P

PARADIGM

See **Kuhn, Thomas**

PARSIMONY

The principle of parsimony—or simplicity (treated here as an equivalent concept)—is also known as Occam's (or Ockham's) razor, after William of Occam, the medieval philosopher who said that plurality is not to be assumed without necessity and that what can be done with fewer assumptions is done in vain with more (Wood 1996, 20–22).

Scientists and philosophers often claim that the *parsimony* of a theory is relevant to deciding whether the theory is true or approximately true or would make accurate predictions. How this can be so is a central puzzle in the epistemology of science. It is not puzzling that people find parsimonious theories aesthetically attractive and easy to understand and manipulate. What requires elucidation is not the pragmatic value but the *epistemic* value of parsimony.

Just as people are said to be parsimonious when they are abstemious in how they spend money, a theory is parsimonious when it is tightfisted with respect to the entities, processes, or events it postulates. There is no cutoff separating theories that are parsimonious from theories that are not; rather, the difference is a matter of degree. The fundamental idea is comparative: One theory is more parsimonious than another. For example, if one theory postulates causes A and B to explain an observed effect E, while a second theory postulates only cause A and does not mention B, the latter theory is more parsimonious.

One epistemically significant feature of this difference is that if A and B are mutually independent, then according to probability theory the conjunction $(A \wedge B)$ will be less probable than A. Does this mean that parsimony and probability always coincide? In what follows, it will be seen that a number of philosophers have strenuously denied this. And even in the case at hand, there is reason to be careful about the suggestion. The second theory is "agnostic" about the relevance of B. But now consider a third theory, which asserts that A is a cause of E and denies that B is a cause. This third theory is "atheistic" about B and is more parsimonious than the first theory. However, probability theory does not say that $(A \wedge \neg B)$—that is, $(A \wedge \neg B)$—is more probable than $(A \wedge B)$. The hypothesis that there is at least one cause of E is more probable than the hypothesis that there are at least two causes, but there is no a priori reason to think that exactly one cause is more probable than exactly two. Parsimony has an obvious link with probability when a logically stronger hypothesis is compared with a hypothesis that is simpler and logically weaker; however, when two theories are mutually incompatible, the connection is anything but obvious.

The giants of the Scientific Revolution frequently referred to the importance of parsimony and its cognates. In *De revolutionibus orbium caelestium*, Copernicus emphasizes that his heliocentric theory differs from Ptolemy's geocentric theory in that the Ptolemaic system requires an independent model for the motion of each planet, whereas the Copernican system unifies the models for the different planets by including a common Earth/sun component in each. Copernicus remarks that his approach "follow[s] Nature, who producing nothing vain or superfluous often prefers to endow one cause with many effects" (Kuhn 1957, 176–179). Newton ([1686] 1953), in *Principia mathematica*, states as his first rule of reasoning in philosophy that

> we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity and affects not the pomp of superfluous causes. (3)

Leibniz ([1686] 1973, 11) defended parsimony as a criterion in scientific reasoning by appeal to his doctrine that God created the best of all possible worlds; our world is perfect because it is "at the same time the simplest in hypotheses and the richest in phenomena." For all these thinkers, the methodological principle rests on an ontological foundation. The principle of parsimony should be used in reasoning because nature is simple, and nature is simple because God made it so.

With the falling away of divine design as an acceptable justification of methodological principles, a fissure appeared in the foundation of scientific inference. If the principle of parsimony cannot be justified by tracing it back to a parsimonious creator, what could its justification be? Does the justification of the principle require any substantive assumptions about the natural world? Or is the principle just part and parcel of what it means to be "rational," which we are required to be no matter what the world is like? If the theological account is the thesis, its antithesis is the idea that the principle of parsimony is purely methodological. In between these two extremes, there is much room for accounts that reject both.

Local Versus Global Accounts

Most attempts to explain the epistemic relevance of parsimony treat the problem globally. They assume that if parsimony is epistemically relevant across a range of problems involving inference, the reason for its relevance must always be the same. However, it is worth pondering the possibility that the justification for using a principle of parsimony may vary from problem to problem. Perhaps parsimony needs to be understood not globally but locally (Sober 1990).

