different." If there are no confounds, they really are perceptually different – not just to us, but to chimpanzees. Of course, an experiment can always inadvertently introduce a confound. For example, it would be a mistake to always have the opaque tunnel on the left side of the box and always have the silent trapdoor on the left as well. The two-winged experiment I have described may need to be fine-tuned to avoid this kind of flaw.

I am well aware that the two-winged experiment can fail to provide evidence that favors the mind-reading hypothesis even if chimpanzees do have the ability to mind-read. This will happen if the two alternating tasks are too difficult. If they are too difficult, the chimpanzees will guess at random, and there will be no association between right answers on one wing and right answers on the other. The same problem can arise if the two tasks are too easy; the chimpanzees will then each achieve near-perfect scores on each and again there will be no violation of screening-off. It is middle-range problems – problems that are neither too hard nor too difficult – that have the best shot at producing failures of screening-off and thus providing evidence for mind-reading.[24]

Blackbox inference provides a couple of lessons concerning the principle of parsimony. Adding an intervening variable to a model may reduce or increase the total number of arrows in a diagram, but that depends on how the variables are represented, not on what the model in fact says. Features of a representation must not be confused with features of the propositions that the representation expresses. It is the semantic features of a hypothesis, not its syntactic features, that are epistemically relevant. In Whiten's

[23] Heyes (2015) makes the important point that behavior-reading hypotheses should not be concocted out of thin air or be based on an uncritical use of folk psychology; they should have some antecedent and independent plausibility in cognitive science.

[24] There is an analogy here with phylogenetic inference. If the probabilities of change on branches of the candidate trees you consider are too small (= 0), then all trees have the same likelihood (namely 0) if there are both apomorphic and plesiomorphic traits in leaf taxa; if the branch transition probabilities are too high (e.g., around ½ in a two-state drift model), the likelihoods are again virtually indistinguishable. It is "middle-range" values that are best.

arrow diagrams, the addition of an intervening variable changes the model from affirming that causes screen-off effects from each other to denying that this screening-off relation obtains. This point has nothing to do with arrow counting, and it opens an important avenue for bringing data to bear on the question of when intervening variables ought to be introduced. The claim of screening-off is a null hypothesis and so it is more parsimonious than the claim that there is no screening-off, since the latter claim involves more adjustable parameters.[25] Parsimony is relevant to comparing these two, but parsimony is on the side of affirming screening-off, not denying it. It is interesting that the problem of comparing common cause with separate cause explanations, discussed in Chapter 2, turns out to differ in important respects from the problem of comparing two common cause models that differ over whether an intervening variable should be introduced. Parsimony pertains to both, but it does so in different ways.

I pointed out a couple of times in this chapter that a hypothesis that postulates a causal chain from *C* to *E*, and one that postulates a causal chain from *C* to *I* to *E*, are evidentially indistinguishable if your evidence is merely the observed association between *C* and *E*. There is a connection here to Newton's famous dictum *hypotheses non fingo* – "I do not feign hypotheses" – in the *General Scholium* that was added to the second edition (1713) of his *Mathematical Principles of Natural Philosophy*:

> I have not as yet been able to discover the reason for these properties of gravity from phenomena, and I do not feign hypotheses. For whatever is not deduced from the phenomena must be called a hypothesis; and hypotheses, whether metaphysical or physical, or based on occult qualities, or mechanical, have no place in experimental philosophy.

Newton is here declining to add a conjecture concerning the mechanism that permits objects to exert gravitational attraction on each other. To add it without evidence would be to introduce a "mere hypothesis." Newton's idea applies to the question of whether one should add *I* to the causal chain linking *C* to *E* if your only evidence is the observed association of *C* and *E*. Even if you fervently believe that there is no action at a distance, it is pointless

[25] Recall from Chapter 2 that I take a null hypothesis to be one that says that there is *no difference* between two or more quantities. The label has nothing to do with which hypothesis you are trying to refute.

to interpolate causes unless you have empirical evidence for their existence. To prefer the proposition that *C* causes *E* over the hypothesis that there is a causal chain from *C* to *I* to *E* is to abide by the razor of silence, not the razor of denial. The former hypothesis feigns no hypothesis about an intervening variable.[26]

[26] After inventing the example of the blackbox in Figure 4.5 in September 2013, I realized that the example reflects a debt I owe to Fred Dretske – my mentor, colleague, and friend – who died in July of that year and who liked to draw analogies between minds and mechanical devices like thermometers. Fred was an engineer by training, so he came by such analogies naturally.

5 Parsimony in philosophy

> Our craving for generality has [as one] main source: our preoccupation with the method of science. I mean the method of reducing the explanation of natural phenomena to the smallest possible number of primitive natural laws; and, in mathematics, of unifying the treatment of different topics by using a generalization. Philosophers constantly see the method of science before their eyes, and are irresistibly tempted to ask and answer questions in the way science does. This tendency is the real source of metaphysics, and leads the philosopher into complete darkness.[1]
>
> Ludwig Wittgenstein, *The Blue Book,* p. 18

Naturalisms

For many philosophers and scientists, the word "naturalism" conjures up a metaphysical and a methodological thesis. Both concern objects that are out there in "nature," meaning things that exist in space and time. The contrast is with *super*natural entities; if such things exist, they exist outside of space and time:

> *Metaphysical naturalism*: the only things that exist are things in nature.
> *Methodological naturalism$_s$*: scientific theories should not postulate the existence of things that are outside of nature.

I put an "s" subscript on the second naturalism to mark the fact that it gives advice about doing science. These two naturalisms are the ones that get cited in discussions of the conflict between evolutionary theory and creationism. Evolutionary biologists often say that their theory obeys the requirements of methodological naturalism but is silent on the metaphysical question. They

[1] This quotation, from Ludwig Wittgenstein's *Blue and Brown Books* (New York: Harper and Row, 1965[1958], p. 18), is reprinted with the permission of John Wiley and Co.

further contend that creationism rejects both these naturalisms; here they are helped by creationists themselves, who often express their belief in a supernatural deity and argue that methodological naturalism$_s$ is a shackle from which science needs to break free. Although it is worth inquiring further into this interpretation of evolutionary theory and creationism, I won't do so here.[2] Rather, I am interested in a third naturalism. Like the second, it is methodological, but it is aimed at the practice of philosophy, not of science (hence the "p" subscript that I use to label it):

> *Methodological naturalism*$_p$: philosophical theories should be evaluated by the same criteria that ought to be used in evaluating theories in natural science.

There are trivial similarities linking science and philosophy that lend a superficial plausibility to this naturalistic thesis. For example, scientists are right to care about logical consistency and so are philosophers. However, there is another context in which naturalism$_p$ is far from obvious. As discussed in preceding chapters, scientists often evaluate their theories by seeing how parsimonious they are. In the present chapter, I'll examine a number of parsimony arguments that philosophers have constructed to evaluate philosophical theories. Taken at face value, this similarity seems to be grist for the naturalist's mill. But is the justification for using parsimony in philosophy really the same as the justification for using parsimony in science? In fact, it is a non-trivial question whether parsimony arguments are ever appropriate in philosophy, as the quotation from Wittgenstein shows.

One way to be a naturalist$_p$ about parsimony is to claim that parsimony arguments in science and parsimony arguments in philosophy have no justification whatever; they are on the same footing, namely none. I argued in Chapter 2 that this nihilistic verdict is profoundly mistaken with respect to science. Sometimes the epistemic credentials of parsimony arguments are clear. Bayesians should recognize that parsimony can reflect prior probabilities and likelihoods; frequentists should recognize that parsimony can play a role in model selection. The question for this chapter is whether any of these

[2] Mathematized evolutionary theory quantifies over numbers; if numbers are what Platonists say they are (entities that exist outside of space and time), then evolutionary theory violates methodological naturalism$_s$ (Sober 2011b). If "intelligent design theory" is formulated without specifying whether the postulated designer is supernatural, does it thereby obey methodological naturalism$_s$? I discuss this question in Sober (2007).

legitimate scientific uses of parsimony underwrite any of the parsimony arguments that are deployed in philosophy.

The principle of parsimony has figured in numerous philosophical debates. Here are the ones I'll discuss in this chapter:

- theism versus atheism – the problem of evil
- the mind-body identity theory versus dualism
- epiphenomenalism about the mental versus the thesis that mentalistic properties are causally efficacious
- moral realism versus anti-realism
- nominalism versus Platonism
- solipsism versus the hypothesis of an external world
- the problem of induction

In addition to investigating these seven topics, I'll make a few detours into neighboring epistemological territory.

My survey of this material will be incomplete in at least two respects. First, I do not pretend that the two parsimony paradigms that I think make sense in science are exhaustive; maybe there is more to scientific parsimony than is dreamt of in my philosophy. And second, even if a good parsimony argument in science must conform to one of the two paradigms, that would not settle whether naturalism$_p$ is correct in what it says about parsimony in philosophy. If there are parsimony arguments in philosophy that do not measure up to the justified applications of parsimony that are found in science, the naturalist$_p$ will say that these philosophical arguments should be committed to the flames, but the anti-naturalist$_p$ has the option of suggesting that parsimony should play a role in philosophy that is undreamt of in science. I will not try to decide who is right here. However, I confess that I often raise an eyebrow at philosophical parsimony arguments that float free from the scientific parsimony arguments whose justifications I can understand.

Atheism and the problem of evil

Atheists sometimes think that their thesis that there is no God is justified by Ockham's razor. Since they are atheists, not agnostics, their slicing away of the God hypothesis involves using the razor of denial, not the razor of silence. Is this application of Ockham's razor justified? In particular, do either of the parsimony paradigms discussed in Chapter 2 underwrite the atheistic conclusion? Here I want to consider the first paradigm – the idea that parsimony sometimes mirrors likelihoods. Is the greater parsimony of atheism over theism epistemically relevant because there are observations that favor atheism over theism in the sense of the law of likelihood?

In Chapter 1, I discussed Leibniz's thesis that we live in the best of all possible worlds. Leibniz's thesis provides a theistic response to the problem of evil: if God exists and is all-powerful, all-knowing, and all-good (all-PKG), how can there be so much evil? Atheists often respond to this question by arguing that the non-existence of an all-PKG God can be deduced from the fact that there is so much evil. This argument for atheism fails. There is no contradiction in the supposition that an all-PKG God exists alongside so much evil; God may permit evils to exist for reasons we cannot fathom. Atheists have a response to this logical point; they propose an *evidential* argument from evil that concludes that the evils we observe are evidence against the existence of an all-PKG God (Rowe 1979). This argument is more modest in its aspirations than the argument that attempts to prove that an all-PKG God cannot exist. The evidential argument can be formulated as having a likelihood inequality as its conclusion:

$$\Pr(S \mid \text{there is no all-PKG God}) > \Pr(S \mid \text{an all-PKG God exists}).$$

In this inequality, *S* is a fairly detailed summary of the kinds and quantities of evil that exist, not the bland statement that some evils exist. The question is not why there is some evil rather than none at all.[3]

Wykstra (1984) claims that the evidential argument makes the following assumption:

> If an all-PKG God had a reason for permitting horrendous evils to exist, human beings would know what those reasons are.

He rejects this assumption and advocates *skeptical theism*, which is the thesis that God exists, but we have little or no access to the reasons God has for permitting evils to exist. Is Wykstra right that the evidential argument from evil requires so strong a premise? Perhaps the argument is consistent with

[3] Howard-Snyder (1996) is a useful anthology of work on the evidential problem of evil. Draper (1989) represents the argument in terms of likelihoods. Tooley (2013) provides a useful introduction.

our having considerable uncertainty about what God's motives might be in allowing different evils to exist. We will see.

When defenders of the evidential argument point to an event *E* (as Rowe does when he considers a fawn that dies a slow, agonizing, and unobserved death in a forest fire) and claim that it is evidence against the existence of an all-PKG deity, skeptical theists reply that *E* might have good-making properties of which we are unaware that show that *E* isn't bad, all things considered. In saying this, skeptical theists are discussing the possibility of what I'll call "hidden compensating benefits." The skeptical theist isn't claiming to know what those benefits are; the claim is merely that such hidden benefits might exist. This suggestion has been criticized; the critics argue that if the possibility of hidden compensating benefits undercuts the evidential argument concerning the existence of God, it also undercuts ordinary moral judgments about the conduct of human beings. The claim is that skeptical theism entails moral skepticism, a conclusion that theists (and non-theists) usually want to avoid (Almeida and Oppie 2003; Jordan 2006).

To investigate what the commitments are of the evidential argument from evil, I want to consider a likelihood representation of the argument that concerns three propositions. The first two are about an event *E* that we know has occurred:

(*O*) Having considered the matter carefully, we conclude that event *E* would be horrendously bad, all things considered, unless there were good-making properties of *E* that compensate for *E*'s very substantial bad-making properties. Since we do not know of any good-making properties that *E* has that come close to outweighing the bad-makers, we conclude that *E* is horrendously bad, all things considered.

(*B*) Event *E* is horrendously bad, all things considered.

(*N*) There is no all-PKG deity.

Here is the argument:

$$\begin{array}{l} \Pr(O \mid B) > \Pr(O \mid notB) \\ \underline{\Pr(B \mid N) > \Pr(B \mid notN)} \\ \Pr(O \mid N) > \Pr(O \mid notN) \end{array}$$

Interpreted in terms of the law of likelihood, the premises say that *O* favors *B* over *notB* and that *B* favors *N* over *notN*. The conclusion is then drawn that *O* favors *N* over *notN*. This argument is invalid. However, if we add the following screening-off condition to the premises, the augmented argument is valid (Shogenji 2003):

(SO) $\Pr(O \mid B) = \Pr(O \mid B\&N)$ and $\Pr(O \mid notB) = \Pr(O \mid notB\&N)$.

Indeed, the argument remains valid if SO is weakened by replacing both occurrences of " = " with "≤" (Roche 2012).

$$\begin{tikzcd}
O \arrow[r, "+"'] \arrow[rr, bend right, "?"'] & B \arrow[r, "+"'] & N
\end{tikzcd}$$

To evaluate this likelihood argument and to understand why the argument as originally stated (without the screening-off assumption) is invalid, I'll rewrite the argument by using the Bayesian concept of confirmation:

$$\begin{array}{l}
\Pr(B \mid O) > Pr(B) \\
\Pr(N \mid B) > \Pr(N) \\
\hline
\Pr(N \mid O) > \Pr(N)
\end{array}$$

Each statement in this argument says that one proposition raises the probability of another. This Bayesian argument is equivalent to the likelihood argument. As noted in Chapter 2, there is a formal equivalence between the law of likelihood and Bayesian confirmation theory:

> *X* favors *Y* over *notY* (in the sense of the law of likelihood) if and only if *X* confirms *Y* (in the sense of Bayesian confirmation theory).

This Bayesian argument, like its likelihood counterpart, is invalid; the relation of probability raising is not transitive. To see why, consider the accompanying arrow diagram. Arrows with pluses below them represent probability raising; the arrow with a question-mark leaves open whether the probability is raised, lowered, or stays the same. Notice that there are two paths from *O* to *N*; one is indirect in that it passes through *B*; the other is direct. To assess whether *O* raises the probability of *N*, you need to consider both paths. The indirect path is positive because its two components are positive. If the direct path is negative, the overall effect of *O* on *N* may be negative. I hope this diagram makes it clear why probability raising isn't a transitive relation. However, a weakened screening-off assumption makes the Bayesian argument valid by assuring transitivity. It says:

$$\text{(Weakened-SO)}\ \Pr(N \mid B) \leq \Pr(N \mid B\&O)\ \text{and}\ \Pr(N \mid notB) \leq \Pr(N \mid notB\&O).$$

Weakened-SO is equivalent to the screening-off condition that I stated in connection with the likelihood version of the evidential argument from evil.

Why think that Weakened-SO is true? Let's start with its first conjunct. I think that $\Pr(N \mid B) = 1$. Since 1's in probability theory are sticky (p. 79), $\Pr(N \mid B\&O) = 1$ as well. The idea here is that if there are events that are horrendously bad, all things considered, then an all-PKG God does not exist. This is why theists have so often responded to the problem of evil by invoking the possibility of hidden compensating benefits. Now let's turn to the second conjunct in Weakened-SO. We can write that condition equivalently as:

$$\Pr(\text{An all-PKG God exists} \mid notB) \geq \Pr(\text{An all-PKG God exits} \mid notB\&O).$$

Suppose *notB* is true (the event *E* is *not* horrendously bad all things considered). Given this, what is the impact of proposition *O* (that, after careful consideration, we see lots of bad-making properties in event *E*, but no good-making properties that come close to compensating for the bad, and conclude that *E* is bad, all things considered) on the probability of an all-PKG God's existing? My view is that *O* does not raise the probability of such a God's existing, given that *notB* is true. If an event isn't bad, all things considered, our failing to grasp this fact isn't evidence for the existence of an all-PKG God.

Consider the first premise in the evidential argument from evil, either in the likelihood version or the Bayesian version. This premise says that *O* is evidence for *B*: if careful deliberation leads us to think that event *E* has many horrendous bad-making properties but we see no good-making properties that come close to outweighing the bad-makers, then that is evidence that *E* is horrendously bad all things considered. This premise is pretty minimal; denying it is tantamount to endorsing moral skepticism, which skeptical theists (and many others!) usually want to avoid. This is why I think that Wykstra is mistaken in his claim about what the evidential argument requires.

The evidential argument from evil is modest. It doesn't aim to prove the non-existence of an all-PKG God, nor even that there probably is no such being. To reach the latter conclusion, prior probabilities are needed and I see no way to defend a choice of priors. In addition, the fact that a given event counts against the existence of an all-PKG God does not preclude the possibility that other lines of evidence might count in favor.[4] Note also that the evidential argument won't convince people who are moral skeptics or who believe that good and bad can exist only if there is a God.

