# 7

# Let's razor Ockham's razor

## 1. INTRODUCTION

When philosophers discuss the topic of explanation, they usually have in mind the following question: Given the beliefs one has and some proposition that one wishes to explain, which subset of the beliefs constitutes an explanation of the target proposition? That is, the philosophical 'problem of explanation' typically has bracketed the issue of how one obtains the beliefs; they are taken as given. The problem of explanation has been the problem of understanding the relation '$x$ explains $y$'. Since Hempel (1965) did so much to canonize this way of thinking about explanation, it deserves to be called 'Hempel's problem'.

The broad heading for the present essay departs from this Hempelian format. I am interested in how we might justify some of the explanatory propositions in our stock of beliefs. Of course, issues of theory confirmation and acceptance are really not so distant from the topic of explanation. After all, it is standard to describe theory evaluation as the procedure of 'inference to the best explanation'. Hypotheses are accepted, at least partly, in virtue of their ability to explain. If this is right, then the epistemology of explanation is closely related to Hempel's problem.

I should say at the outset that I take the philosopher's term 'inference to the best explanation' with a grain of salt. Lots of hypotheses are accepted on the testimony of evidence even though the hypotheses could not possibly be explanatory of the evidence. We infer the future from the present; we also infer one event from another simultaneously occurring event with which the first is correlated. Yet the future does not explain the present; nor can one event explain another that occurs simultaneously with the first. Those who believe in inference to the best

explanation may reply that they do not mean that inferring $H$ from $E$ requires that $H$ explain $E$. They have in mind the looser requirement that $H$ is inferrable from $E$ only if adding $H$ to one's total system of beliefs would maximize the overall explanatory coherence of that system. This global constraint, I think, is too vague to criticize; I suspect that 'explanatory coherence' is here used as a substitute for 'plausibility.' I doubt that plausibility can be reduced to the concept of explanatoriness in any meaningful way.

Another way in which philosophical talk of 'inference to the best explanation' is apt to mislead is that it suggests a gulf between the evaluation of explanatory hypotheses and the making of 'simple inductions'. Inductive inference, whether it concludes with a generalization or with a prediction about the 'next instance', often is assumed to markedly differ from postulating a hidden cause that explains one's observations. Again, I will merely note here my doubt that there are distinct rules for inductive and abductive inference.

Although I am not a card-carrying Bayesian, Bayes' theorem provides a useful vehicle for classifying the various considerations that might affect a hypothesis's plausibility. The theorem says that the probability that $H$ has in the light of evidence $E$ ($P[H/E]$) is a function of three quantities:

$$P(H/E) = P(E/H)P(H)/P(E).$$

This means that if one is comparing two hypotheses, $H_1$ and $H_2$, their overall plausibility (posterior probability) is influenced by two factors:

$$P(H_1/E) > P(H_2/E) \text{ iff } P(E/H_1)P(H_1) > P(E/H_2)P(H_2).$$

$P(H)$ is the prior probability of $H$ – the probability it has before one obtains evidence $E$. $P(E/H)$ is termed the *likelihood* of $H$; the likelihood of $H$ is not $H$'s probability, but the probability that $H$ confers on $E$.

Likelihood is often a plausible measure of explanatory power. If some hypothesis were true, how good an explanation would it provide of the evidence ($E$)? Let us consider this as a comparative problem: We observe $E$ and wish to know whether one hypothesis ($H_1$) would explain $E$ better than another hypothesis ($H_2$) would. Suppose that $H_1$ says that $E$ was to be expected, while $H_2$ says that $E$ was very improbable. Likelihood judges $H_1$ better supported than $H_2$; it is natural to see this judgment as reflecting one dimension of the concept of explanatory power.[1]

Hypotheses we have ample reason to believe untrue may nonetheless be explanatory. They may still have the property of being such that

*IF* they were true, they would account well for the observations. This judgment about antecedent plausibility the Bayesian tries to capture with the idea of prior probability.

There is little dispute about the relevance of likelihood to hypothesis evaluation; nor is there much dispute as to whether something besides the present observations can influence one's judgment about a hypothesis' overall plausibility. The main matter of contention over Bayesianism concerns whether hypotheses always have well-defined priors. The issue is whether prior probability is the right way to represent judgments about antecedent plausibility.

When the hypotheses in question describe possible outcomes of a chance process, assigning them prior probabilities is not controversial. Suppose a randomly selected human being has a red rash; we wish to say whether it is more probable that he has measles or mumps. The prior probability of a disease is just its population frequency. And the likelihoods also are clear; I can say how probable it would be for someone to have the red rash if he had measles and how probable the symptom would be if he had mumps. With these assignments of priors and likelihoods, I can calculate which posterior probability is greater.

Do not be misled by the terminology here. The prior probabilities in this example are not knowable *a priori*. The prior probability of the proposition that our subject has measles is the probability we assign to that disease when we do not know that he happens to have a red rash. The fact that he was randomly drawn from a population allows us to determine the prior probability by observing the population.

Matters change when the hypotheses in question do not describe the outcomes of chance processes. Examples include Newton's theory of gravity and Darwin's theory of evolution. A Bayesian will want to assign these prior probabilities and then describe how subsequent observations modify those initial assignments. Although likelihoods are often well-defined here, it is unclear what it would mean to talk about probabilities.