As an example, consider the long-standing use of parsimony as a criterion for inferring phylogenetic relationships in evolutionary biology (Sober 1988). Given a set of observed similarities and differences that characterize a set of species, how are these data to be used to figure out which species are closely related and which are related more distantly? A standard procedure is to find the phylogenetic tree that requires the smallest number of changes in "character state" to explain the data. This methodology assumes that the species are genealogically related and proceeds to identify the most parsimonious hypothesis concerning what the pattern of relatedness is. However, there is a prior question about phylogeny: Why think that the observed species have any common ancestors? Perhaps each species can be traced back to a separate origin.

The role of parsimony in answering this question can be understood by examining Crick's (1968) argument that the near universality of a single genetic code among the organisms now on Earth is evidence that they are all genealogically related. Crick says that the shared genetic code is arbitrary—one

among a large number of viable mappings of nucleotide triplets onto amino acids. However, once an organism uses a given code, its fitness is likely to be compromised if it or its descendants modify the code already in place. Stabilizing selection then makes it highly probable that descendants will use the same genetic code as their ancestors. These biological assumptions (which Crick summarizes in the phrase "frozen accident") entail that the universality of the code would be very surprising if the organisms now on Earth were not genealogically related (e.g., were products of 27 separate startups) but is precisely what one should expect if all life can be traced back to a single progenitor. Because of this difference, Crick concludes that the observed universality strongly favors one hypothesis over the other. Notice that Crick's argument compares the likelihood of two hypotheses:

> P (the code is now universal | all current life traces back to a single progenitor) > P (the code is now universal | current life traces back to 27 original progenitors and no fewer).

Here 'likelihood' is used in the technical sense introduced by R. A. Fisher (1925): The likelihood of a hypothesis is the probability it confers on observations, not the probability of the hypothesis, given the observations. The likelihood of H is $P(O|H)$; its posterior probability is $P(H|O)$. According to the law of likelihood, the observations differentially support the hypothesis of higher likelihood (Edwards 1972; Hacking 1965; Royall 1997).

The hypothesis that life can be traced back to a single progenitor is simpler than the hypothesis that it has 27 separate startups (since $1 < 27$). Crick's argument thus provides an example in which the principle of parsimony has a justification in terms of likelihood. However, the connection of likelihood and parsimony in this instance depends on specifically biological assumptions about the genetic code—that it is arbitrary and that it is subject to stabilizing selection. If parsimony has a rationale based on likelihood in inferential problems that arise in other sciences, different empirical assumptions will be required to show that this is so. But more important, there seem to be problems in which parsimony cannot be justified in terms of likelihood; in these problems, likelihood and parsimony are actually at odds.

The inferential task of curve fitting provides an example. Consider the following experiment. A sealed pot is put on a stove. Attached to the pot are a thermometer and a device that measures how much pressure the gas inside exerts on the walls of the pot. The pot is heated to various temperatures, and the resulting pressures are observed. Each temperature reading with its associated pressure reading can be represented as a point in a coordinate system (see Figure 1). The problem is to use these observations to determine the general relationship between temperature and pressure for this system. Each hypothesis about this general relationship takes the form of a curve. Which curve is most plausible, given the observations?

One factor that scientists take into account is goodness of fit. A curve that comes close to the data fits them better than a curve that is more distant. If goodness of fit were the only relevant consideration, scientists would always choose curves that pass exactly through the data points. But they do not do this—and even if they did, the question would remain how to choose among the infinity of curves that fit the data perfectly. Another consideration apparently influences their decisions, and this is simplicity. Often, extremely bumpy curves are thought to be complex, whereas smoother curves are thought to be simpler. Scientists sometimes reject an extremely bumpy curve that fits the data perfectly in favor of a smoother curve that fits the data slightly less well. Scientists care about both goodness of fit and simplicity, which influence how they choose curves in the light of data. However, these two desiderata conflict: Increasing simplicity typically involves reducing goodness of fit.