The word "parsimony" need not be used in stating the evidential argument from evil. However, atheism is more parsimonious than the hypothesis that an all-PKG God exists, and the evidential argument from evil is right that atheism has the higher likelihood, relative to the observation that there are lots of events that have the properties described by proposition *O*. We have here yet another case in which parsimony mirrors likelihood.[5]

The argument from evil assumes that God, if there is such a being, must be all-PKG. If that assumption is dropped, theists find it easy to explain why there is so much evil in the world. But suppose you don't just drop this specific assumption; you leave the concept of God so unspecified that the hypothesis that God exists says nothing about the observations we might ever make. Not only does your God hypothesis not deductively entail any observation; it doesn't even confer a probability on any observation, not even when supplemented by independently plausible auxiliary assumptions. Atheists often yearn to apply Ockham's razor in this instance, where the slicing away they have in mind involves denial, not silence. This use of the razor is difficult to defend. If "God exists" is untestable, presumably "God does not exist" is too.[6] If the two propositions are on the same epistemic footing, what justification can there be for treating them asymmetrically? Agnosticism is more tenable than atheism in this instance.

[4] For example, the fine-tuning argument says that the fact that the physical constants in our universe permit life to exist favors the hypothesis that our universe was made by God over the hypothesis that the constants received their values by mindless chance. See Sober (2009a) for discussion.

[5] Consider the hypothesis that God is responsible for all the observational truths. This is a grand unifying theory, but the suggestion that this theory is very simple begins to cloud when you formulate it as a model with adjustable parameters. Consider the observations we have made of the state of a series of dichotomous observational variables $\pm O_1$, $\pm O_2, \ldots,$ $\pm O_n$ (where n is large). The God model has n adjustable parameters, one for each variable. This model can fit the data perfectly, but it is not simple. It is like a model of the tosses of a coin that says that each toss has its own probability of landing heads. This model does not predict new data when fitted to old, so model selection criteria like AIC do not apply.

[6] As noted in Chapter 2, Bayesians concur with this symmetry thesis, but Popper does not.

Absence of evidence and evidence of absence[7]

The razor of denial says that if you don't need to postulate the existence of *X* to explain anything, then you should deny that *X* exists. Although this principle is stated in terms of explanation, not evidence, it is a close cousin to the following:

> If you have no evidence that *X* exists, then you should believe that *X* does not exist.

If believing that *X* does not exist requires that you have evidence that *X* does not exist, then this principle clashes with a slogan that scientists love to invoke: *absence of evidence isn't evidence of absence*.[8] Which of these principles is right? In fact, both are over-stated. Sometimes absence of evidence *is* evidence of absence, though often it is not. To see why, consider two example arguments that Douglas Walton discusses in his 1996 book, *Arguments from Ignorance*:

> I do not have any evidence that it is raining here and now.
> ————————————
> It is not raining here and now.

> I do not have any evidence that there is a storm on the surface of Jupiter now.
> ————————————
> There is no storm on the surface of Jupiter now.

Though neither argument is deductively valid, it is easy to turn them into valid arguments by adding a premise. Both arguments have the form:

> I do not have any evidence that *p* is true.
> ————————————
> *p* is false.

Just add the premise

(E_1) If *p* were true, then I would have evidence that *p* is true.

This further premise may be true in the case of the rain. Suppose, as in Walton's example, that I am sitting in a house with a tin roof and that I'd hear the characteristic pitter-patter if rain were falling. In contrast, it is easy

[7] Here I borrow material from Sober (2009a).

[8] As evidentialism requires – an idea I discussed in Chapter 1.

to imagine that E_1 is false in the case of the storm on Jupiter; suppose, instead, that

(E_2) If p were true, then I would have no evidence that p is true.

The Jupiter example is enough to show that the motto "absence of evidence isn't evidence of absence" is sometimes true, and the rain example is enough to show that it is sometimes false. The fact of the matter is that E_1 is true of some propositions in some circumstances and the same goes for E_2. This point leaves open what should be said about cases in which E_1 and E_2 are *both* false. When it is a matter of chance whether you'll have evidence as to whether a target proposition is true, a likelihood analysis can be undertaken (Sober 2009a).

Just as it may be useful to think about gremlins if you want to be clear on the difference between probabilities and likelihoods (p. 73), it may help to think about storms on Jupiter if you want to be careful about the razor of denial. The fact that you don't need to postulate a storm on Jupiter to explain what you have observed is *no reason whatever* to deny that there is a storm going on there. The razor of denial is over-stated.

The mind/body problem

Mind/body identity theorists sometimes argue for their theory, and against mind/body dualism, by invoking a principle of parsimony (Smart 1959; Brandt and Kim 1967). They do so by asking the reader to suppose that a perfect mind/brain correlation has been empirically discovered. To use an example much in vogue in the 1950s and 1960s, suppose that the c-fibers in a person's brain fire precisely when he or she is in pain.[9] How would this observation bear on the two philosophical theories? The identity theory advances the claim that

(*IT*) For every mental property M, there exists a physical property P such that $M = P$.

Dualism denies this proposition. Neither *IT* nor its negation mentions pain or c-fiber firings.

[9] I mention the brain here so that the identity theory doesn't run afoul of Bennett's (2003) point that c-fibers firing in a petri dish probably won't be accompanied by pain. It is c-fiber firing in an intact brain that is said to do the trick.

Is *IT* more parsimonious than its negation? When identity theorists and dualists survey what properties exist, they come up with different lists. Dualists count mental and physical properties separately. Identity theorists regard this as double counting and count physical properties alone. The identity theorist's list therefore comprises a proper subpart of the dualist's. This may sound like a parsimony comparison of the two mind/body theories, but notice how disconnected this numerology is from the two parsimony paradigms described in Chapter 2. When parsimony is a surrogate for likelihood, the more parsimonious hypothesis confers on the observations a higher probability than the less parsimonious hypothesis does. When parsimony is a consideration in model selection, you fit models to old data and attempt to estimate how accurately the models will predict new data. Property counting seems irrelevant. A way forward may be found by setting aside the lofty generality of *IT* and its negation and focusing on the specific case at hand, pain and c-fiber firing. I'll start by exploring what the model selection paradigm has to say about the relevance of parsimony to the question of whether pain and c-fiber firing are identical properties. After that, I'll turn to the likelihood paradigm. I hope it is clear that this example from the 1950s is only an example. The points I'll make about it apply to more up-to-date materialist characterizations of what pain is.

	The brain monitor says "c-fibers are firing."	The brain monitor says "no c-fibers are firing."
S pushes the button that says "pain."	f_1	f_2
S pushes the button that says "no pain."	f_3	f_4

An identity theory for the relation of pain and c-fiber firing will say that the two properties are identical; a dualist theory will say that they are distinct. What does each theory say about a simple experiment in which you trace each of several subjects through a sequence of trials? At each step, the subject (*S*) is either in pain or is not, and the subject's c-fibers are firing or they are not. You monitor pain at each step by having the subject push a button that says "pain" or a button that says "no pain." You monitor c-fiber firing via a brain scan of some sort. For each subject you therefore observe the values of four frequencies ($f_1 \ldots f_4$), which are represented by the cells in the preceding table. Each cell entry represents how frequently this or that conjunction is true.

	c-fiber firing	no c-fiber firing
pain	p_1	p_2
no pain	p_3	p_4

So far I have described the experiment in terms of what you observe. The next step is to construct models of the subjects' inner states. These inner states are described in the accompanying table; there are four conjoint states to consider, and each has one of four probabilities ($p_1 \dots p_4$). The four probabilities sum to one. The identity theory for pain and c-fiber firing says that two of the four conjoint events cannot occur, whereas dualism doesn't rule out any of them. This means that the identity theory has fewer adjustable parameters:

(Identity) $p_1 + p_4 = 1$

(Dualism) $p_1 + p_2 + p_3 + p_4 = 1$

Identity has one adjustable parameter, since your estimate for the value of p_1 in that model automatically provides an estimate for the value of p_4 (and *vice versa*). Dualism has three adjustable parameters, for the same reason.

How can these two models be connected with the data from your experiment? As I have emphasized, you don't observe the subject's pain and c-fiber firings. Rather, you observe *S*'s button pushes and the outputs of *S*'s brain scans. If your observations were error-free, the observed frequencies would furnish maximum likelihood estimates of the probabilities in these two models. But your observations may be subject to error. Here's a simple model that represents the possibility of error:

(*R*) Pr(*S* pushes the "no pain" button | *S* is in pain) = r_1
Pr(*S* pushes the "pain" button | *S* is not in pain) = r_2
Pr(the brain monitor says "c-fibers are not firing" | *S*'s c-fibers are firing) = r_3
Pr(the brain monitor says "c-fibers are firing" | *S*'s c-fibers are not firing) = r_4

R is a four-parameter model of error. I said before that Identity has one adjustable parameter and Dualism has three. But once the possibility of error is taken into account, the numbers change. Identity&*R* has 1&4 = 5 adjustable parameters and Dualism&*R* has 3&4 = 7. Notice that your data provide the values of four frequencies (which sum to 100 percent) and that both of these models have more than four adjustable parameters. This means that Identity&*R* and Dualism&*R* both fail to be *identifiable*. There is no such thing as *the* maximum likelihood estimate of the parameters in each, since many such assignments tie for first place. Using AIC on a model requires that the model be identifiable, as explained in Chapter 2. We are in a pickle.

The way out of the pickle is to have independent estimates of the probabilities in *R*. How often do people erroneously push the pain button? How often does the brain scan erroneously say that c-fibers are firing? Perhaps other experiments can provide answers to these questions. Let's suppose that our best estimates are that the errors have low (but positive) probabilities. The upshot is that Identity will have a better AIC score than Dualism if the two frequencies f_2 and f_3 are close to zero.[10]

Functionalism is a third theory of the mind/body relation. It was much discussed in the 1960s and 1970s and was widely thought to be superior to both the identity theory and dualism. It still is popular. Functionalism says that mental properties *supervene* on physical properties and that mental properties are *multiply realizable*. Supervenience means that individuals *cannot* differ in their mental states unless they differ physically; multiple realizability means that individuals *can* differ physically even when they are in the same mental state. Applying this functionalist format to the relation of pain and c-fiber firing, we obtain a third model:

(Functionalism) $p_1 + p_2 + p_4 = 1$

This model says that $p_3 = 0$, reflecting the model's claim that pain supervenes on c-fiber firing. This commitment means that the probability is zero that a person is not in pain if his or her c-fibers are firing. Functionalism allows for the possibility that p_2 is greater than zero, since multiple realizability means that individuals may be in pain even when their c-fibers are not firing. The Functionalism model has two adjustable parameters, so it is more parsimonious than Dualism but less parsimonious than Identity. If the frequency f_2 isn't close to zero, but f_3 is, Functionalism will have a better AIC score than both Identity and Dualism.[11]

[10] How close is close enough? That depends on sample size.

[11] The three mind/body models I have described are nested. Identity entails Functionalism, and Functionalism entails Dualism. If you prefer that the models not be nested, you can stipulate that a model's adjustable parameters are all positive. This stipulation will not affect the AIC analysis, as noted in Chapter 2 (p. 143). Notice also that

Setting aside this model selection analysis, I now want to consider whether the law of likelihood can be used to assess whether the identity theory is better than dualism. This means considering the following inequality:

Pr(pain reports and c-fiber scans are positively associated | being in pain $=$ having one's c-fibers fire) $>$

Pr(pain reports and c-fiber scans are positively associated | being in pain $\neq$ having one's c-fibers fire).[12]

It is fairly straightforward to say that the identity claim renders a positive association highly probable (if observational errors have low probabilities). What is harder to evaluate is how probable the association is under the dualist hypothesis that the properties are not identical. There are many "ways" that the non-identity claim could be true. To assess the likelihood of the non-identity hypothesis, you'd need to consider how probable dualism says each of these many ways is. I see no hope for doing this in a defensible manner. This problem resembles the one described in Chapter 2 concerning the average likelihood of a model that says that two fields of corn might have different average heights (p. 138). It also is similar to the problem of computing the average likelihood of Simon Newcomb's model of gravitation, which he constructed in order to cope with data on the precession of the perihelion of Mercury (p. 127). This is why I think that parsimony considerations in the mind/body problem make more sense when you use model selection ideas than when you use the law of likelihood.

It is a standard point in discussion of the mind/body problem that Cartesian dualism is logically compatible with there being a perfect association between mental and physical states. The conclusion is then drawn that observing a perfect association between the two isn't evidence that discriminates between dualism and materialism. This argument is fallacious; from the fact that O is logically compatible with H_1 while O is entailed by H_2, it does not follow that O fails to discriminate between H_1 and H_2. This doesn't follow if you use the law of likelihood *or* if you use model selection ideas. It is a mistake to think that the

I have construed pain and c-fiber firing as dichotomous, rather than quantitative, characteristics. It would be interesting to explore mind/body models that describe the relationship between the intensity of pain and the frequency of c-fiber firings.

[12] The observation is that pain reports and c-fiber scans are positively associated to some high degree, say 95 percent, not just that they are positively associated to some degree or other.

relation of logical compatibility is the be-all and end-all of assessing evidential import. In Chapter 2, I discussed the models NULL and DIFF. NULL predicts that when you sample corn plants from the two fields, the sample means will probably be very close together. DIFF predicts no such thing, but it can accommodate sample means that are close together if that is what you happen to observe. In Chapter 1, I discussed how Copernican astronomy predicts regularities that the Ptolemaic system can only accommodate. The difference between prediction and accommodation is epistemically significant in these scientific cases, and it also is significant in the mind/body problem.

Is dualism really committed to the probability model I called "Dualism"? What about a version of dualism that asserts that pain and c-fiber firing are non-identical but then adds that the probability of each without the other is zero? Here dualism buys into the probability model that the identity theory advances, but departs from the identity theory purely at the level of metaphysics. Model selection theory cannot adjudicate here. However, before you claim that this "hard problem" is the real problem, think about what dualists will say if new observations strongly suggest that the probabilities of pain without c-fiber firing and of c-fiber firing without pain are positive. Dualists will be quick to point out that their position is not cast in doubt by this result. This suggests that the version of dualism just described (in which the dualist embraces the identity theory's probability model) doesn't really capture what dualism involves. This is why I think that my four-parameter model of dualism is closer to dualism's actual commitments.[13]

As noted at the start of this section, discussion of the identity theory and dualism often begins with the idea that we observe various mind/brain "correlations." It is sometimes unclear whether these authors use "correlation" to denote a relationship among probabilities that goes beyond the data at hand, or merely use it to refer to a relationship among sample frequencies. I have used "correlation" to denote the former and "association" to denote the latter. Keeping with that usage, I want to say that it is a mistake to think that the data before us consist of *correlations*. We observe *associations* in a finite

[13] Of course, it is also true that the identity theory, as represented solely by IT, is not committed to the claim that being in pain and having one's c-fibers fire are identical properties. What I am describing is the *application* of IT to any candidate pair of mental and physical properties and the application of IT's negation to that same pair.

sample, not probabilistic correlations that hold in the past, present, and future. It also is important to recognize that we do not observe the pains of others and their c-fiber firings; what we observe is their behaviors (including verbal reports) and the outputs of various brain scans.[14] These observations provide fallible evidence for the occurrence of pain and the occurrence of c-fiber firings. Once our conception of the observations is whittled down to size, it can fit within the framework of model selection theory, which requires that we view our present data as providing guidance about what future data sets will be like. We can't do this if we start with the "observation" that there is a perfect probabilistic correlation.

Smart (1959, pp. 155–156) compares the face-off between dualism and the identity theory with the contest between the theory that "the universe just began in 4004 BC, with sediment in the rivers, eroded cliffs, fossils in the rocks and so on" and a geological theory that postulates an ancient earth in which rivers, cliffs, and fossils are gradually formed. The former is the theory that Philip Gosse (1810–1888) presented in his 1857 book *Omphalos*, shorn of its mention of God (Gould 1985). Smart says that the Gossean theory "offends against the principles of simplicity and parsimony" because it postulates "far too many brute and inexplicable facts."[15] Smart's point in proposing this analogy is that the mind/brain identity theory is more belief-worthy than dualism because the identity theory explains observations that dualism cannot.

I have two reservations about Smart's analysis. First, every theory contains postulates that the theory does not explain. If a theory has several such, the number of unexplained explainers can be reduced to one by postulating a common cause that covers them all. This is often not a good idea; Newton's remark "I do not feign hypotheses" is important to bear in mind. Though the

[14] This is a remark about the epistemology of these experiments, not a comment about perception. It is the experiment that dictates what counts as an observation and what counts as an inference from those observations. In another context, it might be perfectly correct to say that you see that someone is in pain. I'll discuss the theory-neutrality of observation.

[15] This is the second appeal to parsimony that Smart (1959) makes. In the first (pp. 142–143), he offers an inductive argument that extrapolates from the past triumphs of materialist reductionism in science to a claim about what one ought to expect in the future. This argument from the history of science is interesting in that Smart thinks of the future's resembling the past as an instance of parsimony or simplicity, an idea that connects with Hume's views about induction, which I discussed in Chapter 1 and will return to later in the present chapter.

de-theologized Gossean theory that Smart describes offers no explanation of what existed in 4004 BCE, Gosse's original theory does not have this deficiency. Gosse said that the world had those features because God put them there. I am not saying that the God hypothesis is good; rather, my point is that the God hypothesis, if true, would explain the state of the world some 6,000 years ago. Second, I suspect that the focus on explanatoriness fails to get at the heart of the matter. The central question is whether the observations discriminate between the identity theory and dualism. If they do not, it doesn't matter that the identity theory provides a better explanation of those observations (Roche and Sober 2013; Sober 2015). The answer to the more fundamental question about evidence is that the observed association *can* point to an epistemically relevant difference between the identity model and a dualist model. AIC explains why.