Bayesians sometimes go the subjective route and take prior probabilities to represent an agent's subjective degrees of belief in the hypotheses. Serious questions can be raised as to whether agents always have precise degrees of belief. But even if they did, the relevance of such prior probabilities to scientific inquiry would be questionable. If two agents have different priors, how are they to reach some agreement about which is more adequate? If they are to discuss the hypotheses under consideration, they must be able to anchor their probability assignments to something objective (or, at least, intersubjective).

Another Bayesian reaction to the problem of priors has been to argue that they are objectively determined by some *a priori* consideration. Carnap (1950) looked to the structure of the scientist's language as a source of logically defined probabilities. But since scientists can expand or contract their languages at will, it seems implausible that this strategy will be successful. More recently, Rosenkrantz (1977), building on the work of Jaynes (1968), has argued that prior probabilities can be assigned *a priori* by appeal to the requirement that a correct prior should be invariant under certain transformations of how the variables are defined. I will not discuss this line of argument here, except to note that I do not think it works.[2] Prior probabilities, I will assume, are not assignable *a priori*.

I am not a Bayesian, in the sense that I do not think that prior probabilities are always available. But the Bayesian biconditional stated above is nonetheless something I find useful. It is a convenient reminder that hypothesis evaluation must take account of likelihoods and also of the hypotheses' antecedent plausibility. Only sometimes will the latter concept be interpretable as a probability.

Notice that the Bayesian biconditional does not use the word 'explanation'. Explanations have likelihoods; and sometimes they even have priors. This means that they can be evaluated for their overall plausibility. But there is no *sui generis* virtue called 'explanatoriness' that affects plausibility.[3] Likewise, Bayes's theorem enshrines no distinction between induction and abduction. The hypotheses may be inductive generalizations couched in the same vocabulary as the observations; or the hypotheses may exploit a theoretical vocabulary that denotes items not mentioned in the description of the observational evidence.[4] Bayesianism explains why the expression 'inference to the best explanation' can be doubly misleading.

Not only does the Bayesian biconditional make no mention of 'explanatoriness'; it also fails to mention the other epistemic virtues that philosophers like to cite. Parsimony, generality, fecundity, familiarity – all are virtues that do not speak their names. Just as Bayesianism suggests that explanatoriness is not a *sui generis* consideration in hypothesis evaluation, it also suggests that parsimony is not a scientific end in itself. When parsimoniousness augments a hypothesis' likelihood or its prior probability, well and good. But parsimony, in and of itself, cannot make one hypothesis more plausible than another.

To this it may be objected that scientists themselves frequently appeal to parsimony to justify their choice of hypotheses. Since science is a paradigm (perhaps *the* paradigm) of rationality, the objection contin-

ues, does not this mean that a theory's parsimoniousness must contribute to its plausibility? How much of twentieth-century discussion of simplicity and parsimony has been driven by Einstein's remark in his 1905 paper that his theory renders superfluous the introduction of a luminiferous aether? Removing the principle of parsimony from the organon of scientific method threatens to deprive science of its results.

This objection misunderstands my thesis. I do not claim that parsimony never counts. I claim that when it counts, it counts because it reflects something more fundamental. In particular, I believe that philosophers have hypostatized parsimony. When a scientist uses the idea, it has meaning only because it is embedded in a very specific context of inquiry. Only because of a set of background assumptions does parsimony connect with plausibility in a particular research problem. What makes parsimony reasonable in one context therefore may have nothing in common with why it matters in another. The philosopher's mistake is to think that there is a single global principle that spans diverse scientific subject matters.

My reasons for thinking this fall into two categories. First, there is the general framework I find useful for thinking about scientific inference. Probabilities are not obtainable *a priori*. If the importance of parsimony is to be reflected in a Bayesian framework, it must be linked either with the likelihoods or with the priors of the competing hypotheses. The existence of this linkage is always a contingent matter that exists because some set of *a posteriori* propositions governs the context of inquiry. The second sort of reason has to do with how I understand the specific uses that scientists have made of the principle of parsimony. These case studies also suggest that there is no such thing as an *a priori* and subject matter invariant principle of parsimony.

The idea that parsimony is not a *sui generis* epistemic virtue is hardly new. Popper (1959) claims that simplicity reflects falsifiability. Jeffreys (1957) and Quine (1966) suggest that simplicity reflects high probability. Rosenkrantz (1977) seeks to explain the relevance of parsimony in a Bayesian framework. A timeslice of my former self argued that simplicity reduces to a kind of question-relative informativeness (Sober 1975).

What is perhaps more novel in my proposal is the idea that parsimony be understood locally, not globally. All the theories just mentioned attempt to define and justify the principle of parsimony by appeal to logical and mathematical features of the competing hypotheses.