A curve represents a deterministic relationship between temperature and pressure; it maps x-values onto unique y-values. However, a curve plus an error distribution represents a probabilistic relationship: Any x-value is associated with a distribution of possible y-values, each with its own probability

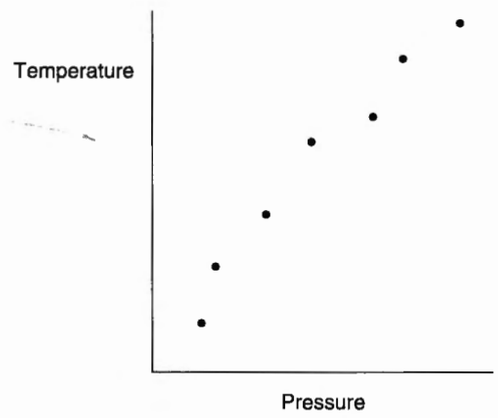

Fig. 1. Data gathered from an experiment in which a pot on a stove is raised to different temperatures and the pressure is recorded.

(density). In the example at hand (Figure 1), the concept of curve plus error distribution is more plausible, since the data are the joint product of the true underlying relationship of temperature and pressure and the measurement errors introduced by the imperfections of the thermometer and the pressure gauge. A standard model of error effects a connection between goodness of fit and likelihood: If one curve fits the data better than another, then the former confers a higher probability on the data. Given the data set depicted in Figure 1, a straight line will have a lower likelihood than a sufficiently complex curve that passes exactly through each data point. Thus, even if simplicity has a rationale based on likelihood in Crick's argument, simplicity and likelihood apparently conflict in the context of curve fitting.

Simplicity and Parsimony

It seems natural to say that curves differ in their simplicity. But what would it mean to say that they differ in parsimony? Parsimony involves paucity of postulation, but how does the idea of abstemiousness apply to curve fitting? Curves are visual representations of equations. For example, a straight line is a representation of a linear equation, which has the form

$$(\text{LIN}) y = a + bx$$

and a parabola is a representation of a quadratic equation, which has the form

$$(\text{PAR}) y = a + bx + cx^2$$

where x and y are the independent and dependent variables, respectively, and a, b, and c are adjustable parameters. In such equations, the adjustable parameters represent existential quantifiers—for example, (LIN) says that there exist values for a and b such that $y = a + bx$. Therefore, (LIN), which makes two "existence claims," may seem more parsimonious than (PAR), which makes three. This point pertains to equations of the form (LIN) and (PAR), not to a specific straight line and a specific parabola (an important distinction, which will come up again). It is worth asking whether simplicity and parsimony in their vernacular meanings always come to the same thing. However, as noted at the outset, the present discussion follows the conventional practice of treating them as equivalent.

Bayesianism

Bayesianism is not the same as Bayes's theorem. The theorem says that the conditional probability $P(H \mid O)$—the probability of H, given O—is a function of three other quantities:

$$P(H \mid O) = P(O \mid H)P(H)/P(O).$$

This theorem is a consequence of the standard definition of conditional probability: $P(H \mid O) = P(H \wedge O)/P(O)$. Bayesianism is a philosophical position, not a mathematical truth; in its strongest form it asserts that the epistemic notion of plausibility can be understood in terms of the mathematical concept of probability and, furthermore, that all the epistemic concepts bearing on empirical inquiry can be understood in terms of the probabilistic relationships described by Bayes's theorem. A double application of this theorem yields the following comparative principle:

$$P(H_1 \mid O) > P(H_2 \mid O) \text{ if and only if } P(O \mid H_1)P(H_1) > P(O \mid H_2)P(H_2).$$

This biconditional makes it clear that Bayesianism can use exactly two ingredients in explaining how parsimony is able to render one hypothesis more plausible than another in light of a set of observations. If parsimony influences plausibility, it must do so through prior probabilities, likelihoods, or both. If the relevance of simplicity cannot be accommodated in one of these two ways, then either simplicity is epistemically irrelevant or (strong) Bayesianism is mistaken. As noted previously in connection with curve fitting, likelihood can be maximized by making one's hypothesis sufficiently complex; this seems to leave Bayesianism only one alternative: If simplicity in such cases influences the plausibility of a hypothesis, it must do so because simpler theories have higher prior probabilities. This led Jeffreys (1957) to introduce a "simplicity postulate," according to which the complexity of an equation is measured by summing its variables, exponents, and parameters. This simplicity ordering is then said to provide an ordering of the prior probabilities of the hypotheses.