Brandt and Kim (1967, pp. 533–534) agree that parsimony is a tie breaker in the mind/body problem, but claim that the relevance of parsimony to this philosophical problem is importantly different from the relevance of parsimony to a scientific problem. They say that Smart's analogy between the mind/body problem and geology "is pernicious in that it lends, or at least tends to lend, a false air of scientific respectability to what is essentially a philosophical and speculative interpretation." Smart embraces, while Brandt and Kim reject, naturalism$_p$ as a thesis about parsimony. The model selection analysis I have described leads me to side with Smart.

Placing the mind/body identity theory, functionalism, and dualism within the context of model selection criteria like AIC means comparing the contending models for their predictive accuracies. Metaphysicians may balk at this, proclaiming that they don't care about predictive accuracy and want only to figure out which philosophical theory is true. In reply, I return to a theme from Chapter 2: model selection criteria aim at finding fitted models that are *close* to the truth. AIC has a connection with instrumentalism, but it also is connected to realism. Metaphysicians should not disdain the finding that experiments of the kind I have described may indicate that an identity model is *closer to the truth* than a dualist model.

The causal efficacy of the mental

Assuming that human beings and other organisms have psychological properties, why think that their possessing those properties causes behavior? If the behavior is caused by the organism's having various physical properties, isn't it unparsimonious to claim that there are psychological causes as well? Why postulate two causes when one will do?[16] Should physicalists avoid the luxury of superfluous causes and assert that our having the beliefs, desires, and sensations we do are epiphenomenal correlates of behavior, not causes?

	A	$notA$
B	$x+a+b$	$x+b$
$notB$	$x+a$	x

As discussed in Chapter 2, the idea that a one-cause model is more parsimonious than a two-cause model finds a natural representation within model selection theory. There I considered the case of two dichotomous properties *A* and *B* that each may be a cause of the dichotomous property *E*. The probability of *E*, conditional on different combinations of ±*A* and ±*B*, is shown in the accompanying 2×2 table. A model that says that both *A* and *B* are (or may be) causes of *E* will take the form.

(TWO)
$$\begin{aligned}&\Pr(E \mid A\&B) - \Pr(E \mid notA\&B) = a\\&\Pr(E \mid A\&notB) - \Pr(E \mid notA\&notB) = a\\&\Pr(E \mid A\&B) - \Pr(E \mid A\&notB) = b\\&\Pr(E \mid notA\&B) - \Pr(E \mid notA\&notB) = b\end{aligned}$$

This model allows that varying the state of ±*A* while holding fixed the state of ±*B* may make a difference in the probability of *E*, and that the same is true of varying the state of ±*B* while holding fixed the state of ±*A*.[17] In contrast, the model that says that only ±*A* is (or might be) a cause of *E* takes the form:

(ONE)
$$\begin{aligned}&\Pr(E \mid A\&B) - \Pr(E \mid notA\&B) = a\\&\Pr(E \mid A\&notB) - \Pr(E \mid notA\&notB) = a\\&\Pr(E \mid A\&B) - \Pr(E \mid A\&notB) = 0\\&\Pr(E \mid notA\&B) - \Pr(E \mid notA\&notB) = 0\end{aligned}$$

[16] These questions are in the background of Kim's (1993, 1996) discussion of nonreductive physicalism, and they trace back to Nozick's smart Martian problem, discussed in Dennett (1980) and in Sober (1999c).

[17] For simplicity, I assume that this model says that ±*A* and ±*B* are related additively. To allow for a non-additive relationship, an additional parameter, an interaction term, would need to be introduced. Recall the example about classroom size and teacher experience in Chapter 2.

This model says that varying the state of B while controlling the state of A makes no difference in the probability of E. TWO has two adjustable parameters while ONE has one. TWO will fit frequency data at least as well as ONE does, but ONE is more parsimonious. Depending on the data, a model selection criterion like AIC might award ONE the better score.

As an example, consider the question of how smoking and asbestos exposure are each related to lung cancer. Perhaps both are causes, or just one of them is. Indeed, there is the even simpler null model that says that neither of them makes a difference in the risk of lung cancer. Frequency data can be gathered that allow the models to be compared. The data you need here requires that some people smoke and others do not and that some people are exposed to asbestos and others are not. Without frequency data pertaining to the four cells of the 2×2 table, there is no fitting of models to data, and no estimating of predictive accuracy.

Can this format be applied to the problem at hand in which a purely physical explanation of a behavior is compared with an explanation that postulates both physical and mental causes? Sometimes this comparison is straightforward. Suppose you want to consider whether studying for an examination helps students do better and also whether drinking coffee has a positive effect. You could run an experiment to see whether there are two causes here, or one, or zero. However, this is not the kind of case that the parsimony argument for epiphenomenalism invites you to contemplate. You are asked to consider a *physically complete* explanation and then ask whether it makes sense to supplement this physically complete story with the postulation of mentalistic causes. Now there is a difficulty. The ONE and TWO models requires you to consider what would happen if each putative causal factor were varied while holding the other fixed. This can't be done in the case at hand if mental properties *supervene* on physical properties. Here supervenience means that it is *impossible* to vary an individual's mental characteristics while holding fixed the individual's complete physical state.[18] Because of this, some of the conditional probabilities described in ONE and TWO will fail to be well-defined (Shapiro and Sober 2007). If they aren't well-defined, it won't be possible to estimate the values of the relevant parameters, and so it won't be possible

[18] I am using a thesis of *strong* supervenience. It isn't merely that every psychological difference *happens* to be accompanied by some physical difference or other. The thesis is that this pattern *must* obtain, given the laws of nature.

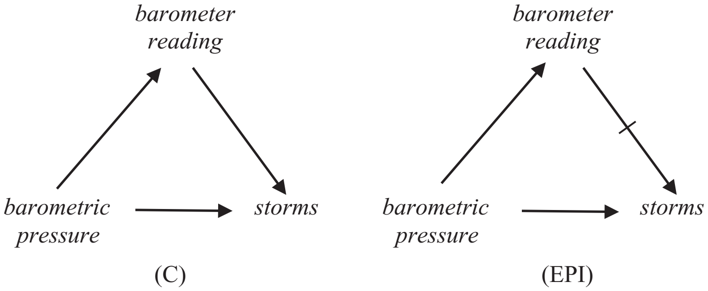


Figure 5.1

to compute AIC scores for the models. The parsimony argument for epiphenomenalism, in which a one-cause model cites the physical property *P* and a two-cause model describes *P* and the mental property *M* (where *M* supervenes on *P*), looks pretty bad when it is viewed through the lens of model selection theory.

How, then, should you test the hypothesis that various mental characteristics are causes of behavior against the hypothesis that they are mere epiphenomenal correlates? A mundane and often-used example provides useful guidance. Why think that barometer readings don't cause storms? The two hypotheses you need to consider are depicted in Figure 5.1. They agree that barometric pressure is a common cause of barometer readings and storms, and both predict that barometer readings and storms will be correlated. However, they disagree about whether barometric pressure *screens-off* barometer readings from storms. The epiphenomenalist model (EPI) says it does:

Pr(storm | high barometric pressure & the barometer reads high)
= Pr(storm | high barometric pressure & the barometer reads low).

Pr(storm | low barometric pressure & the barometer reads high)
= Pr(storm | low barometric pressure & the barometer reads low).[19]

The causal hypothesis (C) denies these equalities; it says that barometer readings affect the probability of storms even when you control for barometric pressure. This difference between the two hypotheses can be translated into the language of model selection theory. The epiphenomenalist model is a null hypothesis; it says that there is no difference between various probabilities.

[19] For the sake of simplicity, I here treat screening-off as a relation among dichotomous propositions, rather than in terms of the relations among continuous variables.

In contrast, the causal model has adjustable parameters that attach to various probability differences. Frequency data can be used to estimate adjustable parameters and a model selection criterion like AIC can be applied. Parsimony is relevant here, and the epiphenomenalist model *is* more parsimonious.

Why does the screening-off argument against mental causation go up in smoke while the screening-off argument against barometer readings' causing storms goes smoothly? The reason is that I considered the *supervenience bases* of mental properties in the one case but the *common causes* of barometer readings and storms in the other. The relevant conditional probabilities are not defined in the former context, but they are in the latter. This opens the door to a new and better version of the problem of mental causation. We should consider how mental properties and behaviors are related to their physical *common causes*, not how mental properties are related to their supervenience bases. Epiphenomenalist and causal models are both legitimate hypotheses, and the fact that mental states are correlated with behaviors isn't evidence that favors one over the other. Epiphenomenalism *is* the more parsimonious model, but that does not suffice to show that it is better.[20] What one needs are data (Shapiro and Sober 2007).

Moral realism

Moral realism is the thesis that some normative ethical propositions are true and their truth is independent of anyone's thinking that they are true or being inclined (in some appropriate circumstance) to approve of them.[21] So defined, moral realism is incompatible with the divine command theory, moral relativism, and some forms of existentialism. According to realism, what makes an action wrong isn't the fact that God, society, or an individual disapproves of it. Moral realism, as I understand it, does not say that ethical

[20] Epiphenomenalism is more parsimonious than the thesis that mental properties are causally efficacious for the same reason that the behavior-reading model is more parsimonious than the mind-reading model discussed in the previous chapter. In both problems, the assertion of screening-off has fewer adjustable parameters than its denial.

[21] It is *normative* ethical propositions that are at issue here. It is not in dispute that there are descriptive propositions about morality that are true. The contrast is between "torture is wrong" and "torture was thought to be morally permissible by the administration of George W. Bush in the aftermath of 9/11."

truths are independent of human psychology; hedonistic utilitarianism, for example, can be given a realist interpretation even though it says that moral statements are made true by facts about pleasure and pain.

Harman (1977) argues against moral realism by way of a parsimony argument.[22] He contends that there is no need to postulate the existence of an independent realm of normative ethical truths if we wish to explain human thought and behavior. A person's upbringing suffices to do the explaining. Harman concludes, not that we should remain silent on whether independent moral truths exist, but that there are no such things. Ruse and Wilson (1986) advance a similar argument, suggesting that evolutionary theory suffices to explain why we have the ethical beliefs and feelings we do and we therefore should be anti-realists about ethical truths.[23] Both arguments deploy the razor of denial, not the razor of silence.

Parsimony arguments against moral realism are sometimes given a *causal* formulation. We know that there are various descriptive facts (including facts about upbringing and evolution) that are causes of human thought and action. Why postulate a set of normative ethical facts, whose truth is independent of anyone's say-so, as a second set of causes? If normative ethical facts *supervene* on descriptive facts (for example, as utilitarianism maintains), then this parsimony argument runs into the same trouble that derails the parsimony argument against mental causation just discussed. It is a mistake to expect supervening causes to have causal powers that go beyond those exhibited by their supervenience bases (Sturgeon 1984; Kim 1993; Sober 1999c). This follows from a manipulationist account of causation; when variable *X*

[22] Harman holds that normative ethical statements are sometimes true, but when they are true, they are true because they reflect our inclinations to approve or disapprove of various actions; this makes Harman a *relativist*, not a *realist*.

[23] Street (2006) gives an evolutionary argument against moral realism, but her target differs from that of Harman and Ruse and Wilson. Her main thesis is that evolutionary considerations show that true normative ethical propositions would be unknowable if moral realism is right in its claim that they are true independently of our inclinations to assent to them. This particular argument does not appeal to Ockham's razor. Even so, Street (p. 129) does make a parsimony argument against moral realism. Joyce (2006) is critical of some parsimony arguments against moral realism, but settles on a version that he thinks works (pp. 209–211); it involves the razor of silence, not the razor of denial. Joyce's conclusion is that we aren't justified in believing any moral proposition; he is making an epistemic claim that does not deny the strictly metaphysical thesis that I am considering.

supervenes on variable *Y*, *X* has a causal impact on *Z* only if *Y* also has a causal impact on *Z* (Shapiro and Sober 2007).

What if we drop the causal version of this parsimony argument and stick to the concept of explanation? Harman says that upbringing provides the best explanation of why we think and act as we do and that the realist's postulate of an independent realm of normative ethical facts isn't needed to do the explaining. Sturgeon (1984) responds that normative ethical facts *are* sometimes explanatory; for example, Sturgeon thinks Hitler's starting World War II is explained by the fact that Hitler was morally depraved. Although I disagree with Harman's argument, I have two hesitations concerning Sturgeon's response to it. First, "moral depravity" is a mixed concept, in that it combines descriptive and normative elements. It may be that the descriptive content is doing all the explanatory work. Sturgeon's proposed explanation of why Hitler started World War II may be like the suggestion that the lemonade dropped in temperature because you dropped an ice cube into it; the suggestion may be correct, though it is the iciness and not the cubical shape that did the causal work. My second hesitation is that even if Hitler's moral depravity is explanatory, it isn't clear that this is the *best* explanation of why Hitler started the war. Harman could grant that the moral depravity has some explanatory oomph and still maintain that it is a very poor explanation. Many historians follow the practice of excluding moral judgments from the explanations they give, and I do not propose to tell them that they are making a mistake.

The reply I prefer to Harman's parsimony argument against moral realism is to reject the requirement that normative ethical facts, if they exist, should explain human thought and behavior. Ethics is in a different line of work from psychology. Psychology has the job of explaining human thought and behavior. Normative ethical propositions have the job of telling us how we *ought* to act, not of explaining why we act as we do (Sober 1990a; Shafer-Landau 2007). The distinction I have in mind also needs to be drawn between logic and psychology. Let logical realism be the view that there are true normative propositions about what we ought to conclude from given premises and that these propositions are true independently of anyone's say-so or inclinations to assent to them. Logical realism may or may not be correct, but it is a mistake to evaluate it as if logic were a descriptive psychological theory.

If normative ethical propositions should not be expected to explain our observations of what human beings think and do, are there other observations that these propositions should be able to explain? That depends on what you mean by *observation*. It has become standard in philosophy of science to say that observations are "theory laden." This means that the propositions that scientists properly treat as observation statements are knowable only by agents who have a relevant theory in their possession. A physicist who looks at the screen of a cloud chamber can observe that an electron is moving from right to left, but a person with no knowledge of physics will be unable to see that this is what is happening. The thesis that observations are theory-laden is sometimes thought to lead to relativism, but, in fact, no such dire consequence is in the offing. Observations can be theory-laden and still provide a neutral basis for discriminating between competing hypotheses. If an observation statement O is to help you discriminate between theories T_1 and T_2, it must be possible for you to know that O is true without already having to believe either T_1 or T_2. However, that leaves it open that knowing O may require using some other theory, T_3. What matters in science is that observations be *relatively* theory-neutral, not that they be *absolutely* theory-neutral (Sober 2008a, 2008b).

Given this, I see no problem with regarding some normative ethical propositions as observation statements. Harman (1977) agrees and gives a good example. You see a group of people set fire to a cat and the thought leaps to mind that what they are doing is wrong. If the judgment that this act is wrong is an observation statement, you can ask different ethical theories to explain why it is true and then evaluate those theories by the quality of the explanations they provide. If so, it is a mistake to maintain that normative ethical propositions aren't needed to explain *anything*; some normative propositions are needed to explain others. I do not suggest that this is an argument for moral realism; there is no denying that it begs the question against moral nihilism, which is the view that normative ethical statements are never true. Still, if some normative propositions are true, others may be needed to explain them; a parallel claim applies to descriptive propositions (Sturgeon 1984).

In summary, the parsimony argument against moral realism goes wrong on three fronts. First, the model-selection rationale for preferring models that postulate one cause over models that postulate two depends on the possibility of varying each putative cause while holding fixed the other, but this cannot be done when one candidate cause supervenes on the other. Second, normative ethical propositions should not be evaluated by their ability to explain descriptive propositions about human thought and behavior. And third, even though normative ethical propositions aren't needed to explain what we think and do, it doesn't follow that they aren't needed to explain anything.

Misinterpreting screening-off

As just explained, anti-realists about morality sometimes argue that you don't need to postulate moral facts (understood in the way realists propose) if you are to explain human thought and action; evolution and upbringing suffice. Epiphenomenalists about the mental have argued similarly, claiming that you don't need to postulate mental causes if you are to explain human behavior; neurophysiological causes suffice. The razor of denial is then trotted out, and two philosophical theses are said to be refuted – moral realism on the one hand and the causal efficacy of the mental on the other.

There are variant versions of these two arguments that bypass Ockham's razor and appeal to a criterion of explanatory relevance — or rather, of irrelevance:

> I is explanatorily irrelevant to E if there exists a proposition C that is explanatorily relevant to E and $\Pr(E \mid C) = \Pr(E \mid C\&I)$. .

In Chapter 2, I defined screening-off as a relation that holds among three *variables* (p. 75). The term is also sometimes used to describe a relation among three *propositions*. So, noting this departure from the previous definition, we can describe the above criterion as saying that I is explanatorily irrelevant to E if C is explanatorily relevant to E and C screens-off I from E.

The example of the barometer, depicted in Figure 5.1, makes this criterion sound like it is on the right track. The barometer reading is not explanatorily relevant to why there is a storm and the criterion just formulated seems to show why: the barometric pressure explains the storm and the pressure screens-off the barometer reading from the storm. The same point holds for an example in Chapter 2 from Mendelian genetics about two siblings and their parents: offspring 1's genotype is not explanatorily relevant to offspring 2's genotype because the parental genotype explains 2's genotype and the parental genotype screens-off 1's genotype from 2's.

Anti-realists about morality and epiphenomenalists about the mental do not explicitly commit to the screening-off criterion of explanatory irrelevance just stated. The usual pattern is that they make assertions about explanatory irrelevance without bothering to describe how that concept should be understood. Even so, it is worth recognizing that screening-off is a poor criterion for explanatory irrelevance. It seems fine when a common cause screens-off one effect from the other, but it has unacceptable consequences elsewhere. Consider a causal chain that goes from a distal cause (*D*) to a proximate cause (*P*) to an effect (*E*) where *P* screens-off *D* from *E*. For example, suppose I dial your telephone number, your telephone rings, and then you pick up. The first causes the second, and the second causes the third. Before the days of caller ID, the ringing of your phone screened-off my calling from your picking up. Given that your phone is ringing, the probability of your picking up is the same, whether or not I dialed your number. However, it simply isn't true that my dialing your number is explanatorily irrelevant to your picking up. In a $D \rightarrow P \rightarrow E$ causal chain, we should not conclude that *D* is explanatorily irrelevant to *E* just because *P* is relevant and *P* screens-off *D* from *E*.