An exclusive focus on these features of hypotheses is inevitable, if one hopes to describe the principle of parsimony as applying across entirely disjoint subject matters. If the parsimoniousness of hypotheses in physics turns on the same features that determine the parsimoniousness of hypotheses in biology, what could determine parsimoniousness besides logic and mathematics? If a justification for this globally defined concept of parsimony is to be obtained, it will come from considerations in logic and mathematics. Understanding parsimony as a global constraint on inquiry thus leads naturally to the idea that it is *a priori* justified. My local approach entails that the legitimacy of parsimony stands or falls, in a particular research context, on subject matter specific (and *a posteriori*) considerations.[5]

In what follows I will discuss two examples of how appeals to parsimony have figured in recent evolutionary biology. The first is George C. Williams' use in his landmark book *Adaptation and Natural Selection* of a parsimony argument to criticize hypotheses of group selection. The second is the use made by cladists and many other systematic biologists of a parsimony criterion to reconstruct phylogenetic relationships among taxa from facts about their similarities and differences.

Williams' (1966) parsimony argument against group selection encountered almost no opposition in the evolution community. Group selection hypotheses were said to be less parsimonious than lower-level selection hypotheses, but no one seems to have asked why the greater parsimony of the latter was any reason to accept them as true.

Cladistic parsimony, on the other hand, has been criticized and debated intensively for the last twenty years. Many biologists have asserted that this inference principle assumes that evolution proceeds parsimoniously and have hastened to add that there is ample evidence that evolution does no such thing. Cladists have replied to these criticisms and the fires continue to blaze.

My own view is that it is perfectly legitimate, in both cases, to ask why parsimony is connected with plausibility. I will try to reconstruct the kind of answer that might be given in the case of the group selection issue. I also will discuss the way this question can be investigated in the case of phylogenetic inference.

I noted earlier that the Bayesian biconditional suggests two avenues by which parsimony may impinge on plausibility. It may affect the prior probabilities and it may affect the likelihoods. The first biological example takes the first route, while the second takes the second.

## 2. PARSIMONY AND THE UNITS OF SELECTION CONTROVERSY

Williams' 1966 book renewed contact between two disciplinary orientations in evolutionary biology that should have been communicating, but did not, at least not very much. Since the 1930s, population geneticists – pre-eminently Fisher (1930), Haldane (1932), and Wright (1945) – had been rather sceptical of the idea that there are group adaptations. A group adaptation is a characteristic that exists because it benefits the group in which it is found. Evolutionists have used the word 'altruism' to label characteristics that are disadvantageous for the organisms possessing them, though advantageous to the group. Population geneticists generally agreed that it is very difficult to get altruistic characteristics to evolve and be retained in a population. Field naturalists, on the other hand, often thought that characteristics observed in nature are good for the group though bad for the individuals. These field naturalists paid little attention to the quantitative models that the geneticists developed. These contradictory orientations coexisted for some thirty years.

Williams (1966) elaborated the reigning orthodoxy in population genetics; but he did so in English prose, without recourse to mathematical arguments. He argued that hypotheses of group adaptation and group selection are often products of sloppy thinking. A properly rigorous Darwinism should cast the concept of group adaptation on the same rubbish heap onto which Lamarckism had earlier been discarded.

Williams deployed a variety of arguments, some better than others. One prominent argument was that group selection hypotheses are less parsimonious than hypotheses that claim that the unit of selection is the individual or the gene.

This argument begins with the observation that 'adaptation is an onerous principle', one that a scientist should invoke only if driven to it. Flying fish return to the water after sailing over the waves. Why do they do this? Williams claims that there is no need to tell an adaptationist story. The mere fact that fish are heavier than air accounts for the fact that what goes up must come down. The thought that flying fish evolved a specific adaptation for returning to the water, Williams concludes, is unparsimonious and so should be rejected.

This idea – that it is more parsimonious to think of the fish's return to the water as a 'physical inevitability' rather than as an adaptation – is only part of the way the principle of parsimony applies to evolutionary explanations. Williams invokes Lloyd Morgan's rule that lower-level explanations are preferable to higher-level ones; Williams takes this to

mean that it is better to think of a characteristic as having evolved for the good of the organism possessing it than to view it as having evolved for the good of the group. The principle of parsimony generates a hierarchy: purely physical explanations are preferable to adaptationist explanations, and hypotheses positing lower-level adaptations are preferable to ones that postulate adaptations at higher levels of organization.

Before explaining in more detail what Williams had in mind about competing units of selection, a comment on his flying fish is in order. I want to suggest that parsimony is entirely irrelevant to this example. If flying fish return to the water because they are *heavier than air,* then it is fairly clear why an adaptationist story will be implausible. Natural selection requires variation. If being heavier than air were an adaptation, then some ancestral population must have included organisms that were heavier than air and ones that were lighter. Since there is ample room to doubt that this was ever the case, we can safely discard the idea that being heavier than air is an adaptation. My point is that this reasoning is grounded in a fact about how natural selection proceeds and a plausible assumption about the character of ancestral populations. There is no need to invoke parsimony to make this point; Ockham's razor can safely be razored away.

Turning now to the difference between lower-level and higher-level hypotheses of adaptation, let me give an example of how Williams' argument proceeds. Musk oxen form a circle when attacked by wolves, with the adult males on the outside facing the attack and the females and young protected in the interior. Males therefore protect females and young to which they are not related. Apparently, this characteristic is good for the group, but deleterious for the individuals possessing it. A group selection explanation would maintain that this wagon-training behaviour evolved because groups competed against other groups. Groups that wagon train go extinct less often and found more daughter colonies than groups that do not.