Popper (1959) argued that this postulate is incompatible with the axioms of probability. It assigns (LIN) a higher prior probability than (PAR), but this is impossible, since (LIN) entails (PAR). Howson (1988) replied that the problem can be evaded by stipulating that the parameters in a model have nonzero values. Instead of comparing (LIN) and (PAR), we should compare (LIN*) and (PAR*), which stipulate that $a, b, c \neq 0$. These models are disjoint, not nested, so assigning (LIN*) a higher prior probability is consistent with the axioms.

This suggestion raises two new questions. First, why should the original problem—comparing (LIN)

and (PAR)—be ignored? Should it be said that these two models are not in competition, because they are compatible? If so, scientific practice needs to change, since scientists often compare nested models.

Second, why should (LIN*) be assigned a higher prior probability than (PAR*)? Why think that $c = 0$ is more probable than $c \neq 0$? If probabilities are merely degrees of subjective belief, it is undeniable that someone might have greater confidence in the hypothesis that $c = 0$. However, it is puzzling why, in the absence of evidence, one should feel this way. If a sharp pin is dropped on a line a mile long, would you bet that the pin will land exactly at the beginning of the line or that it will land somewhere else? In the absence of information concerning how the pin is dropped, it is hard to see why you should bet on the first probability—yet this is precisely what Jeffreys's simplicity postulate recommends.

Another problem with this postulate has to do not with its correctness but with its completeness: It imposes an ordering of prior probabilities without providing specific values. This is important in inferential problems when the more complex hypothesis has a higher likelihood. If H_1 has the higher likelihood and H_2 has the higher prior probability, which has the higher posterior probability? Determining how simplicity trades off against likelihood requires more than a simplicity ordering.

Although Jeffreys held out no hope of getting likelihood and parsimony to coincide, later Bayesians saw a way to reopen the question. To grasp their idea, it is important to understand the difference between a model (which contains at least one adjustable parameter) and a specific hypothesis (which contains none). In this regard (LIN) is a model, but $y = 2 + 3x$ is not; it is a specific linear hypothesis. In effect, a model is a disjunction of specific hypotheses. When it was noted earlier that a sufficiently complex equation will fit the data better than a simpler equation, the point pertained to specific hypotheses. However, what would it mean to talk about the likelihood of a model? It is clear how $y = 2 + 3x$ "probabilifies" the data (once an error distribution is specified). But what probability does (LIN) confer on them? The answer is that the likelihood of (LIN) is the average likelihood of the set of straight lines ($i = 1, 2, \ldots$):

$$P(\text{data} \mid \text{LIN}) = \Sigma_i P(\text{data} \mid \text{straight line } i)\, P(\text{straight line } i \mid \text{LIN}).$$

The first term in this summation makes sense, but what should we make of the second? If the relation between temperature and pressure in the example of a pot on a stove is linear, what probabilities do the different specific linear hypotheses have? Schwarz (1978) approached this problem by thinking about the ratio of the average likelihoods of two models, assuming that there is a flat, uniform distribution over parameter values in each model. He derived the following result, which came to be known as the Bayesian information criterion (BIC):

$$\log[P(\text{data} \mid \text{model } M)] = \log\{P[\text{data} \mid L(M)]\} - (k/2)\log(N),$$

where $L(M)$ is the likeliest member of model M, N is the number of data, and k is the number of parameters in M. Notice that BIC includes a penalty term for complexity. If the best-fitting straight line and the best-fitting parabola fit the data in our example about equally well, (PAR) will have the lower estimated average likelihood because it is more complex. Complexity is relevant to estimating the average likelihoods of models, so Jeffreys's recourse to "priors" in his simplicity postulate is not, as it turns out, the only Bayesian approach to the problem.

One virtue of Schwarz's analysis is its avoidance of the criticism already noted—that it seems arbitrary and implausible, if not contradictory, to assign simpler models higher prior probabilities. (Nonetheless, questions can be raised about the assumed flat prior distribution of the values a parameter might have in a model.) Another virtue is that BIC specifies an exact quantitative rule for trading off simplicity and the likelihood of $L(M)$; it describes how much of a gain in one is required for a given loss in the other, if there is to be a net improvement in the estimated average likelihood of the model. However, there is a fly in the ointment. Schwarz's derivation uses improper priors (i.e., priors that do not sum to unity) in such a way that his derivation is not invariant under reparameterization (Forster and Sober 1994). Subsequent Bayesian work derives BIC so as to avoid this defect: The strategy is to use some of the data to transform the initial, improper priors into proper "posteriors"; thereafter, the rest of the data are taken into account to compute the final, average likelihood. (For further discussion, see Wasserman 2000.)