This point is relevant to an influential argument against reductionism advanced by Hilary Putnam. The reductionism that Putnam aims to refute says the following:

> (Micro-reductionism) Every event has a micro-explanation, and if the event also has a macro-explanation, the micro-explanation is better.

Putnam (1975) attacks this thesis by presenting a simple example about a board and a peg. The board has two holes. One is round and is a bit more than 1 inch in diameter while the other is square and is a bit more than 1 inch on each side. The peg is a cube that is 1 inch on each edge. The peg fits through the square hole, but not the round one. Why? Putnam contends that the correct explanation is given by the macro-dimensions just cited. He claims, in addition, that a detailed micro-description of the configuration of all the many molecules in the peg and board is either not an explanation, or is a terrible explanation, because the micro-story involves lots of irrelevant detail. Putnam concludes from this that micro-reductionism is false. Although Putnam does not explicitly embrace a criterion that describes what explanatory irrelevance is, the screening-off criterion seems to deliver the verdict he wants about the example.[24] If the screening-off criterion is rejected, as I think it should be,

[24] In fact, this appearance is deceiving. Consider the MACRO and the MICRO descriptions of the peg and board system. Each screens-off the other from the observation *O* that the peg passes through one hole but not through the other in the sense that $\Pr(O \mid \text{MACRO}) = \Pr(O \mid \text{MACRO \& MICRO}) = \Pr(O \mid \text{MICRO})$.

we face a question: is there a plausible criterion of explanatory irrelevance that allows Putnam to draw his anti-reductionist conclusion (Sober 1999b)? I leave this as a problem for the reader to ponder and now turn to a different argument in which screening-off is misinterpreted.

The argument I want to examine is Alvin Plantinga's much-discussed criticism of "evolutionary naturalism," by which he means the conjunction of evolutionary theory and atheism.[25] Plantinga doesn't argue that this conjunction is false; rather, his thesis is that no one can rationally believe it. According to Plantinga, if you believe evolutionary naturalism, this belief of yours instructs you to distrust everything you believe, including evolutionary naturalism itself; belief in evolutionary naturalism is therefore self-undermining. What people ought to believe, Plantinga suggests, is evolutionary theory conjoined with theism; this conjunction, he says, does not shoot itself in the foot.

Plantinga (2011, p. 315) develops his argument against evolutionary naturalism by endorsing an idea of Patricia Churchland's:

> Boiled down to essentials, a nervous system enables the organism to succeed in the four F's: feeding, fleeing, fighting, and reproducing. The principal chore of nervous systems is to get the body parts where they should be in order that the organism may survive . . . Improvements in sensorimotor control confer an evolutionary advantage: a fancier style of representing is advantageous *so long as it is geared to the organism's way of life and enhances the organism's chances of survival*. Truth, whatever that is, definitely takes the hindmost. (Churchland 1987, p. 548)

My interpretation of Churchland is that she is making a point about screening-off:

(CSO) Pr(M evolves | M promotes survival and reproductive success & M is reliable) =
Pr(M evolves | M promotes survival and reproductive success & M is not reliable).[26]

[25] Plantinga's formulation of this argument has evolved. His most recent version is in Plantinga (2011, pp. 307–346); this is the version I'll discuss.

[26] Although I am using a characterization of screening-off that differs from the two described in Chapter 2, it is equivalent to the others; $\Pr(X \mid Z\&Y) = \Pr(X \mid Z\&\text{not}Y)$ precisely when $\Pr(X \mid Z\&Y) = \Pr(X \mid Z)$, provided that all the conditional probabilities are well-defined.

Here *M* is a mental mechanism that generates beliefs from sensory inputs. Plantinga takes "*M* is reliable" to mean that most of the beliefs that *M* generates are true (or approximately so). According to Churchland, if *M* promotes survival and reproductive success, the probability of *M*'s evolving is the same, whether or not *M* is reliable. Churchland's point might be put metaphorically by saying that natural selection "cares" about survival and reproduction, not about getting organisms to have true beliefs. Here as elsewhere, the price of metaphor is eternal vigilance.[27]

CSO has nothing special to do with human cognitive equipment. A screening-off thesis also can be formulated about zebra leg morphology (*L*):

(ZSO) Pr(*L* evolves | *L* promotes survival and reproductive success & *L* helps zebras to outrun predators) =
Pr(*L* evolves | *L* promotes survival and reproductive success & *L* does not help zebras to outrun predators).

CSO and ZSO are on the same page; if natural selection doesn't "care" about getting human beings to have true beliefs, it also doesn't "care" about getting zebras to outrun predators. But watch out for the metaphor of caring! ZSO does not undercut the hypothesis that *L* evolved because it helped zebras evade predators nor does CSO undercut the hypothesis that *M* evolved because it helped our ancestors obtain true beliefs. The CSO and ZSO equalities do not rule out these two inequalities:

Pr(*M* evolves | *M* is reliable) > Pr(*M* evolves | *M* not reliable).

Pr(*L* evolves | *L* helps zebras outrun predators) >
Pr(*L* evolves | *L* does not help zebras to outrun predators).

Nor does CSO entail that

Pr(*M* is reliable | *M* evolved and God does not exist) is low,

which is the conclusion that Plantinga erroneously draws from Churchland – that evolutionary naturalism says that it is very improbable that our cognitive mechanisms for constructing beliefs are reliable (Fitelson and Sober 1998). Churchland's idea concerning what natural selection "cares" about needs to be understood carefully.

[27] Lewontin (2001) attributes this precept to Arturo Rosenbleuth and Norbert Weiner.

Nominalism and Platonism about mathematics

Even though William of Ockham did not use his razor to defend his nominalism, other philosophers, including the seventeenth century inventor of the razor metaphor, have thought that the tool and the *ism* are connected. Quine (1953a, p. 4) was part of that long tradition. Because of his "taste for desert landscapes," he wanted to exclude abstract entities from his ontology.[28] However, Quine came to believe that this austere ideal is unattainable, owing to the fact that mathematized science can't do without abstract entities. Here I want to consider how the two parsimony paradigms discussed in Chapter 2 bear on the contest between nominalism and Platonism. This debate can be considered in connection with the existence of universals and also with respect to the existence of mathematical objects (e.g., numbers). Platonism claims that properties and numbers exist and that they exist outside of space and time; nominalists deny the existence of properties and numbers (at least as these are Platonistically conceived). I will focus on numbers.

Nominalists about mathematics pursue a dual strategy – translate what you can and deny the truth of the rest. The first line of attack is to show that various mathematical propositions can be paraphrased into a nominalistically acceptable language. Consider, for example, the fact that

(Apples) There are exactly two apples in the basket.

Nominalists suggest that Apples be understood as follows:

(*N*) There exist physical objects x and y such that x is an apple in the basket and y is an apple in the basket and $x \neq y$, and for all z, if z is an apple in the basket, then $z = x$ or $z = y$.

An alternative construal of Apples is Platonistic:

(*P*) There exists a number n such that $n =$ the number of apples in the basket and $n = 2$.

N commits to the existence of physical objects while *P* commits to the existence of numbers. If *N* is a good enough paraphrase of Apples, then we don't need to assert that *P* is true to say what we want to say about them apples. However,

[28] Goodman and Quine (1947, p. 105) say that their rejection of abstract entities stems from a "basic intuition."

that provides no reason to think that *P* is false (Alston 1958). Here again, the distinction between the razor of silence and the razor of denial is key.

Nominalism as a metaphysical thesis clearly involves denial, not silence, but does it involve *likelihoods*? If Apples is true, perhaps we can take that finding to be an observation statement, and *P* and *N* can be viewed as competing hypotheses. However, *P* and *N* both *entail* Apples, so their likelihoods are the same. This means that the ratio of their posterior probabilities is identical with the ratio of their priors. But what priors should we assign? Nominalists get what they want if *N* has a high prior and *P* a low one; Platonists get what they want with the opposite ordering. I see no basis for defending either assignment.

Even if *N* captures what Apples says, it is widely recognized that much of mathematics cannot be paraphrased in this way. Consider, for example:

(Prime) There are infinitely many prime numbers.

This is not equivalent to a claim about marks on paper or about what human beings can achieve by various numerical calculations. Nor is Prime equivalent to

(Prime*) If numbers exist, then there are infinitely many primes.

Perhaps the best nominalist response to statements like Prime is fictionalism, the thesis that this and other existence claims in mathematics are false though conditional claims like Prime* are true (Field 1989, Balaguer 2001). Fictionalists may seek to bolster their position by pointing out that mathematicians establish conditional results like Prime*, not unconditional statements like Prime. This *is* a defensible epistemic position, but it hardly shows that Prime is false. If we know that Prime* is true, why not be agnostic about the logically stronger Prime? The agnosticism I have in mind is inspired by Carnap (1950). Mathematics assumes a framework of numbers just as physics assumes a framework of physical objects. We *assume* these frameworks; we can't point to non-question begging evidence that these frameworks are true *or* that they are false. Absent such evidence, perhaps we should withhold belief, but also withhold disbelief.

If the parsimony argument for nominalism is hard to justify, the same is true of what has become a standard Platonist reply to it – the indispensability argument of Quine (1953b) and Putnam (1971). Their point is not just that mathematized natural science needs to quantify over numbers. That, after all, allows us to regard the existence of numbers as a useful (indeed, an indispensable) fiction. Rather, the indispensability argument claims that the empirical evidence that confirms a scientific theory also confirms the purely mathematical consequences that the theory has. This argument relies on a form of *epistemological holism*: when a whole theory gets confirmed, so does each of its parts, even the parts that are propositions of pure mathematics.[29] For example, the observations that confirm relativity theory are said to confirm the existence of numbers, since relativity theory entails that numbers exist. This indispensability argument goes wrong in two ways (Sober 1993, 2011b). First, it is guilty of *selective attention*; if the empirical success of relativity theory confirms the existence of numbers, why doesn't the empirical failure of many other mathematized theories disconfirm the existence of numbers? Second, I suspect that the indispensability argument falls into the trap of assuming a faulty principle about confirmation, one that Hempel (1965) called the *special consequence condition*. The special consequence condition says that if *E* confirms *T*, and *T* has the logical consequence *C*, then *E* confirms *C*. It has been known for a long time that the special consequence condition is wrong. Here is a simple example that I've used before to explain why: you are playing poker and wonder whether the card you are about to be dealt will be the Jack of Hearts. The dealer is a bit careless and so you catch a glimpse of the card at the top of the deck before it is dealt to you. You see that it is red. The fact that it is red confirms the hypothesis that the card is the Jack of Hearts, not in the sense of proving that the card will be the Jack of Hearts, but in the Bayesian sense of raising the probability that it will be. The hypothesis that the card will be the Jack of Hearts entails that the card will be a Jack. However, the fact that the card is red does not confirm the hypothesis that the card will be a Jack.

Friends of the indispensability argument have not endorsed the special consequence condition explicitly, but I do think it is implicit in what they say. This means that they need to articulate an epistemology that steers clear of that error and still sustains their argument about mathematics. Hellman (1999) and Colyvan (2001) have taken steps in that direction, but there is much work to be done (Sober 2011b). It is not enough to invoke the Quinean idea that rational belief revision proceeds by "minimum mutilation" – by making the "smallest" change in one's total system of belief that renders the system

[29] This is a distributive holism; holism can also be given a non-distributive formulation (Sober 2000).

logically consistent with what one observes. Smallness needs to be spelled out. Quine says that we ought to abandon less "central" beliefs before we abandon beliefs that are more central. Centrality requires clarification. Also, the idea that logical consistency between observation and theory is all there is to the evidence relation is widely recognized to be inadequate. In any event, the minimum mutilation idea is not enough to justify the indispensability argument.

Why does the nominalist's parsimony argument go so badly wrong? Is it because its subject matter (numbers and universals) is *a priori*, whereas the two parsimony paradigms discussed in Chapter 2 each involve empirical observations? That is not the central issue. Recall the relaxed understanding of what an observation is that I discussed earlier in this chapter in connection with moral realism. Perhaps we can take this one step further. If a physicist can see that an electron is moving across a cloud chamber, can a mathematician see that the number 13 is prime? The word "see" has to do with vision in the first case, but not in the second; blind mathematicians are under no handicap in connection with their apprehension of this mathematical fact. If we nonetheless treat such singular judgments about particular numbers as observations, we can think of their relation to competing generalizations about numbers in something like the format we use to think about the relation of singular observations to generalizations in the empirical sciences. Of course, mathematicians strive to find proofs or disproofs of these generalizations, but before they reach that happy terminus, they form defeasible judgments about the relative plausibility of different generalizations that fit the "observations" they have made of the properties of individual numbers. These plausibility judgments are influenced by background knowledge, just as is true in the empirical sciences. They also might be influenced by the relative simplicity of competing generalizations. If so, it would be worth investigating how the conceptualizations of simplicity that mathematicians use in evaluating the plausibility of mathematical conjectures connect with the ones used in empirical sciences.[30] This is a project for the epistemology of

[30] Mathematicians are well aware that inferences of the form "conjecture G holds for the first several million integers, so it holds for all of them" have failed. A famous example involves the functions $li(x)$ and $\pi(x)$; with the integers arrayed on the x-axis, $li(x)$ is the "logarithmic integral" (an estimate of the number of prime numbers less than a given value) and $\pi(x)$ is the number of primes less than x. It once was thought that, for all x, $li(x) > \pi(x)$, and this had been proved up to $x = 20{,}000{,}000$. Then Littlewood (1914) proved that $\pi(x)$ crosses $li(x)$ and become larger at some point and,

mathematics. The nominalist's parsimony argument raises somewhat different questions.

There is a modest something that remains despite my complaints concerning parsimony arguments against mathematical Platonism. Nominalists deny that numbers exist; their justification for this denial is something I have found wanting. But the razor of silence stands at the ready. Perhaps the right response to Platonism is withholding belief, not committing to disbelief.[31]

Solipsism

To present the problem of solipsism, I will use the first-person, and I invite you to do the same. Solipsism (as it pertains to me) is the thesis that *the only things that exist are my mind and the mental states I have*; for you, solipsism is something different, though you can use the same words to express it. The idea that there is an external world – a world of physical objects that exists above and beyond my mind and my experiences – is regarded by solipsism as a mistake. Solipsism sounds about as crazy as any philosophical thesis could be. The challenge is to see if it can be proven wrong, or if that isn't possible, to show that it is less plausible than the hypothesis of an external world. If that can't be done, maybe I'll need to reconcile myself to thinking that my belief in an external world, though it is natural and even irresistible, has no rational justification. I might be obliged to view the external world hypothesis in the same way that Hume viewed the principle of the uniformity of nature, a topic I discussed in Chapter 1.

Rather than trying to *prove* that solipsism is false, or even show that it is less probable than the alternative, I want to consider something more modest. Maybe there is evidence that discriminates between solipsism and the hypothesis of an external world. G. E. Moore (1939) attempted to prove that solipsism is false and by holding up his hand and asserting "here is a hand." Moore talked of proof rather than evidence, but I want to consider whether my belief that I have a hand is evidence that there is an external world. The suggestion faces a dilemma. If "hand" means something extramental, then it begs the question to cite the existence of a hand as evidence

furthermore, that the two functions cross infinitely many times. Given results like this one, maybe number theorists are more wary of simplicity (qua "uniformity") as a principle of inference than physicists are.

[31] See Huemer (2009) for further discussion of the relevance of parsimony to the mind/body problem and to nominalism versus Platonism.

against solipsism. Alternatively, if "hand" just means a kind of regularity in the flow of my experience, then it is hard to see how the existence of *that* sort of hand is evidence for the external world hypothesis. The challenge is to show how evidence for something outside can come from what is inside.

John Locke (1632–1704) gives four arguments against solipsism in his *Essay Concerning Human Understanding* (1689, Book 4, Chapter 11). The first three strike me as unpersuasive, but the fourth is more interesting:

> Our senses assist one another's testimony of the existence of outward things, and enable us to predict. Our senses in many cases bear witness to the truth of each other's report, concerning the existence of sensible things without us. He that sees a fire, may, if he doubt whether it be anything more than a bare fancy, feel it too; and be convinced, by putting his hand in it.[32]

Locke's argument links with the discussion of common cause and separate cause explanation in Chapter 2. Just as the matching student essays discussed in Chapter 2 bear witness to a common cause (the existence of a file on the Internet from which both students plagiarized), so agreement between my fiery visual and tactile impressions bears witness to something that is external to both – a fire that is out there in the physical world.

Locke's common cause argument for an external world is similar to an argument elaborated by Hans Reichenbach in his 1937 book *Experience and Prediction.* Reichenbach addresses the competition between solipsism and the external world hypothesis by proposing an analogy; see Figure 5.2, which comes from *Experience and Prediction* (p. 117). Reichenbach asks us to imagine

> a world in which the whole of mankind is imprisoned in a huge cube, the walls of which are made of sheets of white cloth, translucent as the screen of a cinema but not permeable by direct light rays. Outside this cube there live birds, the shadows of which are projected on the ceiling of the cube by the sun rays; on account of the translucent character of this screen, the shadow-figures of the birds can be seen by the men within the cube. The birds themselves cannot be seen, and their singing cannot be heard. To introduce the second set of shadow-figures on the vertical plane, we imagine a system of mirrors outside the cube which a friendly ghost has constructed in such a way

[32] The premises for Locke's first three arguments are: (i) people who are blind do not have visual sensations, and the visual experiences of sighted people are not caused by their having eyes; (ii) I have visual experiences that I cannot turn off and on by acts of will; (iii) some experiences are painful whereas the memories I have of them are not.