Williams rejected this hypothesis of group adaptation. He proposes the following alternative. In general, when a predator attacks a prey organism, the prey can either fight or flee. Selection working for the good of the organism will equip an organism with optimal behaviours, given the nature of the threatening predator. If the threat comes from a predator that is relatively small and harmless, the prey is better off standing its ground. If the threat is posed by a large and dangerous predator, the prey is better off running away. A prediction of this idea is that there are some predators that cause large prey to fight and small prey to flee. Williams proposes that wolves fall in this size range; they

make the male oxen stand their ground and the females and young flee to the interior. The group characteristic of wagon-training is just a statistical consequence of each organism's doing what is in its own self-interest. No need for group selection here; the more parsimonious individual-selection story suffices to explain.

Williams' book repeatedly deploys this pattern of reasoning. He describes some characteristic found in nature and the group selection explanation that some biologist has proposed for it. Williams then suggests that the characteristic can be explained purely in terms of a lower-level selection hypothesis. Rather than suspending judgment about which explanation is more plausible, Williams opts for the lower-level story, on the grounds that it is more parsimonious.

Why should the greater parsimony of a lower-level selection hypothesis make that hypothesis more plausible than an explanation in terms of group selection? Williams does not address this admittedly philosophical question. I propose the following reconstruction of Williams' argument. I believe that it is the best that can be done for it; in addition, I think that it is none too bad.

Williams suggests that the hypothesis of group selection, if true, would explain the observations, and that the same is true for the hypothesis of individual selection that he invents. This means, within the format provided by the Bayesian biconditional of the previous section, that the two hypotheses have identical likelihoods. If so, the hypotheses will differ in overall plausibility only if they have different priors. Why think that it is antecedently less probable that a characteristic has evolved by group selection than that it evolved by individual selection?

As noted earlier, an altruistic characteristic is one that is bad for the organism possessing it, but good for the group in which it occurs. Here good and bad are calculated in the currency of fitness – survival and reproductive success.[6] This definition of altruism is illustrated in the fitness functions depicted in Figure 1.1 (Essay 1). An organism is better off being selfish (S) than altruistic (A), no matter what sort of group it inhabits. Let us suppose that the fitness of a group is measured by the average fitness of the organisms in the group; this is represented in the figure by $\bar{w}$. If so, groups with higher concentrations of altruists are fitter than groups with lower concentrations.

What will happen if S and A evolve within the confines of a single population? With some modest further assumptions (e.g., that the traits are heritable), we may say that the population will evolve to eliminate altruism, no matter what initial composition the population happens to have.

For altruism to evolve and be maintained by group selection, there must be variation among groups. An ensemble of populations must be postulated, each with its own local frequency of altruism. Groups in which altruism is common must do better than groups in which altruism is rare.

To make this concrete, let us suppose that a group will fission into a number of daughter colonies once it reaches a certain census size. Suppose that this critical mass is 500 individuals and that the group will then divide into 50 offspring colonies containing 10 individuals each. Groups with higher values of $\bar{w}$ will reach this fission point more quickly, and so will have more offspring; they are fitter. In addition to this rule about colonization, suppose that groups run higher risks of extinction the more saturated they are with selfishness. These two assumptions about colonization and extinction ground the idea that altruistic groups are fitter than selfish ones – they are more reproductively successful (i.e., found more colonies) and they have better chances of surviving (avoiding extinction).

So far I have described how group and individual fitnesses are related, and the mechanism by which new groups are founded. Is that enough to allow the altruistic character to evolve? No it is not. I have omitted the crucial ingredient of *time*.

Suppose we begin with a number of groups, each with its local mix of altruism and selfishness. If each group holds together for a sufficient length of time, selfishness will replace altruism within it. Each group, as Dawkins (1976) once said, is subject to 'subversion from within'. If the groups hold together for too long, altruism will disappear before the groups have a chance to reproduce. This means that altruism cannot evolve if group reproduction happens much more slowly than individual reproduction.

I have provided a sketch of how altruism can evolve by group selection. One might say that it is a 'complicated' process, but this is not why such hypotheses are implausible. Meiosis and random genetic drift also may be 'complicated' in their way, but that is no basis for supposing that they rarely occur. The rational kernel of Williams' parsimony argument is that the evolution of altruism by group selection requires a number of restrictive assumptions about population structure. Not only must there be sufficient variation among groups, but rates of colonization and extinction must be sufficiently high. Other conditions are required as well. This coincidence of factors is not impossible; indeed, Williams concedes that at least one well documented case has been found (the evolution of the *t*-allele in the house mouse). Williams' parsi-

mony argument is at bottom the thesis that natural systems rarely exemplify the suite of biological properties needed for altruism to evolve by group selection.[7]

Returning to the Bayesian biconditional, we may take Williams to be saying that the prior probability of a group selection hypothesis is lower than the prior probability of an hypothesis of individual selection. Think of the biologist as selecting at random a characteristic found in some natural population (like musk oxen wagon-training). Some characteristics may have evolved by group selection, others by lower-level varieties of selection. In assigning a lower prior probability to the group selection hypothesis, Williams is making a biological judgment about the relative frequency of certain population structures in nature.