Popper and Falsifiability

Popper (1959) proposed a demarcation criterion that separates scientific from nonscientific statements:

The former are falsifiable. A falsifiable statement is one that is incompatible with a finite conjunction of observation statements. Falsifiable statements do not have to be false; rather, they have the nice property that observation can disprove them if, in fact, they are untrue.

Just as falsifiability separates science from nonscience, so degree of falsifiability distinguishes some scientific statements from others. The (LIN) model can be falsified by three data points, but not by any smaller number. A single data point, or any pair of data points, can be supplied with a straight line that passes through them exactly. On the other hand, (PAR) requires at least four data points to be falsified. This means that (LIN) is more falsifiable than (PAR).

Popper saw this as the key to understanding simplicity in science. Simpler theories are easier to falsify: If they are false, fewer data are required to show this. Popper turns Jeffreys's simplicity postulate on its head; whereas Jeffreys thinks that simpler theories are more probable, Popper thinks that simplicity goes with greater content: Simpler theories say more and hence are less probable.

It is clear that more falsifiable hypotheses have a pragmatic virtue: It is easier for us to prove them false, if they are false. The principal reservation philosophers have had regarding Popper's analysis is that it fails to account for the epistemic significance of parsimony. Why should predictions be based on simpler models rather than on more complicated models that fit the data equally well? It is here that Popper aligns himself with the skeptic and in opposition to the Bayesian. There is no assurance that our best hypotheses are true or even probably true. All that can be said is that they so far have evaded our best attempts to disprove them. Simplicity can provide no guarantee of truth or of probable truth, for the simple reason that nothing can.

There are further problems with Popper's account of simplicity. First, although it entails that (LIN) is simpler than (PAR), it does not have this consequence when a specific straight line and a specific parabola are compared. Each can be falsified by a single data point, so the two are equally falsifiable; this means that Popper must say they are equally simple. In addition, Popper's notion of degrees of falsifiability is restricted to hypotheses that have deductive consequences (perhaps in conjunction with auxiliary assumptions) about observations. If the hypotheses in question confer only probabilities on the data, they are not falsifiable. Since observation is virtually always subject to error, this is a large gap in Popper's theory.

Akaike and Selecting Models

The Bayesian approach to selecting models is not the only game in town. Before Schwarz (1978) proved his result, Akaike (1973) provided an alternative treatment (see also Sakamoto, Ishiguro, and Kitagawa 1986; Burnham and Anderson 1998). In fact, Akaike's contribution was twofold: He described a goal for selecting models—predictive accuracy—and he proved a theorem concerning how the predictive accuracy of a model can be estimated (Forster and Sober 1994).

How might a model like (LIN) be used to make a prediction about the pressure in our pot if the pot is brought to a certain temperature? A specific linear hypothesis, such as $y = 2 + 3x$, makes a prediction about the y-values that will be associated with newly observed x-values, but what does (LIN) tell us to expect? The answer is that (LIN) makes predictions by a two-step process. First, one uses old data to estimate the maximum likelihood values of the parameters in (LIN); then one uses this fitted model to predict new data. Thus, from the old data and (LIN) one obtains L(LIN), the likeliest member of (LIN); it is L(LIN) that makes a definite prediction about new data.

How well will L(LIN) predict new data? That depends, of course, on the true underlying relationship between temperature and pressure. In addition, since different data sets drawn from the same underlying distribution may differ, L(LIN) may make fairly accurate predictions about some data sets and rather inaccurate predictions about others. Because data sets may vary, it makes sense to define the predictive accuracy of a model as its average performance across multiple data sets.