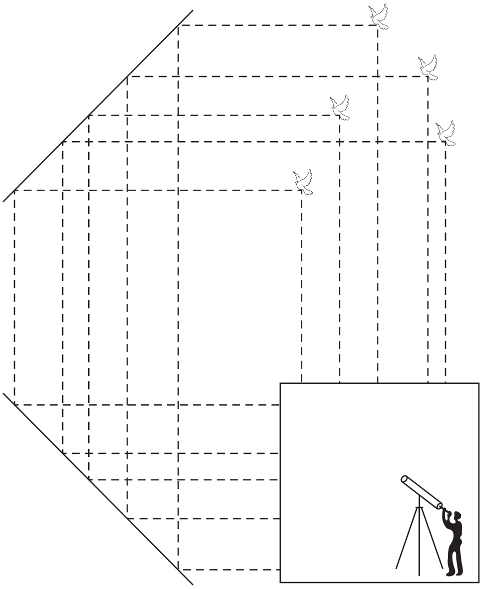


Figure 5.2[33]

> that a second system of light rays running horizontally projects shadow-figures of the birds on one of the vertical walls of the cube [as shown in the accompanying figure] . . . As a genuine ghost this invisible friend of mankind does not betray anything of his construction, or of the world outside the cube, to the people within; he leaves them entirely to their own observations and waits to see whether they will discover the birds outside. He even constructs a system of repulsive forces so that any near approach toward the walls of the cube is impossible for the men; any penetration through the walls, therefore, is excluded, and men are dependent on the observation of the shadows for all statements they make about the "external" world, the world outside the cube. (Reichenbach 1937, pp. 115–116)

Reichenbach (p. 154) takes the epistemological relationship of objects observed (shadows on the surfaces of the cube) to objects merely inferred (the birds outside) to be the same as the relationship of one's sensations to the physical objects that cause them. He comments (pp. 163ff.), correctly in my view, that the analogy is not subverted by the fact that we see physical objects though we rarely, if ever, see our own experiences; light bounces off

[33] This figure is reprinted by permission of University of Chicago Press.

trees, not our inner sensory states. He also is right that there is no need to assume, in addressing the problem of the external world, that our knowledge of our own sensations is absolutely certain; this is inessential, just as the problem of the cubical world does not require that we be absolutely certain about the properties that the shadow-figures have.

After describing the cubical world and its inhabitants, Reichenbach poses a question: "Will these men discover that there are things outside their cube different from the shadow-figures?" He says that initially they do not, but after a while, a gifted individual ("a Copernicus")

> will direct telescopes to the walls and will discover that the dark spots have the shape of animals; and what is more important still, that there are corresponding pairs of black dots, consisting of one dot on the ceiling and one dot on the side wall, which show a very similar shape. If a_1, a dot on the ceiling, is small and shows a short neck, there is a corresponding dot a_2 on the side wall which is also small and shows a short neck; if b_1 on the ceiling shows long legs (like a stork), then b_2 on the side wall shows on most occasions long legs also. It cannot be maintained that there is always a corresponding dot on the other screen but this is generally the case. (pp. 117–118)

What interests Reichenbach about these dots on the walls of the cube is the "correspondence" that obtains between the "internal motions" of pairs:

> If the shade a_1 wags its tail, then the shade a_2 also wags its tail at the same moment. Sometimes there are fights among the shades; then, if a_1 is in a fight with b_1, a_2 is always simultaneously in a fight with b_2. (p. 118)

Reichenbach says that the hero of his story, Copernicus,

> will surprise mankind by the exposition of a very suggestive theory. He will maintain that the strange correspondence between the two shades of one pair cannot be a matter of chance but that these two shades are nothing but effects caused by one individual thing situated outside the cube within free space. He calls these things "birds" and says that these are animals flying outside the cube, different from the shadow-figures, having an existence of their own, and that the black spots are nothing but shadows. (p. 118)

The logic that leads Copernicus to discover that there are birds outside the cube is supposed to be the same as the logic that each of us can use to show that solipsism is false. But exactly how is this analogy supposed to work?

Can the problem posed by the cubical universe be analyzed in terms of the law of likelihood? The goal would be to show that the observed association of the two shadows on the walls of the cube is something we'd expect if they have a common cause, but it would be very surprising ("an improbable coincidence") if the shadows originated independently. The analogy would then lead to the conclusion that the associations I see in my experience favor the hypothesis that an external world exists over the hypothesis of solipsism. Whether or not this likelihood argument is successful, it is not how Reichenbach comes at the problem. What he talks about is the *probability* of there being a common cause outside the cube, and not just about the probability of the observations under that hypothesis (pp. 120–121).

Reichenbach's argument veers away from the law of likelihood when he considers what his adversary, "the positivist," might say about the cubical universe. The positivist grants that the common cause hypothesis and the coincidence hypothesis "furnish different consequences within the domain of our observable facts" (p. 122). But the positivist then claims that there is a third hypothesis that is predictively equivalent to the hypothesis of common cause. This is the hypothesis that the two correlated shadows are related to each other as cause to effect. Reichenbach says that although the common cause and the cause/effect hypothesis both predict that the shadows on the walls of the cube will be correlated, the common cause hypothesis is nonetheless superior. He puts his point in the mouth of "the physicist," who

> simply states that, wherever he observed simultaneous changes in dark spots like these, there was a third body different from the spots; the changes happened, then, in the third body and were projected by light rays to the dark spots which he used to call shadow-figures . . . Whenever there were corresponding shadow-figures like the spots on the screen, there was in addition, a third body with independent existence; it is therefore highly probable that there is also such a third body in the case in question. (p. 123)

Reichenbach's point is that there are associated events that occur *inside* the cube that resemble the shadows on the walls of the cube. These inside events consist of pairs of correlated shadows; the cube's inhabitants *observe* that the inside shadows in a pair are *usually* caused by a common cause (a physical object inside the cube that is casting both shadows); the cubists also see that correlated shadows inside the cube are *rarely* related to each other as cause to effect.

Reichenbach has a footnote on the next page (p. 124) that confirms this interpretation of his argument. He says that the common cause hypothesis has a higher posterior probability than the cause/effect hypothesis because the former has the higher prior and in spite of the fact that the two hypotheses confer the same probability on the observed association. So the reason the observed association of the shadows on the walls of the cube makes it more probable that birds exist outside the cube than that the shadows cause each other is that the existence of birds outside has the higher prior probability.

How does the positivist's third hypothesis about the cubical universe bear on the problem of solipsism? In Locke's example, you observe that fiery visual impressions are associated with fiery tactile impressions and initially you consider two hypotheses – that they have a common cause and that they are independent. These two hypotheses make different predictions. The association is what you'd expect if the common cause hypothesis were true, but the association would be very improbable if the two events were independent. You then consider the positivist's third hypothesis – that one of the fiery impressions causes the other. This too predicts the association, but it does so in a way that is consistent with solipsism. Can Reichenbach appeal to prior probabilities to undermine this third option? This seems dubious, given that sensations often cause other sensations.

In fact, there is something Reichenbach can say to address the positivist's third hypothesis in both problems. Reichenbach can focus his argument on shadows on the walls of the cube that change their trajectories *simultaneously* and on fiery tactile and visual impressions that are also *simultaneous*. If cause must precede effect, we can rule out the explanation that says that one item in a pair causes the other. This takes us back to the two options that Reichenbach initially considered, and, with the assumptions about common cause and separate cause explanations enumerated in Chapter 2, we have a likelihood argument in favor of the former. Does that show that the external world hypothesis is better supported by the experiences I have than solipsism is?

There is a catch. The fact that my experience favors the common cause hypothesis leaves open *where* that common cause is to be found. Locke's association of fiery visual and tactile sensations favors a common cause, but whether that common cause is inside the mind or outside is a further question. That further question can be answered if we assume that the mind is an open book. If I am introspectively aware of *everything* that happens in my mind, then the fact that introspection fails to reveal a common cause shows that if there is one, it must be outside.[34] The past half century has not been kind to the open book assumption; the cognitive revolution has made it abundantly clear that there are processes and representations in the mind to which we have little or no introspective access. For this reason, the likelihood argument against solipsism fails (Sober 2011c).

On the face of it, solipsism seems more parsimonious than the hypothesis that my mind is embedded in an external world of physical objects. On the other hand, it also seems that the external world hypothesis is able to unify what would otherwise be the blooming buzzing confusion of my experience. So maybe, on balance, the external world hypothesis makes for a simpler overall picture of the world. Both these thoughts need to be scrutinized. Maybe the first merely reflects the simple fact that X is a subset of $X + Y$. However, that provides no reason for doubting the existence of Y. Once again we must attend to the difference between the razor of denial and the razor of silence. As for the suggestion that the external world hypothesis is unifying, the challenge is to show why that unifying power is epistemically relevant. Reichenbach rose to this challenge and he was half right. An association that I notice in the flow of my experiences *does* favor a common cause over a separate cause explanation (if the seven assumptions enumerated in Chapter 2 are acceptable). But that is not enough.

My criticism of Reichenbach's argument against solipsism connects with themes from earlier chapters. It is one thing to infer the *existence* of a common cause, something different to infer what *characteristics* that common cause possesses. An instance of this distinction was discussed in Chapter 3: inferring common ancestry and inferring the character state of a common ancestor are distinct problems. The first appearance of this distinction was in Chapter 2, where I derived the greater likelihood of a common cause over a separate cause explanation from Reichenbachian assumptions. To deduce the likelihood inequality from those assumptions, it is essential that the competing hypotheses *not* specify the states of the causes they postulate.

[34] Locke comes close to asserting the open-book thesis when he writes that "consciousness . . . is inseparable from thinking, and as it seems to me essential to it: it being impossible for anyone to perceive, without perceiving, that he does perceive. When we see, hear, smell, taste, feel, meditate, or will any thing, we know that we do so . . . consciousness always accompanies thinking" (II.27.ix).

Given the failure of the Locke/Reichenbach argument, what becomes of solipsism? Instead of seeking other empirical arguments for the external world hypothesis, we should cast our net more widely. Perhaps the hypothesis of an external world is a framework assumption that we adopt for pragmatic reasons, not because we have any evidence that it is true. I mentioned in the previous section that mathematicians do not *prove* that numbers exist; rather, they *assume* that numbers exist and then prove theorems about them. Maybe the same point applies to physicists; they *assume* that there is an external physical world and then develop theories about its contents. Theorems in mathematics and theories in physics are subject to test, but the tests never discharge the framework assumptions that are used to formulate them. This is Carnap's (1950) position. He argued that the question of whether numbers exist and the question of whether physical objects exist are *external* questions; they differ epistemologically from internal questions such as "Are there prime numbers greater than 1,000,000?" and "Do electrons exist?" The failure of Reichenbach's common cause argument for an external world does not prove that Carnap was right, but the Carnapian alternative is worth considering.[35]

The problem of induction

In Chapter 1, I discussed Hume's claim that all inductive inferences rest on the principle of the uniformity of nature (PUN) – the assumption that the future will resemble the past. Kant agrees with Hume. Kant thinks that this assumption is mandated by reason, but Hume thinks that it is an assumption that cannot be justified at all. Despite this difference, the two philosophers agree that PUN is presupposed by every inductive argument that draws a conclusion about the future from the observations we have made about the past. For both Hume and Kant, this simplicity postulate is central to the way we construct and evaluate our beliefs about the world.

Most objections to Hume's skeptical thesis have attempted to show that PUN can be justified. I think these attempted justifications have all failed and that no such justification is possible. However, I think that Hume's thesis that all inductions presuppose PUN is wrong. Even so, Hume was right about something important.

[35] To appreciate Carnap's position, you need to understand why the special consequence condition is mistaken; I discussed this earlier in the present chapter.

To make the problem concrete, let's use one of Hume's examples. Human beings have observed over many centuries that the Sun has risen each day. It seems altogether reasonable to infer from this data that the Sun will rise tomorrow. Hume says that this inference presupposes PUN — that the future will resemble the past. But what does PUN actually say? Here are two interpretations:

(Strong)	The future will resemble the past in all respects.
(Weak)	The future will resemble the past in some respects.

As the labels suggest, Strong is too strong and Weak is too weak. Surely we don't actually believe Strong; rather, we are confident that the future will differ from the past in some respect or other. What is more, Strong can't be right, since it involves a contradiction.[36] On the other hand, Weak isn't substantial enough to tell us whether we should think that the Sun will rise tomorrow (Salmon 1967, pp. 42–43, p. 52). It is worth noting that this jaundiced view of PUN also applies to the suggestion that Ockham's razor depends on the assumption that "nature is simple."

Although Hume gives PUN a centrality that it does not deserve, he was right that if the belief that the Sun will rise tomorrow is justified, the justification must include elements that aren't part of what we have observed to date. If you are a Bayesian, Hume's point should be obvious. If you want to show that the Pr(the Sun will rise tomorrow | we have observed for centuries that the Sun has risen each day) is high, you need to be able to say something about the value of the prior probability Pr(the Sun will rise tomorrow). The value of this prior isn't something you have observed. Notice, however, that what you need here is a prior probability for a particular proposition about the Sun, not some sweeping principle about nature in general.

You don't need to be a Bayesian to buy Hume's point. To see why, let's reformulate Hume's thesis so that it isn't about the probability of hypotheses. Suppose you toss a coin 1,000 times and obtain 500 heads. A frequentist may advise you to use maximum likelihood estimation; if you do, your

[36] To use Goodman's (1955) famous example, if past and future emeralds are the same with respect to color, they can't be the same with respect to gruler. Just as green and blue are colors, grue and bleen are grulers. An object *x* at time *t* is grue precisely when *x* is green at *t* and *t* is before the year 2050 or *x* is blue at *t* and *t* is not before the year 2050. Bleen can be defined in tandem. These aren't Goodman's definitions, but they serve to make the point.

estimate is that the coin's probability of landing heads is $p = 0.5$. Frequentists who give this advice will attempt to justify it by noting that the method of maximum likelihood estimation has various desirable operating characteristics – for example, maximum likelihood estimation is *statistically consistent* (a topic discussed in Chapter 3). This means that if you use this method again and again, on bigger and bigger samples, the probability that your estimate will be as close as you like to the true value converges on one as your sample size approaches infinity. The pertinent Humean point about this claim concerning statistical consistency is that the claim's justification depends on assumptions that go beyond the observations you have made to date. You are assuming that the coin is an *i.i.d.* system – that past, present, and *future* tosses are independent of each other and there is a single probability of heads that applies to all past, present, and *future* tosses.

What is right in Hume's skeptical argument about induction is that each inductive inference from past observations to a hypothesis about the future depends on a third element – a background assumption.[37] However, it does not follow that there is a single background assumption that all such inductive inferences must display. The mistake of inferring the latter from the former is an instance of *the birthday fallacy* (Sober 1988). This is the mistake of thinking that the following argument is valid:

Everyone has a birthday.
———————————
There is a single day on which every person was born.

Every inductive inference involves assumptions additional to the observational premises, but that doesn't mean that every inductive inference makes the same assumption – for example, that PUN is true.[38]

Most of the topics discussed in this chapter involve a philosophical problem that someone has sought to solve by appeal to a principle of simplicity.

[37] Understood in this way, Hume's discovery complements a point that Duhem (1914) made, that physical theories do not, by themselves, entail observations, but do so only when conjoined with auxiliary assumptions (Sober 2008b).

[38] Wittgenstein (1921) connects the problem of induction to the concept of simplicity in the *Tractatus* (6.363-6.3631): "The process of induction is the process of assuming the *simplest* law that can be reconciled with our experiences. This process, however, has no logical foundation, but only a psychological one. It is clear that there are no grounds for believing that the simplest eventuality will in fact be realized."

The problem of induction is not like this. Rather, simplicity is part of the traditional *setting* of the problem, not its solution. You are told that a particular simplicity postulate, PUN, is an assumption in every inductive inference; your challenge is to say what justification PUN might have. The problem, thus formulated, has a false presupposition.

Concluding comments

Philosophers who embrace naturalism$_p$ often feel entitled to use the principle of parsimony to evaluate philosophical theories because scientists use the same principle to evaluate scientific theories. The main point of this chapter is that this line of thought needs to be treated with caution. Naturalists need to consider whether parsimony arguments in philosophy really do have the same structure as the good parsimony arguments that are advanced in science. The fact that the word "parsimony" gets used in both science and philosophy is not enough.

Philosophical parsimony arguments sometimes conform to the likelihood paradigm, sometimes to the model selection paradigm, and sometimes to neither. The "success stories" described in this chapter concern the problem of evil and the mind/body problem. The evidential argument from evil concludes that some of the evils we observe favor atheism over the hypothesis that an all-PKG God exists; here the more parsimonious hypothesis is the one with the higher likelihood. In the mind/body problem, the identity theory is more parsimonious than dualism, and model selection theory helps explain why this difference in parsimony is epistemically relevant. That's the good news. The bad news is that other parsimony arguments – arguments that oppose mental causation, moral realism, mathematical Platonism, and solipsism – are flawed. Model selection ideas explain when and why a one-cause model is better than a two-cause model, but this account applies only when each candidate cause can be varied while holding the other fixed. This cannot be done if one of the putative causes supervenes on the other, which is why the parsimony argument against mental causation goes down in flames. The parsimony argument against moral realism also goes wrong. The argument says that postulating an independent realm of normative ethical truths is justified only if that postulate is needed to explain descriptive propositions about human thought and action. The problem is that ethics and psychology are in different lines of work. As for the parsimony argument against

Platonistic entities, it is true that the sentence "there are two apples in the basket" can be paraphrased without saying that numbers exist, and it also is true that mathematicians *assume* that numbers exist but never *prove* that they do. However, neither of these truths constitutes a good reason to deny that numbers (as understood by Platonists) exist. It is the razor of silence, not the razor of denial, that can be put to work here. Finally, the problem of solipsism has important connections with a likelihood comparison of common cause and separate cause explanations; correlated experiences often are evidence for a common cause, but it is a further step to conclude that that common cause is outside the mind. These unsuccessful parsimony arguments on philosophical topics are not accommodated by the two parsimony paradigms I described in Chapter 2, but that, by itself, does not show that these arguments are misguided. Perhaps we should reject naturalism$_P$. Or perhaps we need to delve deeper into science to uncover new and convincing parsimony paradigms. Maybe so, but my guess is that there is no rescuing these parsimony arguments. This is because I think there are flaws in these arguments that can be identified without needing to understand much about Ockham's razor.