In the ten years following Williams' book, a number of evolutionists investigated the question of group selection from a theoretical point of view. That is, they did not go to nature searching for altruistic characteristics; rather, they invented mathematical models for describing how altruism might evolve. The goal was to discover the range of parameter values within which an altruistic character can increase in frequency and then be maintained. These inquiries, critically reviewed in Wade (1978), uniformly concluded that altruism can evolve only within a narrow range of parameter values. The word 'parsimony' is not prominent in this series of investigations; but these biologists, I believe, were fleshing out the parsimony argument that Williams had earlier constructed.

If one accepts Williams' picture of the relative frequency of conditions favourable for the evolution of altruism, it is quite reasonable to assign group selection explanations a low prior probability. But this assignment cuts no ice, once a natural system is observed to exhibit the population structure required for altruism to evolve. Wilson (1980) and others have argued that such conditions are exhibited in numerous species of insects. Seeing Williams' parsimony argument as an argument about prior probabilities helps explain why the argument is relevant *prima facie,* though it does not prejudge the upshot of more detailed investigations of specific natural systems.

Almost no one any longer believes the principle of indifference (a.k.a. the principle of insufficient reason). This principle says that if $P_1$, $P_2$, . . . $P_n$ are exclusive and exhaustive propositions, and you have no more reason to think one of them true than you have for any of the others, you should assign them equal probabilities. The principle quickly leads to contradiction, since the space of alternatives can be partitioned in different ways. The familiar lesson is that probabilities cannot be ex-

tracted from ignorance alone, but require substantive assumptions about the world.

It is interesting to note how this standard philosophical idea conflicts with a common conception of how parsimony functions in hypothesis evaluation. The thought is that parsimony considerations allow us to assign prior probabilities and that the use of parsimony is 'purely methodological', presupposing nothing substantive about the way the world is. The resolution of this contradiction comes with realizing that whenever parsimony considerations generate prior probabilities for competing hypotheses,[8] the use of parsimony cannot be purely methodological.

## 3. PARSIMONY AND PHYLOGENETIC INFERENCE

The usual philosophical picture of how parsimony impinges on hypothesis evaluation is of several hypotheses that are each consistent with the evidence,[9] or explain it equally well, or are equally supported by it. Parsimony is then invoked as a further consideration. The example I will now discuss – the use of a parsimony criterion in phylogenetic inference – is a useful corrective to this limited view. In this instance, parsimony considerations are said to affect how well supported a hypothesis is by the data. In terms of the Bayesian biconditional, parsimony is relevant because of its impact on likelihoods, not because it affects priors.

Although parsimony considerations arguably have been implicit in much work that seeks to infer phylogenetic relationships among species from facts concerning their similarity and differences, it was not until the 1960s that the principle was explicitly formulated. Edwards and Cavalli-Sforza (1963, 1964), two statistically minded evolutionists, put it this way: 'the most plausible estimate of the evolutionary tree is that which invokes the minimum net amount of evolution'. They claim that the principle has intuitive appeal, but concede that its presuppositions are none too clear. At about the same time, Willi Hennig's (1966) book appeared in English; this translated an expanded version of his German work of 1950. Although Hennig never used the word 'parsimony', his claims concerning how similarities and differences among taxa provide evidence about their phylogenetic relationships are basically equivalent to the parsimony idea. Hennig's followers, who came to be called 'cladists,' used the term 'parsimony' and became that concept's principal champions in systematics.

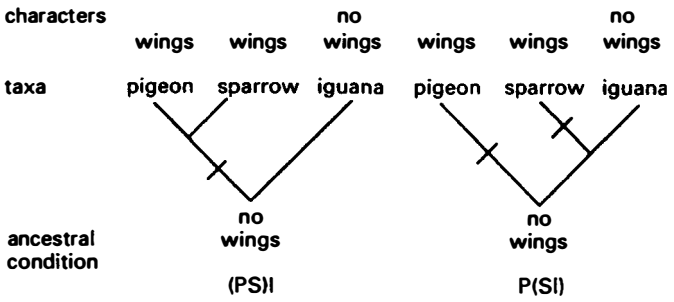


Figure 7.1

Sparrows and pigeons have wings, whereas iguanas do not. That fact about similarity and difference seems to provide evidence that sparrows and pigeons are more closely related to each other than either is to iguanas. But why should this be so?

Figure 7.1 represents two phylogenetic hypotheses and the distribution of characters that needs to be explained. Each hypothesis says that two of the taxa are more closely related to each other than either is to the third. The inference problem I want to explore involves two evolutionary assumptions. Let us assume that the three taxa, if we trace them back far enough, share a common ancestor. In addition, let us assume that this ancestor did not have wings. That is, I am supposing that having wings is the derived (apomorphic) condition and lacking wings is the ancestral (plesiomorphic) state.[10]

According to each tree, the common ancestor lacked wings. Then, in the course of the branching process, the character must have changed to yield the distribution displayed at the tips. What is the minimum number of changes that would allow the (PS)I tree to generate this distribution? The answer is *one*. The (PS)I tree is consistent with the supposition that pigeons and sparrows obtained their wings from a common ancestor; the similarity might be a homology. The idea that there was a single evolutionary change is represented in the (PS)I tree by a single slash mark across the relevant branch.