If maximizing predictive accuracy is the goal, how is this goal to be achieved? How can we tell whether a model will make accurate predictions about new data, given just the single data set that we have at hand? If we opt for the model that best fits the data, we will usually select a fairly complex model. Working scientists know from practical experience that a complex model fitted to old data is often a poor predictor of new data; in such cases, the model is said to "overfit" the data. Sometimes a simpler model, although it does not fit the old data as well, will be a better predictor of new data. A mathematical explanation of this familiar fact is provided by Akaike's (1973) theorem:

$$\text{An unbiased estimate of the predictive accuracy of model } M = \log P\,[\text{data} \mid L(M)] - k,$$

where k is the number of adjustable parameters in the model. We obtain the log-likelihood of the

best-fitting member of the model and then subtract k, which is a penalty for complexity. This estimate is called the Akaike information criterion (AIC) score of the model. Forster and Sober (1994) recommend representing the estimate per datum—that is, multiplying the right-hand side by $1/N$, where N is the number of data; this helps defuse the criticism that AIC is statistically inconsistent (Forster 2002). Although it is intuitive to think about Akaike's framework in the context of curve fitting, it and other criteria for selecting models apply to a far larger range of inference problems, including those that arise in causal modeling (Forster and Sober 1994).

Akaike's theorem, as such, must be considered for the assumptions that go into its proof. First, there is an assumption about the uniformity of nature, which has two parts: (1) It says that the old and new data sets described in the definition of predictive accuracy are drawn from the same underlying distribution, and (2) it assumes that the x-values sampled in different data sets are drawn from a single distribution. For this reason, Forster (2000) describes AIC as addressing the problem of interpolation; the model-selection criterion that would be appropriate for extrapolation is not described by Akaike's theorem, whose proof also requires an assumption about normality; roughly, this says that repeated estimates of a parameter in a model form a normal distribution.

What does it mean to say that AIC is unbiased? If your bathroom scale is unbiased, it may give different readings of what you weigh, but the average of these must be your true weight. If the scale is unbiased, so is the procedure of adding or subtracting 50 percent of what it says, depending on the result of tossing a fair coin. This second estimation procedure also is centered on the true value, but it has higher variance than the procedure that just takes the scale's reading at face value. Similarly, the fact that AIC provides an unbiased estimate of a model's predictive accuracy leaves open whether its estimates have minimum variance. Furthermore, it is not clear that lack of bias should be regarded as a necessary condition for an acceptable estimator. Suppose that a scale has very low variance but is slightly biased; on average, it reads a little too high or a little too low (it is not clear which). Would one decline to use this scale if the alternative is to use a scale that is unbiased but has enormous variance?

In the literature on selecting models, AIC and BIC are often treated as competitors. This is odd, since the two criteria were derived as solutions for different problems. BIC estimates average likelihood; AIC estimates predictive accuracy. This does not mean that they cannot be considered as possible solutions to the same problem; however, to do so involves wrenching one of them from its natural conceptual home. Forster (2002) describes a set of simulations in which AIC is better at estimating predictive accuracy in some circumstances, while BIC is better in others. If we knew in advance where the problem we want to solve is located in parameter space, such simulations might indicate which model-selection criterion to use. However, the sad fact of the matter is that we often do not know enough about the factual setting of a problem for this to be possible.

Akaike's framework and criterion have important implications for the debate concerning realism, empiricism, and instrumentalism. It often turns out that a model known to be false has a higher AIC score than a model known to be true. This means that the goal of finding predictively accurate models differs from the goal of finding true models. If realists maintain that the goal of science is to find true theories, and empiricists maintain that the goal of science is to find empirically adequate theories (van Fraassen 1980), then Akaike's framework and theorem open the door to a third possibility. Instrumentalism, shorn of the faulty philosophy of language, which led it to deny that theories have truth values, becomes an option worth exploring (Sober 2002).

ELLIOTT SOBER

The author thanks Malcolm Forster, Steven Nadler, and Kyle Stanford for helpful discussion.

References

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, 267–281.

Burnham, K., and D. Anderson (1998), *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Crick, F. (1968), "*The Origin of the Genetic Code*," *Journal of Molecular Biology* 38: 367–379.

Edwards, A. (1972), *Likelihood*. Cambridge: Cambridge University Press.