The two success stories I have recounted in this chapter involve appeals to observations. In the mind/body problem, we observe that there is an association between pain reports and c-fiber brain scans; in the problem of evil, we observe that there are events that have lots of bad-making properties, and careful deliberation fails to reveal any good-making properties that could remotely compensate for those bad-makers. These examples suggest that the two parsimony paradigms do not apply to philosophical arguments in which observations play no role. Perhaps, but just as I adopted a "relaxed" attitude to the concept of observation when I suggested that some singular moral judgments are observations and that some mathematical judgments are too, I think it is worth considering the possibility that what philosophers call "intuitions" may sometimes have the same epistemological status that narrow-sense observations have. If intuitions can be treated in this way, there may be applications of the two parsimony paradigms to philosophical problems where empirical observations (in a narrow sense) do not matter.

In addition to considering several parsimony arguments in philosophy, I have addressed a few other issues in this chapter. Scientists like to say that absence of evidence isn't evidence of absence, but this is an overstatement. What is true is that absence of evidence *often* fails to be evidence of absence.

This is worth bearing in mind when philosophers invoke the razor of denial. Another topic was a screening-off criterion for explanatory irrelevance that philosophers often implicitly use; this criterion cannot withstand scrutiny. Finally, the problem of induction is standardly formulated by claiming that all inductive inferences presuppose a single global simplicity postulate. This is a mistake.

I want to conclude by considering the remark of Wittgenstein's that I quoted at the start of this chapter. Wittgenstein sees the philosopher's passion for parsimony as misguided science worship. Why does Wittgenstein think that Ockham's razor has no role to play in philosophy? Right after the sentences I quoted, he says that "it can never be our job to reduce anything to anything, or to explain anything. Philosophy really *is* 'purely descriptive.'" The phrase he puts in scare quotes is the key. If philosophy is to be purely descriptive, what does that rule out? Wittgenstein's point is not that philosophy should never be *normative*. Wittgenstein, both early and late, is happy to rail against philosophical error. His way of doing philosophy is normative with a capital *N*. What Wittgenstein opposes in this passage is philosophical *theorizing*. He thinks that philosophical puzzlement arises from the "bewitchment" of language; once the relevant linguistic facts are described, the puzzles dissolve. Philosophers should act as therapists, not as theorists. Wittgenstein's antipathy to philosophical theorizing goes beyond his complaint that philosophers are guilty of over-simplifying and over-generalizing. That complaint might simply indicate the need for *better* theorizing, not that philosophers should be T-totaling abstainers.

Wittgenstein's assessment of how philosophical puzzlement arises is doubly peculiar. First, he sees its source as exclusively linguistic. Philosophical questions are never really about the way the world is; their source is always to be found in our misunderstanding the language we speak. Second, Wittgenstein says that the way to address philosophical puzzles is by describing, not theorizing. This second suggestion is not entailed by the first. Even if philosophical puzzlement arises from our failure to understand our language, why is theorizing about language always useless? Perhaps our misunderstandings are sometimes due to having a false theory of how language works; one remedy for a bad theory is a better one.

In this book, I have not doubted that philosophical theorizing has a legitimate role to play. The separate sciences have their separate subject matters. Physicists study particles and geneticists study genes, but who is studying the difference between good scientific inference and bad? That task has fallen to statisticians and philosophers of science. It is not an abuse of language to say that their joint enterprise is the elaboration and evaluation of theories of scientific inference. The clash between Bayesianism and frequentism described in Chapter 2 is a clash between rival *theories.*

Theories of scientific inference identify patterns of reasoning that apply across very different scientific subject matters. These patterns of reasoning also apply to problems that are of specifically philosophical interest. For example, consider a point of contact between oncology and philosophy of mind. Oncologists are interested in determining whether smoking cigarettes and inhaling asbestos both cause lung cancer. Philosophers of mind are interested in determining whether mental properties and neurophysiological properties both cause behavior. The principles for testing causal hypotheses apply to both problems. Philosophy is not walled off from those principles any more than oncology is. Notice that the applicability of these principles to both subjects does not mean that the two problems are in every respect the same. I have argued that they are not.

It is a familiar idea in science that a theory can be guilty of over-simplifying. One need only think of the role that simplicity plays in model selection theory. Models gain points by being parsimonious, but they also gain points by fitting the data. The typical situation is that these *desiderata* conflict – more complex theories often fit the data better than their simpler competitors. It is good scientific practice to balance these considerations against each other. Models that oversimplify and thereby fit the data very poorly have gone too far. Wittgenstein's admonition that philosophers should avoid oversimplification is well taken. His claim that philosophy is "purely descriptive" is not. The suggestion that philosophers should never seek to explain, but should stick to describing, is reminiscent of the skeptic's advice that you should never venture beyond describing the observations at hand. As William James (1897) noted, the benefit of skepticism is that you will never have a false belief; the cost is that you will miss out on embracing true ones.

Wittgenstein says that "the real source of metaphysics" (here "metaphysics" is a pejorative term) is the philosopher's penchant for parsimony. Wittgenstein's use of the definite article is significant; it indicates that he has failed to follow his own good advice about the dangers of oversimplification. There are many ways to fall into philosophical error, not just one. Oversimplifying is one pitfall, but there are others. For example, there is the mistake of endorsing distorted images of science and philosophy and concluding, on that basis, that philosophy and science are miles apart.

Wittgenstein's rejection of philosophical theorizing is radical. For those who reject Wittgenstein's rejection, there is work to be done. Here are some philosophical problems, additional to the ones I have reviewed in this chapter, and the philosophers who have appealed to parsimony in their efforts to solve them:

- Can there be time without change (Shoemaker 1969)?
- Do species have essences (Sober 1980)?
- Should a linguistic regularity be explained by a semantic or a pragmatic principle (Grice 1989)?
- What is wrong with skeptical hypotheses about brains in vats (Vogel 1990)?
- Are the values of the physical constants evidence for the existence of God (Swinburne 2009)?
- Do composite objects exist (Sider 2013)?

Do these philosophical parsimony arguments hold water and what do they tell us about naturalism$_P$?

References

In what follows, I usually cite older books by giving their dates of first publication and then mentioning a more recent edition. The pages cited in references to these books in the preceding chapters are from the more recent edition.

Ackermann, R. (1963) "A Neglected Proposal Concerning Simplicity." *Philosophy of Science* 30: 228–235.

Adams, M. M. (1987) *William Ockham*. Notre Dame, IN: University of Notre Dame Press, 2nd edition, 1989.

Akaike, H. (1973) "Information Theory as an Extension of the Maximum Likelihood Principle." In B. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267–281.

Almeida, M. and Oppie, G. (2003) "Skeptical Theism and Evidential Arguments from Evil." *Australian Journal of Philosophy* 81: 496–516.

Alston, W. (1958) "Ontological Commitments." *Philosophical Studies* 9: 8–17.

Andrews, K. (2005) "Chimpanzee Theory of Mind: Looking in All the Wrong Places?" *Mind and Language* 20(5): 521–536.

Aquinas, T. (1945) *Basic Writings of St. Thomas Aquinas*, A. C. Pegis (trans.), New York: Random House.

Aristotle (1995) *The Complete Works of Aristotle: The Revised Oxford Translation* (Vol. 1). J. Barnes (ed.). Princeton University Press.

Arnheim, R. (1954) *Art and Visual Perception*. Berkeley: University of California Press.

Arntzenius, F. (2010) "Reichenbach's Common Cause Principle." *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/fall2010/entries/physics-Rpcc.

Atkinson, Q. and Gray, R. (2005) "Curious Parallels and Curious Connections – Phylogenetic Thinking in Biology and Historical Linguistics." *Systematic Biology* 54: 513–426.

Baker, A. (2003) "Quantitative Parsimony and Explanatory Power." *British Journal for the Philosophy of Science* 54: 245–259.

Balaguer, M. (2001) *Platonism and Anti-Platonism in Mathematics*. Oxford University Press.

Barbrook, A., Howe, C., Blake, N., and Robinson, P. (1998) "The Phylogeny of the Canterbury Tales." *Nature* 394: 839.

Barnes, E. (2000) "Ockham's Razor and the Anti-Superfluity Principle." *Erkenntnis* 53: 353–374.

Bayes, T. (1764) "An Essay Toward Solving a Problem in the Doctrine of Chances." *Philosophical Transactions of the Royal Society of London* 53: 370–418.

Bennett, K. (2003) "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, To Tract It." *Nous* 37: 471–497.

Berger, J. (1985) *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer-Verlag.

Black, M. (1954) *Problems of Analysis*. Ithaca, NY: Cornell University Press.

Blumenfeld, D. (1995) "Perfection and Happiness in the Best Possible World." In N. Jolley (ed.), *The Cambridge Companion to Leibniz*. Cambridge University Press, pp. 382–410.

Box, G. and Tao, G. (1973) *Bayesian Inference and Statistical Analysis*. New York: Wiley, 1992.

Boyd, R. (1990) "Observations, Explanatory Power, and Simplicity – Towards a non-Humean Account." R. Boyd, P. Gasper, and J. D. Trout (eds.), *The Philosophy of Science*. Cambridge, MA: MIT Press, pp. 349–378.

Brandt, R. and Kim, J. (1967) "The Logic of the Identity Theory." *Journal of Philosophy* 64: 515–537.

Buchdahl, G. (1969) *Metaphysics and the Philosophy of Science*. Oxford: Basil Blackwell.

Buckner, C. (2013) "Morgan's Canon, Meet Hume's Dictum – Avoiding Anthropofabulation in Cross-species Comparison." *Biology and Philosophy* 28(5): 853–871.

Burks, A. (1953) "The Presupposition Theory of Induction." *Philosophy of Science* 20: 177–197.

Burnham, K. and Anderson, D. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). New York, NY: Springer-Verlag.

Camin, J. H. and Sokal, R. R. (1965) "A Method for Deducing Branching Sequences in Phylogeny." *Evolution* 19(3): 311–326.

Carnap, R. (1950) "Empiricism, Semantics, and Ontology." *Revue Internationale de Philosophie* 4: 20–40, enlarged edition, 1956.

Cavalli-Sfrorza, L. and Edwards, A. (1967) "Phylogenetic Analysis: Models and Estimation Procedures." *Evolution* 32: 550–570.

Cheney, D. and Seyfarth, R. (1990) *How Monkeys See the World*. University of Chicago Press.

Cherkassky, V. (2013) *Predictive Learning*. Self-published.

Chomsky, N. (1959) "Review of B. F. Skinner's Verbal Behavior." *Language* 35: 26–58.

Churchland, P. S. (1987) "Epistemology in the Age of Neuroscience." *The Journal of Philosophy* 84(10): 544–553.

Clatterbuck, H. (forthcoming) "The Value of Parsimonious Mental Models — Evidence of Chimpanzee Mindreading." *Mind and Language*.

(2015) "Are Humans the Only Theorizers? A Philosophical Examination of the Theory-Theory of Human Uniqueness." Unpublished PhD dissertation, University of Wisconsin, Madison.

Colyvan, M. (2001) *The Indispensability of Mathematics*. New York, NY: Oxford University Press.

Copernicus, N. (1543) *On the Revolutions of the Heavenly Spheres*, A. M. Duncan (trans.). New York, NY: Barnes and Noble, 1976.

Cummins, R. (1975) "Functional Analysis." *Journal of Philosophy* 72: 741–764.

Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray (Facsimile: Cambridge, MA: Harvard University Press, 1964).

(1862) *On the Various Contrivances by which British and Foreign Orchids are Fertilised by Insects, and on the Good Effects of Intercrossing*. London: John Murray.

(1868) *The Variation of Animals and Plants under Domestication*. London: John Murray.

(1871) *The Descent of Man, and Selection in Relation to Sex*. (Facsimile: Princeton University Press, 1981).

(1872a) *Expression of the Emotions in Man and Animals*. London: John Murray.

(1872b) "Letter to A. Hyatt." Darwin Correspondence Database, www.darwinproject.ac.uk/entry-8658.

(1958) *Autobiography*. N. Barlow (ed.). London: Collins.

(1959) *The Origin of Species – a Variorum Edition*. M. Peckham (ed.). Philadelphia: University of Pennsylvania Press.

Dennett, D. C. (1980) "True Believers – the Intentional Stance and Why it Works." *Herbert Spencer Lecture on Scientific Explanation*. Oxford University Press. Expanded version in D. C. Dennett (1989) *The Intentional Stance*. Cambridge, MA: MIT Press, pp. 13–42.

(1981) "Skinner Skinned." In *Brainstorms*. Cambridge, MA: MIT Press, pp. 53–70.

Descartes, R. (1633) *The World, a Treatise on Light*. M. Mahoney (trans.). New York, NY: Abaris Books, 1979.

(1637) *Optics*. In P. Olscamp (ed.), *Discourse on Method, Optics, Geometry, and Meteorology*. Indianapolis: Bobbs-Merrill Publishing, 1965.

(1641) *Meditations on First Philosophy*. J. Cottingham (trans.). Cambridge University Press, 1996.

(1644) *Principles of Philosophy*. In J. Cottingham, R. Stoothoff, and D. Murdoch (eds. and trans.), *The Philosophical Writings of Descartes* (Vol. 1). Cambridge University Press, 1985.

(1662) *Treatise on Man*. In S. Graukroger (ed. and trans.), *The World and Other Essays*. Cambridge University Press, 2004.

de Waal, F. B. M. (1991) “Complementary Methods and Convergent Evidence in the Study of Primate Social Cognition.” *Behaviour* 118: 297–320.

(1999) “Anthropomorphism and Anthropodenial: Consistency in our Thinking about Humans and Other Animals.” *Philosophical Topics* 27: 255–280.

(2009) “Darwin’s Last Laugh.” *Nature* 460: 175.

Dijksterhuis, E. (1961) *The Mechanization of the World Picture*. Oxford University Press.

Doolittle, W. F. (2000) “Uprooting the Tree of Life.” *Scientific American* 282(6): 90–95.

Draper, P. (1989) “Pain and Pleasure – An Evidential Problem for Theists.” *Noûs* 23: 331–350. Reprinted in D. Howard-Snyder (ed.), *The Evidential Argument from Evil*. Bloomington, IN: Indiana University Press, 1996, pp. 12–29.

Duhem, P. (1914) *The Aim and Structure of Physical Theory*. Princeton University Press, 1954.

Duns Scotus, J. (1998) *Questions on the Metaphysics of Aristotle* (Text Series, Number 19, Volume 2). St. Bonaventure, NY: Franciscan Institute Publishers.

Edwards, A. (1972) *Likelihood*. Cambridge University Press.

(2007) “Maximisation Principles in Evolutionary Biology.” In M. Matthen, C. Stephens, D. Gabbay, P. Thagard, and J. Woods (eds.), *Handbook of the Philosophy of Science, Volume 3: Philosophy of Biology*. Amsterdam: North Holland-Elsevier, pp. 335–348.

Edwards, A. and Cavalli-Sforza, L. (1963) “The Reconstruction of Evolution.” *Ann. of Human Genetics* 27: 105.

Einstein, A. (1933) “On the Method of Theoretical Physics.” *Herbert Spencer Lecture*. Oxford University Press.

Eldredge, N. and Cracraft, J. (1980) *Phylogenetic Patterns and the Evolutionary Process*. New York, NY: Columbia University Press.

Farris, J. S. (1983) “The Logical Basis of Phylogenetic Analysis.” In N. Platnick and V. Funk (eds.), *Advances in Cladistics – Proceedings of the 2nd Annual Meeting of the Willi Hennig Society*. New York, NY: Columbia University Press, pp. 7–36. Reprinted in E. Sober (ed.), *Conceptual Issues in Evolutionary Biology*, Cambridge: MIT Press, 1994, pp. 333–362.

Felsenstein, J. (1973) “Maximum Likelihood and Minimum-Step Methods for Estimating Evolutionary Trees from Data on Discrete Characters.” *Systematic Zoology* 22: 240–249.

(1978) “Cases in which Parsimony and Compatibility Methods can be Positively Misleading.” *Sytematic Biology* 27: 401–410.

Field, H. (1989) *Realism, Mathematics and Modality*. Oxford: Blackwell.

Fitelson, B. (1999) “The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity.” *Philosophy of Science* 66: S362–78.

(2011) “Favoring, Likelihoodism, and Bayesianism.” *Philosophy and Phenomenological Research* 83: 666–672.

Fitelson, B. and Sober, E. (1998) "Plantinga's Probability Arguments against Evolutionary Naturalism." *Pacific Philosophical Quarterly* 79: 115–129.

Fitzpatrick, S. (2006) *Simplicity, Science, and Mind*. Unpublished doctoral dissertation, Department of Philosophy, University of Sheffield.

(2009) "The Primate Mindreading Controversy: A Case Study in Simplicity and Methodology in Animal Psychology." In R. Lurz (ed.), *The Philosophy of Animal Minds*. Cambridge University Press, pp. 258–277.

(2013) "Kelly on Ockham's Razor and Truth-Finding Efficiency." *Philosophy of Science* 80: 298–309.

Flombaum, J. and Santos, L. (2005) "Rhesus Monkeys Attribute Perceptions to Others." *Current Biology* 15: 447–452.

Forster, M. (1988) "Unification, Explanation, and the Composition of Causes in Newtonian Mechanics." *Studies in History and Philosophy of Science* 19(1): 55–101.

(2000) "Key Concepts in Model Selection." *Journal of Mathematical Psychology* 44: 205–231.

Forster, M. and Sober, E. (1994) "How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions." *British Journal for the Philosophy of Science* 45: 1–36.