Matters are different when we consider the P(SI) hypothesis. For this phylogenetic tree to generate the data found at the tips, at least two changes in character state are needed. I have drawn two slash marks in the P(SI) tree to indicate where these might have occurred. According to this tree, the similarity between pigeons and sparrows cannot be a homology, but must be the result of independent origination. The term for this is 'homoplasy'.

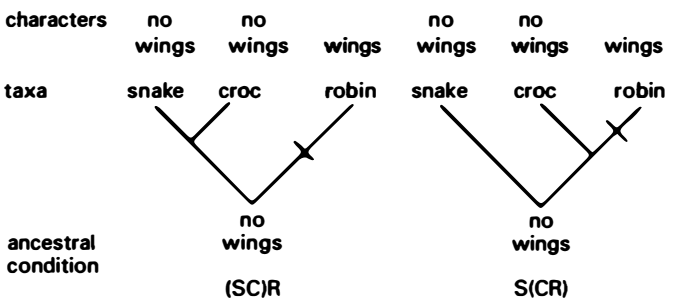


Figure 7.2

The principle of parsimony judges that the character distribution just mentioned supports (PS)I better than it supports P(SI). The reason is that the latter hypothesis requires at least two evolutionary changes to explain the data, whereas the former requires only one. The principle of parsimony says that we should minimize assumptions of homoplasy.

The similarity uniting pigeons and sparrows in this example is a *derived* similarity. Pigeons and sparrows are assumed to share a characteristic that was *not* present in the common ancestor of the three taxa under consideration. This fact about the example is important, because the principle of phylogenetic parsimony entails that some similarities do *not* count as evidence of common ancestry. When two taxa share an *ancestral* character, parsimony judges that fact to be devoid of evidential meaning.

To see why, consider the fact that snakes and crocodiles lack wings, whereas robins possess them. Does the principle of parsimony judge that to be evidence that snakes and crocodiles are more closely related to each other than either are to robins? The answer is *no,* because evolutionists assume that the common ancestor of the three taxa did not have wings. The (SC)R tree displayed in Figure 7.2 can account for this character distribution by assuming a single change; the same is true for the S(CR) tree.

In summary, the idea of phylogenetic parsimony boils down to two principles about evidence:

Derived similarity *is* evidence of propinquity of descent.
Ancestral similarity is *not* evidence of propinquity of descent.

It should be clear from this that those who believe in parsimony will reject *overall* similarity as an indicator of phylogenetic relationships.

For the last twenty years, there has been an acrimonious debate in the biological literature concerning which of these approaches is correct.

Why should derived similarities be taken to provide evidence of common ancestry, as parsimony maintains? If multiple originations were impossible, the answer would be clear. However, no one is prepared to make this process assumption. Less stringently, we can say that if evolutionary change were very improbable, then it would be clear why a hypothesis requiring two changes in state is inferior to a hypothesis requiring only one. But this assumption also is problematic. Defenders of parsimony have been loath to make it, and critics of parsimony have been quick to point out that there is considerable evidence against it. Felsenstein (1983), for example, has noted that it is quite common for the most parsimonious hypothesis obtained for a given data set to require multiple changes on a large percentage of the characters it explains.

I have mentioned two possible assumptions, each of which would *suffice* to explain why shared derived characters are evidence of common ancestry. No one has shown that either of these assumptions is *necessary,* though critics of parsimony frequently assert that parsimony assumes that evolutionary change is rare or improbable.

There is a logical point about the suggestions just considered that needs to be emphasized. A phylogenetic hypothesis, all by itself, does not tell you whether a given character distribution is to be expected or not. On the other hand, if we append the assumption that multiple change is impossible or improbable, then the different hypotheses do make different predictions about the data. By assuming that multiple changes are improbable, it is clear why one phylogenetic hypothesis does a better job of explaining a derived similarity than its competitors. The operative idea here is *likelihood:* If one hypothesis says that the data were hardly to be expected, whereas a second says that the data were to be expected, it is the second that is better supported. The logical point is that *phylogenetic hypotheses are able to say how probable the observations are only if we append further assumptions about character evolution.*

Figure 7.3 allows this point to be depicted schematically. The problem is to infer which two of the three taxa *A, B,* and *C* share a common ancestor apart from the third. Each character comes in two states, denoted by '0' or '1'. It is assumed that 0 is the ancestral form. Character I involves a derived similarity that *A* and *B* possess but *C* lacks.

The branches in each figure are labelled. With the $i$th branch, we can associate a transition probability $e_i$ and a transition probability $r_i$. The

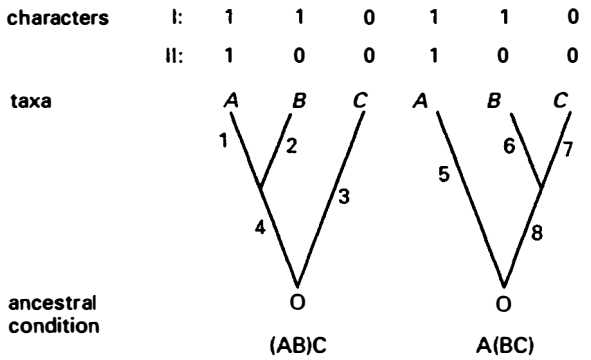


Figure 7.3

former is the probability that a branch will end in state 1, if it begins in state 0; the latter is the probability that the branch will end in state 0, if it begins in state 1. The letters are chosen as mnemonics for 'evolution' and 'reversal'.[11]

Given this array of branch transition probabilities, we can write down an expression that represents the probability of obtaining a given character distribution at the tips, conditional on each of the hypotheses. That is, we can write down an expression full of $e$'s and $r$'s that represents P[110/$(AB)C$] and an expression that represents P[110/$A$(BC)). Which of these will be larger?