Fisher, R. (1925), *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Forster, M. R. (2000), "*Key Concepts in Model Selection: Performance and Generalizability*," *Journal of Mathematical Psychology* 44: 205–231.

——— (2002), "*The New Science of Simplicity*," in A. Zellner, H. Keuzenkamp, and M. McAleer (eds.), *Simplicity, Inference, and Modeling*. Cambridge: Cambridge University Press, 83–119.

Forster, M., and E. Sober (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.
Hacking, I. (1965), *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.
Howson, C. (1988), "*On the Consistency of Jeffreys's Simplicity Postulate and Its Role in Bayesian Inference*," *Philosophical Quarterly* 38: 68–83.
Jeffreys, H. (1957), *Scientific Inference*, 2nd ed. Cambridge: Cambridge University Press.
Kuhn, T. (1957), *The Copernican Revolution*. Cambridge, MA: Harvard University Press.
Leibniz, G. ([1686] 1973), *Discourse on Metaphysics*. LaSalle, IL: Open Court. (First published in 1840.)
Newton, I. ([1686] 1953), "*Rules of Reasoning in Philosophy*," from *Philosophiae naturalis principia mathematica*, in H. Thayer (ed.), *Newton's Philosophy of Nature*. New York: Hafner.
Popper, K. (1959), *Logic of Scientific Discovery*. London: Hutchinson.
Royall, R. (1997), *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, FL: Chapman and Hall.
Sakamoto, Y., M. Ishiguro, and G. Kitagawa (1986), *Akaike Information Criterion Statistics*. New York: Springer.
Schwarz, G. (1978), "*Estimating the Dimension of a Model*," *Annals of Statistics* 6: 461–465.
Sober, E. (1988), *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge, MA: MIT Press.
——— (1990), "*Let's Razor Ockham's Razor*," in D. Knowles (ed.), *Explanation and Its Limits*. Cambridge: Cambridge University Press, 73–94.
——— (2002), "*Instrumentalism, Parsimony, and the Akaike Framework*," *Philosophy of Science* 69: S112–S123.
van Fraassen, B. C. (1980), *The Scientific Image*. New York: Oxford University Press.
Wasserman, L. (2000), "*Bayesian Model Selection and Model Averaging*," *Journal of Mathematical Psychology* 44: 92–107.
Wood, R. (1996), *Ockham on the Virtues*. West Lafayette, IN: Purdue University Press.

See also **Bayesianism; Popper, Karl**

PARTICLE PHYSICS

Because its subject matter is the structure and behavior of the fundamental constituents of the physical world, particle physics involves a unique blend of experimental, theoretical, and philosophical issues. In the 1930s, early particle physics looked at cosmic ray traces in Wilson cloud chambers and results from particle accelerators probing distances of the order of the size of an atomic nucleus (energies of about 100 million electron-volts [MeV]) (Brown and Hoddeson 1983). Particle physics made use of a patchwork of models that included quantum electrodynamics, Enrico Fermi's theory of radioactive β-decay, and Hideki Yukawa's meson theory. Today, the Tevatron at Fermilab collides protons and antiprotons accelerated to 980,000 MeV to create charged jets of strongly interacting particles that have probably not existed in the universe since the big bang. The standard model in particle physics, with its unified theory of three of the four fundamental forces in nature, is the current dominant paradigm. A central conceptual problem of particle physics since its inception has been to define and establish an appropriately tight connection between an account of the basic building blocks of matter (fundamental theory) and experimental predictions (phenomenological models).

Early Particle Physics

Particle physics emerged in the 1930s as a confluence of three distinct fields of physics: nuclear physics, cosmic ray physics, and quantum field theory (QFT). The consensus at the time was that matter was composed entirely of two fundamental particles—negatively charged electrons and positively charged protons—and that there were two fundamental forces of nature: electromagnetism, mediated by the photon, and gravitation. Atoms were thought to be composed of electrons orbiting a nucleus composed of protons and electrons bound tightly together. Early observations of fair-weather "atmospheric electricity" at the turn of the twentieth century had expanded by late 1920s into a well-developed program of research studying cosmic rays, that is, ionizing radiation from outer space. This radiation was thought to consist of high-energy photons.

On the theoretical side, physics by the mid-1920s appeared equally successful and complete. Einstein's