Freeland, S., Knight, R., Landweber, L., and Hurst, L. (2000) "Early Fixation of an Optimal Genetic Code." *Molecular Biology and Evolution* 17: 511–518.

Friedman, M. (1992) "Causal Laws and Foundations of Natural Science." In P. Guyer (ed.), *The Cambridge Companion to Kant*. Cambridge University Press, pp. 161–199.

Froidmont, L. (1649) *Philosophia Christiana de Anima*. Louvain: H. Nempaei.

Gaffney, E. (1979) *An Introduction to the Logic of Phylogenetic Reconstruction*. In J. Cracraft and N. Eldredge (eds.) *Phylogenetic Analysis and Paleontology*. New York: Columbia University Press, pp. 79–112.

Galilei, G. (1632) *Dialogue Concerning the Two Chief World Systems*. S. Drake (trans.). Berkeley, CA: University of California Press, 1953. 2nd revised edition 1967.

Garber, D. (1992) *Descartes' Metaphysical Physics*. University of Chicago Press.

Gascuel, O. and Steel, M. (2010) "Inferring Ancestral Sequences in Taxon-rich Phylogenies." *Mathematical Biosciences* 227: 125–135.

Gaut, B. and Lewis, P. (1995) "Success of Maximum Likelihood Phylogeny Inference in the Four Taxon Case." *Molecular Biology and Evolution* 12: 152–162.

Gayon, J. (1998) *Darwinism's Struggle for Survival: Heredity and the Hypothesis of Natural Selection*. Cambridge University Press.

Ghiselin, M. (1974) "A Radical Solution to the Species Problem." *Systematic Zoology* 23:536–544.

Goldstein, B. and Hon, G. (2005) "Kepler's Move from Orbs to Orbits – Documenting a Revolutionary Scientific Concept." *Perspectives on Science* 13: 74–111.

Goodman, N. (1955) *Fact, Fiction, and Forecast*. Indianapolis: Bobbs-Merrill, 2nd edition, 1965.

Goodman, N. and Quine, W. (1947) "Steps Toward a Constructive Nominalism." *Journal of Symbolic Logic* 12: 105–122.

Gopnik, A. and Meltzoff, A. (1998) *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gottlieb, P. and Sober, E. (forthcoming) "Aristotle on 'Nature Does Nothing in Vain'."

Gould, S. (1985) "Adam's Navel." In *The Flamingo's Smile – Reflections in Natural History*. New York: Norton, pp. 99–113.

Gould, S. and Lewontin, R. (1979) "The Spandrels of San Marco and the Panglossian Paradigm – A Critique of the Adaptationist Programme." *Proceedings of the Royal Society B* 205: 581–598.

Gray, R. and Jordan, F. (2000) "Language Trees Support the Express-train Sequence of Austronesian Expansion." *Nature* 405: 1052–1055.

Grice, P. (1989) *Studies in the Way of Words*. Cambridge: Harvard University Press.

Grünwald, P. (2007) *The Minimum Description Length Principle*. Cambridge, MA: MIT Press.

Guyer, P. (2006) *Kant*. London: Taylor and Francis.

Hacking, I. (1965) *The Logic of Statistical Inference*. Cambridge University Press.

Hájek, A. (2003) "What Conditional Probabilities Could Not Be." *Synthese* 137: 273–323.

Hamilton, W. (1852) "On Causality." In *Discussions in Philosophy, Literature and Education*. London: Longman, Brown, Green, and Longmans.

Hare, B., Call, J., Agnetta, B., and Tomasello, M. (2000) "Chimpanzees Know what Conspecifics Do and Do Not See." *Animal Behaviour* 59: 771–785.

Hare, B., Call, J. and Tomasello, M. (2001) "Do Chimpanzees Know what Conspecifics Know?" *Animal Behaviour* 61: 139–151.

Harman, G. (1965) "The Inference to the Best Explanation." *Philosophical Review* 74: 88–95.

(1977) *The Nature of Morality*. Oxford University Press.

Hayes W. (2007) "Is the Outer Solar System Chaotic?" *Nature Physics* 3(10): 689–691.

Hellman, G. (1999) "Some Ins and Outs of Indispensability — A Modal-Structural Approach." In A. Cantini, E. Casad, and P. Minad (eds.), *Logic and Foundations of Mathematics*, Dordrecht: Kluwer, pp. 25–39.

Hempel, C. (1965) "Studies in the Logic of Confirmation." In *Aspects of Scientific Explanation*, New York, NY: Free Press, pp. 3–51.

Henderson, L., Goodman, N, Tenenbaum, J., and Woodward, J. (2010) "The Structure and Dynamics of Scientific Theories – A Bayesian Perspective." *Philosophy of Science* 77: 172–200.

Hennig, W. (1966) *Phylogenetic Systematics*. Urbana, IL: University of Illinois Press.

Heyes, C. M. (1994) "Cues, Convergence, and a Curmudgeon: a Reply to Povinelli." *Animal Behaviour* 48: 242–244.

(1998) "Theory of Mind in Non-Human Primates." *Behavioral and Brain Sciences* 21: 101–148.

(2015) "Animal Mindreading – What's the Problem?" *Psychonomic Bulletin and Review* 22: 313–327.

Hinde, R. A. (1970) *Animal Behavior: A Synthesis of Ethology and Comparative Psychology*. New York, NY: McGraw Hill.

Hitchcock, C. and Sober, E. (2004) "Prediction versus Accommodation and the Risk of Overfitting." *British Journal for the Philosophy of Science* 55:1–34.

Holden, C. (2002) "Bantu Language Trees Reflect the Spread of Farming across Sub-Sahara Africa – a Maximum Parsimony Analysis." *Proceedings of the Royal Society of London (Series B)* 269: 793–799.

Hotopp, J. (2011) "Horizontal Gene Transfer between Bacteria and Animals." *Trends in Genetics* 27: 157–163.

Howard-Snyder, D. (1996) *The Evidential Argument from Evil*. Bloomington, IN: Indiana University Press.

Howson, C. (1988) "On the Consistency of Jeffreys' Simplicity Postulate and its Role in Bayesian Inference." *Philosophical Quarterly* 38: 68–83.

(2001) "The Logic of Bayesian Probability." In D. Corfield and J. Williamson (eds.), *Foundations of Bayesianism*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Hübener, W. (1983) "Occam's Razor not Mysterious." *Archiv für Begriffsgeschichte*. 27: 73–92.

Huemer, M. (2009) "When is Parsimony a Virtue?" *Philosophical Quarterly* 59: 216–236.

Hull, D. (1978) "A Matter of Individuality." *Philosophy of Science* 45: 335–360.

Hume, D. (1739–1740) *A Treatise of Human Nature*. D. F. Norton and M. J. Norton (eds.). New York, NY: Oxford University Press, 2000.

(1748) *Enquiry Concerning Human Understanding*. New York, NY: Pearson, 1995.

(1779) *Dialogues Concerning Natural Religion*. N. K. Smith (ed.). Oxford University Press, 1935.

James, W. (1897) "The Will to Believe." In *The Will to Believe and Other Essays in Popular Philosophy*. New York, NY: Longmans Green & Co, pp. 1–31.

(1907) "The Present Dilemma in Philosophy." In *Pragmatism*. Cambridge, MA: Harvard University Press, 1979.

Janssen, M. (2002) "COI Stories: Explanation and Evidence in the History of Science." *Perspectives on Science* 10: 457–522.

Jefferys, W. and Berger, J. (1992a) "Ockham's Razor and Bayesian Analysis." *American Scientist* 80: 64–72.

(1992b) "Reply to Sober and Forster." *American Scientist* 80: 213–214.

Jeffrey, R. (1965) *The Logic of Decision*. University of Chicago Press, 2nd edition, 1983.

Jeffreys, H. (1931) *Scientific Inference*. London: Macmillan, 2nd edition 1957.

(1939) *A Theory of Probability*. Oxford: Clarendon Press, 3rd edition, 1961.

Jones, R. (1973) "James Clerk Maxwell at Aberdeen, 1856–1860." *Notes and Records of the Royal Society of London* 28: 57–81.

Jordan, J. (2006) "Does Skeptical Theism Lead to Moral Skepticism?" *Philosophy and Phenomenological Research* 72: 403–417.

Joyce, R. (2006) *The Evolution of Morality*. Cambridge, MA: MIT Press.

Kahneman, D. and Tversky, A. (1985). "Evidential Impact of Base Rates." In D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, pp. 153–160.

Kant, I. (1787) *The Critique of Pure Reason*. P. Guyer and A. Wood (trans.). Cambridge University Press, 1998.

(1790) *Critique of the Power of Judgment*. P. Guyer (trans.). Cambridge University Press, 2000.

Karin-D'Arcy, M. R. (2005) "The Modern Role of Morgan's Canon in Comparative Psychology." *International Journal of Comparative Psychology* 18:179–201.

Kaye, S. (2007) "William of Ockham." J. Fieser and B. Dowden (eds.). *The Internet Encyclopedia of Philosophy*, www.iep.utm.edu/ockham.

Kelly, K. T. (2007). "A New Solution to the Puzzle of Simplicity." *Philosophy of Science* 74(5): 561–573.

Kepler, J. (1596) *Mysterium Cosmographicum*. A. M. Duncan (trans.). New York, NY: Abaris Books, 1981.

Kim, J. (1993) *Supervenience and Mind*. Cambridge University Press.

(1996) *Philosophy of Mind*. Boulder, CO: Westview Press.

Kishino, H. and Hasegawa, M. (1990) "Converting Distance to Time: Application to Human Evolution." *Methods in Enzymology* 183: 550–570.

Kleisner, K., Ivell, R., and Flegr, J. (2010) "The Evolutionary History of Testicular Externalization and the Origin of the Scrotum." *Journal of Biosciences* 35(1): 27–37.

Kolmogorov, A. N. (1950) *Foundations of the Theory of Probability*. New York, NY: Chelsea.

Kreuth, H. (2005) *The Philosophy of Karl Popper*. Cambridge University Press.

Kuhn, T. (1957) *The Copernican Revolution*. Cambridge: Harvard University Press.

(1977) "Objectivity, Value Judgment, and Theory Choice." In *The Essential Tension – Selected Studies in Scientific Tradition and Change*. University of Chicago Press, pp. 320–339.

Kulkarni, S. and Harman, G. (2011) "Statistical Learning Theory: A Tutorial." www.princeton.edu/˜harman/Papers/SLT-tutorial.pdf.

Lange, M. (1995) "Spearman's Principle." *British Journal for the Philosophy of Science* 46: 503–521.

Laplace, P. S. (1796) *Exposition of the System of the World*. J. Pond (trans.). London: R. Phillips, 1809.

Laskar, J. (2008) "Chaotic Diffusion in the Solar System." *Icarus* 196: 1–15.

Leibniz, G. W. (1969) *Philosophical Papers and Letters*. L. Loemker (ed. and trans.). Dordrecht: Kluwer, 2nd edition, 1989.

(1686) *Discourse on Metaphysics*. In D. Garber and R. Ariew (eds. and trans.), *Discourse on Metaphysics and Other Essays*. Indianapolis: Hackett, 1991, pp. 1–40.

(1710) *Theodicy*. E. Huggard (trans.). New York: Cosimo Classics, 2010.

Lennox, J. (2001) *Aristotle's Philosophy of Biology*. Cambridge University Press.

Lewis, D. (1973) *Counterfactuals*. Oxford: Basil-Blackwell.

Lewontin, R. (2001) "In the Beginning was the Word." *Science* 291: 1263–1264.

Lewontin, R., and Dunn, L. (1960) "The Evolutionary Dynamics of a Polymorphism in the House Mouse." *Genetics* 45: 705–722.

Lipton, P. (1991) *Inference to the Best Explanation*. London: Routledge, 2nd edition 2004.

Littlewood, J. E. (1914) "Sur la Distribution des Nombres Premiers." *Comptes Rendus* 158: 1869–1872.

Locke, J. (1689) *An Essay Concerning Human Understanding*. P. Nidditch (ed.). Oxford: Clarendon, 1975.

Lovegrove, B. (2014) "Cool Sperm – Why Some Placental Mammals have a Scrotum," *Journal of Evolutionary Biology*, 27: 801–814.

Lurz, R. (2011) *Mindreading Animals*. Cambridge: MIT Press.

Maas, P. (2010) "Text Genealogy, Textual Criticism and Editorial Technique." In H. Jürgen and P. Maas (eds.), *Wiener Zeitschrift für die Kunde Südasiens*. Vienna: Austrian Academy of Sciences, 52–53: 63–120.

Mach, E. (1898) "On the Economical Nature of Physical Inquiry." In *Popular Scientific Lectures*. Chicago: Open Court, 1986, pp. 186–214.

MacKay, D. (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Malebranche, N. (1680) *Treatise on Nature and Grace*. P. Riley (trans.). Oxford: Clarendon, 1992.

Malthus, T. (1797) *An Essay on the Principle of Population*. Oxford University Press, 2008.

Maupertuis, P. (1746) "Les Lois de Mouvement et du répos déduites d'un principle métaphysique." *Mem. Acade. R. Sci. et Belles Lettres Berlin*, pp. 267–294.

Maxwell, J. C. (1856) "Essays for the Apostles on 'Analogies in Nature.'" In P. Harman (ed.), *Scientific Letters and Papers of James Clerk Maxwell* (Vol. 1). Cambridge University Press, 1990, pp. 376–383.

Maynard Smith, J. (1964) "Group Selection and Kin Selection." *Nature* 201: 1145–1146.

Mayr, E. (1983) "How to Carry Out the Adaptationist Program." *American Naturalist* 121: 24–34.

McDonough, J. (2009) "Leibniz on Natural Teleology and the Laws of Optics." *Philosophy and Phenomenological Research* 3: 505–544.

(forthcoming) "Descartes's Dioptrics and Optics." In L. Nolan (ed.), *The Cambridge Descartes Lexicon*. Cambridge University Press.

Meek, C. and Glymour, C. (1994) "Conditioning and Intervening." *British Journal for the Philosophy of Science* 45: 1001–1021.

Melis, A., Call, J., and Tomasello, M. (2006) "Chimpanzees (Pan troglodytes) Conceal Visual and Auditory Information from Others." *Journal of Comparative Psychology* 120: 154–162.

Mill, J. S. (1865) *An Examination of Sir William Hamilton's Philosophy*. In M. Robson (ed.), *Collected Works of John Stuart Mill* (Vol. 9). University of Toronto Press, 1963.

Miller, N. E. (1959) "Liberalization of Basic S-R Concepts." In S. Koch (ed.), *Psychology: A Study of a Science* (Vol. 1). New York, NY: McGraw Hill, pp. 196–292.

Milne, P. (2003) "Bayesianism v. Scientific Realism." *Analysis* 63: 281–288.

Moore, G. E. (1939) "Proof of an External World." *Proceedings of the British Academy* 25: 273–300.

Morgan, C. L. (1894) *An Introduction to Comparative Psychology*. London: Walter Scott.

(1903) *An Introduction to Comparative Psychology*. London: Walter Scott, 2nd edition.

Mougin, G. and Sober, E. (1994) "Betting Against Pascal's Wager." *Noûs* 28: 382–395.

Myrvold, W. (2003) "A Bayesian Account of the Virtue of Unification." *Philosophy of Science* 70: 399–423.

Nadler, S. (1990) "Deduction, Confirmation, and the Laws of Nature in Descartes' Principia Philosophiae." *Journal of the History of Philosophy* 28: 359–383.

(2008) *The Best of All Possible Worlds – A Story of Philosophers, God, and Evil*. New York, NY: Farrar, Straus, and Giroux.

Neugebauer, Otto (1957) *The Exact Sciences in Antiquity*. Providence, RI: Brown University Press.

Neurath, O. (1921) *Anti-Spengler*. Munich, G.D.W: Callwey.

Newcomb, S. (1895) *The Element of the Four Inner Planets and the Fundamental Constants in Astronomy*. Washington, DC: Government Printing Office, pp. 109–122.

Newton, I. (1687) *The Principia: Mathematical Principles of Natural Philosophy*. I. B. Cohen and A. Whitman (trans.). Berkeley, CA: University of California Press, 1999.

Nolan, D. (1997) "Quantitative Parsimony." *British Journal for the Philosophy of Science* 48: 329–343.

Norton, J. (2000) "'Nature is the Realisation of the Simplest Conceivable Mathematical Ideas': Einstein and the Canon of Mathematical Simplicity." *Studies in History and Philosophy of Modern Physics*, 31: 135–170.

(2003) "A Material Theory of Induction." *Philosophy of Science*, 70: 647–670.

O'Brien, M., Lyman, R., and Glover, D. (2003) *Cladistics and Archaeology*. Salt Lake City, UT: University of Utah Press.

Ockham, W. (1951, 1954) *Summa Logicae* (Vol. I-II). P. Boehner (ed.). St. Bonaventure, NY: Franciscan Institute Press.

(1981) *Quaestiones in librum secundum Sententiarum (Reportatio)*. In G. Gal and R. Wood (eds.), *Opera Theologica* (Vol. V). St. Bonaventure, NY: Franciscan Institute Press.

(1986a) *Scriptum in Librum Primum Sententiarum (Ordinatio)*. In G. J. Etzkorn and F. E. Kelley (eds), *Opera Theologica* (Vol. IV). St. Bonaventure, NY: Franciscan Institute Press.

(1986b) *Tractatus de Corpore Christi*. In C. Grassi (ed.), *Opera Theologica* (Vol. X), St. Bonaventure, NY: Franciscan Institute Press.

Oppenheimer, S. (2006) *The Origins of the British*. London: Robinson.

Paley, William, (1802) *Natural Theology*. Indianapolis: Bobbs-Merrill, 1963.

Palmer, E. (2002) "Pangloss Identified." *French Studies Bulletin* 84: 7–10.

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.

Peckham, M. (1967) *Man's Rage for Chaos*. New York: Schoken Books.