If we assume that changes on all branches have a probability of 0.5, then the two hypotheses have identical likelihoods. That is, under this assumption about character evolution, the character distribution fails to discriminate between the hypotheses; it is evidentially meaningless. On the other hand, if we assume that change is very improbable on branches, 1, 2, and 4, but is not so improbable on branches 5 and 6, we will obtain the result that the first character supports $(AB)C$ *less well* than it supports $A(BC)$. This paradoxical result flies in the face of what parsimony maintains (and contradicts the dictates of overall similarity as well). As a third example, we might assign the $e$ and $r$ parameters associated with each branch a value of 0.1. In this case, $(AB)C$ turns out to be more likely than $A(BC)$, relative to the 110 distribution.

With different assumptions about branch transition probabilities, we get different likelihoods for the two hypotheses relative to character I. A similar result obtains if we ask about the evidential significance of character II, in which there is an ancestral similarity uniting $B$ and $C$ apart from $A$. Parsimony judges such similarities to be evidentially

meaningless, but whether the likelihoods of the two hypotheses are identical depends on the assignment of branch transition probabilities.

If parsimony is the right way to determine which hypothesis is best supported by the data, this will be because the most parsimonious hypothesis is the hypothesis of maximum likelihood. Whether this is so depends on the model of character evolution one adopts. Such a model will inevitably rest on biological assumptions about evolution. One hopes that these assumptions will be plausible and maybe even commonsensical. But that they are biology, not pure logic or mathematics, is, I think, beyond question.

I will not take the space here to explain the results of my own investigation (in Sober 1988b) into the connection between phylogenetic parsimony and likelihood.[12] The more general philosophical point I want to make is this: If parsimony is the right method to use in phylogenetic inference, this will be because of specific facts about the phylogenetic process. The method does not have an *a priori* and subject matter neutral justification.

By now it is an utterly familiar philosophical point that a scientific hypothesis ($H$) has implications (whether deductive or probabilistic) about observations ($O$) only in the context of a set of auxiliary assumptions ($A$). Sometimes this is called Duhem's thesis; sometimes it is taken to be too obvious to be worth naming. It is wrong to think that $H$ makes predictions about $O$; it is the conjunction $H\&A$ that issues in testable consequences.

It is a small step from this standard idea to the realization that quality of explanation must be a three-place relation, not a binary one. If one asks whether one hypothesis ($H_1$) provides a better explanation of the observations ($O$) than another hypothesis ($H_2$) does, it is wrong to think this is a matter of comparing how $H_1$ relates to $O$ with how $H_2$ relates to $O$. Whether $H_1$ is better or worse as an explanation of $O$ depends on further auxiliary assumptions $A$.

Why should a more parsimonious explanation of some observation be better than one that is less parsimonious? This cannot be a matter of the logical relationship between hypothesis and observation alone; it must crucially involve auxiliary assumptions.[13]

## 4. CONCLUDING REMARKS

Philosophical discussion of simplicity and parsimony is replete with remarks to the effect that smooth curves are better than bumpy ones and that postulating fewer entities or processes is better than postulat-

ing more. The natural philosophical goal of generality has encouraged the idea that these are methodological maxims that apply to all scientific subject matters. If this were so, then whatever justification such maxims possess would derive from logic and mathematics, not from anything specific to a single scientific research problem.

Respect for the results of science then leads one to assume that general principles of simplicity and parsimony must be justified. The question is where the global justification is to be found; philosophers have been quite inventive in generating interesting proposals. As a fallback position, one could announce, if such proposals fail, that simplicity and parsimony are *sui generis* constituents of the habit of mind we call 'scientific'. According to this gambit, it is part of what we mean by science that simpler and more parsimonious hypotheses are scientifically preferable. Shades of Strawson on induction.

Aristotle accused Plato of hypostatizing The Good. What do a good general and a good flute player have in common? Aristotle argued that the characteristics that make for a good military commander need have nothing in common with the traits that make for a good musician. We are misled by the common term if we think that there must be some property that both possess, in virtue of which each is good.

Williams argued that we should prefer lower-level selection hypotheses over group selection hypotheses, since the former are more parsimonious. Hennig and his followers argued that we should prefer hypotheses requiring fewer homoplasies over ones that require more, since the former are parsimonious. Following Aristotle, we should hesitate to conclude that if Williams and Hennig are right, then there must be some single property of parsimonious hypotheses in virtue of which they are good.