Penn, D. C., Holyoak, K. J., and Povinelli, D. J. (2008) "Darwin's Mistake: Explaining the Discontinuity between Human and non-Human Minds." *Behavioral and Brain Sciences* 31: 109–178.

Penn, D. C. and Povinelli, D. J. (2007) "On the Lack of Evidence that Chimpanzees Possess Anything Remotely Resembling a 'Theory of Mind.'" *Philosophical Transactions of the Royal Society B* 362: 731–744.

Plantinga, A. (2011) *Where the Conflict Really Lies – Science, Religion, and Naturalism*. New York, NY: Oxford University Press.

Platnick, N. and Cameron, D. (1977) "Cladistic Methods in Textual, Linguistic and Phylogenetic Analysis." *Systematic Zoology* 26: 380–385.

Poincaré, H. (1914) *Science and Method*. F. Maitland (trans.). London: Thomas Nelson.

Popper, K. R. (1959) *The Logic of Scientific Discovery*. London: Routledge Classics, 2002. (English translation of *Logic der Forschung*, 1934.)

(1963) *Conjectures and Refutations*. London: Routledge and Kegan Paul, 3rd edition, 1969.

(1983) *Realism and the Aim of Science*. W. W. Bartley III (ed.). London: Hutchinson.

Posada, D. and Buckley, T. (2004) "Model Selection in Phylogenetics – Advantages of the AIC and Bayesian Approaches." *Systematic Biology* 53: 793–808.

Posada, D. and Crandall, K. (2001) "Selecting the Best-Fit Model of Nucleotide Substitution." *Systematic Biology* 50: 580–601.

Povinelli, D. (1994) "Comparative Studies of Animal Mental State Attribution: a Reply to Heyes." *Animal Behavior* 48: 239–241.

Povinelli, D., Nelson, K., and Boysen, S. (1990) "Inferences about Guessing and Knowing by Chimpanzees (Pan troglodytes)." *Journal of Comparative Psychology* 104: 203–210.

Povinelli, D. and Vonk, J. (2003) "Chimpanzee Minds – Suspiciously Human?" *Trends in Cognitive Sciences* 7: 157–160.

(2004) "We Don't Need a Microscope to Explore the Chimpanzee's Mind." *Mind and Language* 19:1–28.

Ptolemy, C. (150?) *The Almagest*. G. Toomer (trans.), London: Duckworth, 1984.

Putnam, H. (1967) "Psychological Predicates." In W. Capitan and D. Merril. (eds.), *Art, Mind, and Religion*. Pittsburgh, PA: University of Pittsburgh Press, pp. 429–440.

(1971) *Philosophy of Logic*. Boston: Allen and Unwin.

(1975) "Philosophy and Our Mental Life." In *Mind, Language, and Reality, Philosophical Papers* (Vol. 2). Cambridge University Press, pp. 291–303.

Quine, W. (1953a) "On What There Is." In *From a Logical Point of View*. Cambridge, MA: Harvard University Press, pp. 1–19.

(1953b) "Two Dogmas of Empiricism." In *From a Logical Point of View*. Cambridge, MA: Harvard University Press, pp. 20–46.

(1960) *Word and Object*. Cambridge, MA: MIT Press.

Reichenbach, H. (1937) *Experience and Prediction*. University of Chicago Press.

(1956) *The Direction of Time*. Berkeley: University of California Press.

(1958) *The Philosophy of Space and Time*. New York, NY: Dover.

Rényi, A. (1970) *Probability Theory*. New York, NY: Elsevier.

Rescher, N. (1982) *Leibniz's Metaphysics of Nature*. Boston: Kluwer.

Richards, R. (2003) "Character Individuation in Phylogenetic Inference." *Philosophy of Science* 70: 264–279.

Robinson, P. and O'Hara, R. (1996) "Cladistic Analysis of an Old Norse Manuscript Tradition." *Research in Humanities Computing* 4: 115–137.

Roche, W. (2012) "A Weaker Condition for Transitivity in Probabilistic Support." *European Journal for the Philosophy of Science* 2: 111–118.

Roche, W. and Sober, E. (2013) "Explanatoriness is Evidentially Irrelevant, or Inference to the Best Explanation Meets Bayesian Confirmation Theory." *Analysis* 73: 659–668.

Rosen, E. (1959) *Three Copernican Treatises*, 2nd edition. New York, NY: Dover.

Rorty, R. (ed.) (1967) *The Linguistic Turn: Essays in Philosophical Method*. University of Chicago Press.

Rowe, W. (1979) "The Problem of Evil and Some Varieties of Atheism." *American Philosophical Quarterly* 16: 335–341.

Royall, R. (1997) *Statistical Evidence – a Likelihood Paradigm*. Boca Raton, FL: Chapman and Hall.

Ruse, M. and Wilson, E. O. (1986) "Moral Philosophy as Applied Science." *Philosophy* 61: 173–192. Reprinted in E. Sober (ed.), *Conceptual Issues in Evolutionary Biology*. Cambridge, MA: MIT Press, 2006.

Salmon, W. (1967) *Foundations of Scientific Inference*. Pittsburgh, PA: University of Pittsburgh Press.

(1984) *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Santos, L., Nissen, A., and Ferrugia, J. (2006) "Rhesus monkeys (Macaca mulatta) Know What Others can and cannot Hear." *Animal Behaviour* 71: 1175–1181.

Schulte, O. (1999) "Means-Ends Epistemology." *British Journal for the Philosophy of Science* 50: 1–31.

Schwarz, G. (1978) "Estimating the Dimension of a Model." *Annals of Statistics* 6: 461–465.

Shafer-Landau, R. (2007) "Moral and Theological Realism – the Explanatory Argument." *Journal of Moral Philosophy* 4: 311–329.

Shanahan, T. (2004) *The Evolution of Darwinism – Selection, Adaptation, and Progress in Evolutionary Biology*. Cambridge University Press.

Shank, M. (2003) "Rings in a Fluid Heaven – The Equatorium-Driven Physical Astronomy of Guido de Marchia (fl. 1292–1310)." *Centaurus* 45: 175–203.

Shapiro, L. and Sober, E. (2007) "Epiphenomenalism – The Do's and the Don'ts." P. Machamer and G. Wolters (eds.). *Thinking about Causes*. Pittsburgh, PA: University of Pittsburgh Press, pp. 235–264.

Shoemaker, S. (1969) "Time Without Change." *Journal of Philosophy* 66: 363–381.

Shogenji, T. (2003) "A Condition for Transitivity in Probabilistic Support." *British Journal for the Philosophy of Science* 54: 613–616.

Sider, T. (2013) "Against Parthood." In K. Bennett and D. Zimmerman (eds.) *Oxford Studies in Metaphysics*. Vol. 8. Oxford University Press, pp. 237–293.

Skinner. B. (1938) *The Behavior of Organisms: An Experimental Analysis*. New York, NY: Appleton Crofts.

Smart, J. J. C. (1959) "Sensations and Brain Processes." *Philosophical Review* 68: 141–156.

Sober, E. (1980) "Evolution, Population Thinking, and Essentialism." *Philosophy of Science* 47: 350–383.

(1988) *Reconstructing the Past – Parsimony, Evidence, and Inference*. Cambridge, MA: MIT Press.

(1989) "Independent Evidence about a Common Cause." *Philosophy of Science* 56: 275–287.

(1990a) *Core Questions in Philosophy*. New York, NY: Prentice-Hall.

(1990b) "Let's Razor Ockham's Razor." In D. Knowles (ed.), *Explanation and Its Limits*. Cambridge University Press, pp. 73–94.

(1993) "Mathematics and Indispensability." *Philosophical Review* 102: 35–58.

(1994) "Progress and Direction in Evolution." In J. Campbell (ed.), *Creative Evolution?!* Sudbury, MA: Jones and Bartlett Publishers, pp. 19–33.

(1998) "Black Box Inference – When Should an Intervening Variable be Postulated?" *British Journal for the Philosophy of Science* 49: 469–498.

(1999a) "Instrumentalism Revisited." *Crítica* 31: 3–38.

(1999b) "The Multiple Realizability Argument against Reductionism." *Philosophy of Science* 66: 542–564.

(1999c) "Physicalism from a Probabilistic Point of View." *Philosophical Studies* 95: 135–174.

(2000) "Quine's Two Dogmas." *Proceedings of the Aristotelian Society*, Supplementary Volume, 74: 237–280.

(2001) "Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause." *British Journal for the Philosophy of Science* 52: 331–346.

(2004) "Likelihood, Model Selection, and the Duhem-Quine Problem." *Journal of Philosophy* 101: 1–22.

(2007) "Intelligent Design Theory and the Supernatural – the 'God or Extra-Terrestrials' Reply." *Faith and Philosophy* 24: 72–82.

(2008a) "Empiricism." In S. Psillos and M. Curd (eds.), *The Routledge Companion to Philosophy of Science*, pp. 129–138.

(2008b) *Evidence and Evolution – The Logic behind the Science*. Cambridge University Press.

(2009a) "Absence of Evidence and Evidence of Absence – Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads." *Philosophical Studies* 144: 63–90.

(2009b) "Parsimony Arguments in Science and Philosophy — A Test Case for Naturalism$_{\text{P}}$." *Proceedings and Addresses of the American Philosophical Association* 83(2): 117–155.

(2009c) "Parsimony and Models of Animal Minds." In Lurz, R. (ed.), *The Philosophy of Animal Minds*. Cambridge University Press, pp. 237–257.

(2011a) *Did Darwin Write the Origin Backwards?* Amherst, NY: Prometheus Books.

(2011b) "Evolution without Naturalism." In J. Kvanvig (ed.), *Oxford Studies in Philosophy of Religion* (Vol. 3). Oxford University Press, pp. 187–221.

(2011c) "Reichenbach's Cubical Universe and the Problem of the External World." *Synthese* 181: 3–21.

(2011d) "Responses to Comments on *Evidence and Evolution* by B. Fitelson, R. Sansom, and S. Sarkar." *Philosophy and Phenomenological Research* 83: 692–704.

(2012a) "Anthropomorphism, Parsimony, and Common Ancestry." *Mind and Language* 27: 229–238.

(2015) "Two Cornell Realisms – Moral and Scientific." *Philosophical Studies* 172: 905–924.

Sober, E. and Forster, M. (1992): "Lessons in Likelihood – A Critique of Jefferys's and Berger's 'Ockham's Razor and Bayesian Analysis.'" *American Scientist* 80: 212–13.

Sober, E. and Steel, M. (2014) "Time and Knowability in Evolutionary Processes." *Philosophy of Science* 81: 537–557.

Sober, E. and Wilson, D. S. (1998) *Unto Others – the Psychology and Evolution of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Spade, P. and Panaccio, C. (2011) "William of Ockham." *Stanford Encyclopedia of Philosophy*. Fall edition. E. Zalta (ed.), http://plato.stanford.edu/archives/fall2011/entries/ockham.

Spirtes, P., Glymour, C., and Scheines, R. (2001) *Causality, Prediction, and Search*. Cambridge, MA: MIT Press.

Stanley, M. (2012) "By Design — James Clerk Maxwell and the Evangelical Unification of Science." *British Journal for the History of Science* 45: 57–73.

Steel, M. and Penny, D. (2000): "Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics." *Molecular Biology and Evolution* 17: 839–850.

Stigler, S. (1980) "Stigler's Law of Eponymy." In T. Gieryn (ed.), *Science and Social Structure – a Festschrift for Robert K. Merton*. New York, NY: NY Academy of Sciences, pp. 147–57.

Stone, M. (1977) "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion." *Journal of the Royal Statistical Society B* 39: 44–47.

Street, S. (2006) "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127(1): 109–166.

Sturgeon, N. (1984) "Moral Explanations." In D. Copp and D. Zimmerman (eds.), *Morality, Reason, and Truth*. Totowa, NJ: Rowman and Allanheld, pp. 49–78.

Swinburne, R. (2009) *Simplicity as Evidence of Truth*. Milwaukee: Marquette University Press.

Takeuchi, K. (1976) "Distribution of Information Statistics and a Criterion of Model Fitting." *Suri-Kagaku (Mathematical Sciences)* (in Japanese) 153: 12–18.

Theobald, D. L. (2010) "A Formal Test for the Theory of Universal Common Ancestry." *Nature* 465(13): 219–223.

Thorburn, W. (1918) "The Myth of Occam's Razor." *Mind* 27(107): 345–353.

Tishkoff S. A., Reed F. A., Ranciaro A., *et al*. (2007) "Convergent Adaptation of Human Lactase Persistence in Africa and Europe." *Nature Genetics* 39: 31–40.

Titelbaum, M. (2013) *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. New York, NY: Oxford University Press.

Tomasello, M. and Call, J. (2006) "Do Chimpanzees Know what Others See, or Only What They are Looking at?" In S. Hurley and M. Nudds (eds.), *Rational Animals*?. New York, NY: Oxford University Press, pp. 371–384.

Tooley, M. (2013) "The Problem of Evil." E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/sum2013/entries/evil.

True, J. R. and Haag, E. S. (2001) "Developmental System Drift and Flexibility in Evolutionary Trajectories." *Evolution and Development* 3:109–119.

Tuffley, C. and Steel, M. (1997) "Links Between Maximum Likelihood and Maximum Parsimony under a Simple Model of Site Substitution." *Bulletin of Mathematical Biology* 59: 581–607.

Tversky, A. and Kahneman, D. (1982) "Judgments of and by Representativeness." In D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Van Fraassen, B. (1980) *The Scientific Image*. New York, NY: Oxford University Press.

(1982) "The Charybdis of Realism – Epistemological Implications of Bell's Inequality." *Synthese* 52: 25–38.

(1985) "Empiricism in the Philosophy of Science." In P. Churchland and C. Hooker (eds.), *Images of Science. Essays on Realism and Empiricism with a Reply from Bas C. van Fraassen*. University of Chicago Press, pp. 245–308.

Vapnik, V. (2000) *The Nature of Statistical Learning Theory*. New York: Springer.

Vogel, J. (1990) "Cartesian Skepticism and Inference to the Best Explanation." *Journal of Philosophy* 87: 658–666.

Voltaire (1759) *Candide, or Optimism*. C. H. R. Niven (ed.). London: Longman, 1980.

von Luxburg, U. and Schöllkopf, B. (2009) "Statistical Learning Theory: Models, Concepts, and Results." In D. Gabbay, S. Hartmann, and J. Woods (eds.), *Handbook of the History of Logic. Volume 10: Inductive Logic*. Amsterdam: Elsevier, pp. 651–706.

Wade, M. (1978) "A Critical Review of Models of Group Selection." *Quarterly Review of Biology* 53: 101–114.

Wallace, A. R. (1905) *My Life*. London: Chapman and Hall.

Walton, D. (1996) *Arguments from Ignorance*. University Park, PA: The Penn State University Press.

Wasserman, L. (2000) "Bayesian Model Selection and Model Averaging." *Journal of Mathematical Psychology* 44: 92–107.

Wegener, A. (1928) *Die Entstehung der Kontinente und Ozeane*, 4th edition. Translated by J. Biram as *The Origin of Continents and Oceans*. New York: Dover, 1966.

Werdelin, L. and Nilsonne, Å (1999) "The Evolution of the Scrotum and Testicular Descent in Mammals: A Phylogenetic View." *Journal of Theoretical Biology* 196: 61–72.

Whewell, W. (1968) *William Whewell's Theory of Scientific Method*. R. Butts (ed.). Pittsburgh, PA: University of Pittsburgh Press.

(1833) *Astronomy and General Physics Considered with Reference to Natural Theology*. London: W. Pickering.

Wickens, D. (1938) "The Transference of Conditioned Excitation and Conditioned Inhibition from one Muscle Group to the Antagonistic Muscle Group." *Journal of Experimental Psychology* 22: 101–140.

Whitehurst, G. and Chingos, M. (2011) "Class Size: What Research Says and What it Means for State Policy." *Brown Center for Education Policy*, Brookings Institution.

Whiten, A. (1996) "When does Smart Behavior-Reading become Mind-Reading?" In P. Carruthers and P. Smith (eds.), *Theories of Theories of Mind*. Cambridge University Press, p. 277–292.

(2013) "Humans are not Alone in Computing how Others See the World." *Animal Behaviour* 86: 213–221.

Wiley, E. (1975) "Karl Popper, Systematics, and Classification." *Systematic Zoology* 24: 233–242.

(1981) *Phylogenetics – the Theory and Practice of Phylogenetic Systematics*. New York, NY: Wiley.

Williams, G. C. (1966) *Adaptation and Natural Selection*. Princeton, NJ: Princeton University Press.

(1992) *Natural Selection – Domains, Levels, and Challenges*. New York, NY: Oxford University Press.

Winther, R. (2009) "Character Analysis in Cladistics: Abstraction, Reification, and the Search for Objectivity." *Acta Biotheoretica* 57:129–162.

Wittgenstein, L. (1921) *Logisch-Philosophische Abhandlung*. Translated by D. Pears and B. McGuinness as *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul, 1961.

(1958) *The Blue and Brown Books*. New York, NY: Harper and Row, 1965.

Woodward, J. (2013) "Simplicity in the Best Systems Account of Laws of Nature." *British Journal for the Philosophy of Science* 65: 91–123.

Wrinch, D. and Jeffreys, H. (1921) "On Certain Fundamental Principles of Scientific Inquiry." *Philosophical Magazine* 42: 369–390.

Wright, L. (1976) "Functions." *Philosophical Review* 85: 70–86.

Wykstra, S. (1984) "The Humean Obstacle to Evidential Arguments from Suffering: On Avoiding the Evils of 'Appearance.'" *International Journal for Philosophy of Religion* 16: 73–93.

Young, A. (2012) "Discovery of the Law of Refraction." http://mintaka.sdsu.edu/GF/explain/optics/discovery.html.

Yule, G. U. (1926) "Why do we Sometimes get Nonsensical Relations between Time Series? A Study of Sampling and the Nature of Time Series." *Journal of the Royal Statistical Society* 89: 1–64.

Index