Maxims like Ockham's razor have their point. But their force derives from the specific context of inquiry in which they are wielded. Even if one is not a Bayesian, the Bayesian biconditional provides a useful reminder of how parsimony may affect hypothesis evaluation. Scientists may assign more parsimonious hypotheses a higher antecedent plausibility; but just as prior probabilities cannot be generated from ignorance, so assignments of prior plausibility must be justified by concrete assumptions if they are to be justified at all. Alternatively, scientists may assert that more parsimonious hypotheses provide better explanations of the data; but just as scientific hypotheses standardly possess likelihoods only because of an assumed model connecting hypothesis to data, so assessments of explanatory power also must be justified by concrete assumptions if they are to be justified at all.[14]

By suggesting that we razor Ockham's razor, am I wielding the very instrument I suggest we abandon? I think not. It is not abstract numerology -- a formal preference for less over more – that motivates my conclusion. Rather, the implausibility of postulating a global criterion has two sources. First, there are the 'data'; close attention to the details of how scientific inference proceeds in well-defined contexts of inquiry suggests that parsimony and plausibility are connected only because some local background assumptions are in play. Second, there is a more general framework according to which the evidential connection between observation and hypothesis cannot be mediated by logic and mathematics alone.[15]

Admittedly, two is not a very large sample size. Perhaps, then, this paper should be understood to provide a *prima facie* argument for concluding that the justification of parsimony must be local and subject matter specific. This sort of circumspection is further encouraged by the fact that many foundational problems still remain concerning scientific inference. I used the Bayesian biconditional while eschewing Bayesianism; I offered no alternative doctrine of comparable scope. Nonetheless, I hope that I have provided some grounds for thinking that razoring the razor may make sense.

## NOTES

1. Although I agree with Salmon (1984) that a true explanation can be such that the *explanans* proposition says that the *explanandum* proposition had low probability, I nonetheless think that the explanatory power of a candidate hypothesis is influenced by how probable it says the *explanandum* is. See Sober (1987) for further discussion. It also is worth noting that philosophical discussion of explanation has paid little attention to the question of what makes one explanation a better explanation than another. Hempel's problem leads one to seek a yes/no criterion for being an explanation, or for being an ideally complete explanation. It is another matter to search for criteria by which one hypothesis is a better explanation than another.
2. See Seidenfeld's (1979) review of Rosenkrantz's book for some powerful objections to objective Bayesianism. Rosenkrantz (1979) is a reply.
3. Here I find myself in agreement with Van Fraassen (1980), 22.
4. This is why a Bayesian model of theory testing counts against Van Fraassen's (1980) constructive empiricism. According to Van Fraassen, the appropriate epistemic attitude to take towards a hypothesis depends on what the hypothesis is about. If it is strictly about observables, it is a legitimate scientific task to say whether the hypothesis is true or false. If it is at least partly about unobservables, science should not pronounce on this issue.

These strictures find no expression in the Bayesian biconditional. I discuss the implications of the present view of confirmation for the realism/empiricism debate in Essay 6.

5. Miller (1987) also develops a local approach to confirmational issues, but within a framework less friendly to the usefulness of Bayesian ideas.
6. I will ignore the way the concept of inclusive fitness affects the appropriate definition of altruism and, indirectly, of group selection. I discuss this in Sober (1984, 1988c).
7. I believe that this reconstruction of Williams' parsimony argument is more adequate than the ones I suggest in Sober (1981, 1984).
8. Philosophers have sometimes discussed examples of competing hypotheses that bear implication relations to each other. Popper (1959) talks about the relative simplicity of the hypothesis that the earth has an elliptical orbit and the hypothesis that it has a circular orbit, where the latter is understood to entail the former. Similarly, Quine (1966) discusses the relative simplicity of an estimate of a parameter that includes one significant digit and a second estimate consistent with the first that includes three. In such cases, saying that one hypothesis has a higher prior than another of course requires no specific assumptions about the empirical subject at hand. However, it is debatable whether these are properly treated as competing hypotheses; and even if they could be so treated, such purely logical and mathematical arguments leave wholly untouched the more standard case in which competing hypotheses do not bear implication relations to each other.
9. Although 'consistency with the data' is often how philosophers describe the way observations can influence a hypothesis' plausibility, it is a sorry explication of that concept. For one thing, consistency is an all or nothing relationship, whereas the support of hypotheses by data is presumably a matter of degree.
10. I will not discuss here the various methods that systematics use to test such assumptions about character polarity, on which see Sober (1988c). Also a fine point that will not affect my conclusions is worth mentioning. The two evolutionary assumptions just mentioned entail, not just that *a* common ancestor of the three taxa lacked wings, but that this was the character state of the three taxa's *most recent common ancestor.* This added assumption is useful for expository purposes, but is dispensable. I do without it in Sober (1988b).
11. The complements of $e_i$ and $r_i$ do not represent the probabilities of stasis. They represent the probability that a branch will end in the same state in which it began. This could be achieved by any even number of flip-flops.
12. I will note, however, that assigning $e_i$ and $r_i$ the same value for all the $i$ branches is implausible, if the probability of change in a branch is a function of the branch's temporal duration. In addition, there is a simplification in the treatment of likelihood presented here that I should note. The $(AB)C$ hypothesis represents a family of trees, each with its own set of

branch durations. This means that the likelihood of the $(AB)C$ hypothesis, given a model of character evolution, is an *average* over all the specific realizations that possess that technology. See Sober (1988b) for details.

13. Of course, by packing the auxiliary assumptions into the hypothesis under test, one can obtain a new case in which the 'hypotheses' have well-defined likelihoods without the need to specify still further auxiliary assumptions. My view is that this logical trick obscures the quite separate status enjoyed by an assumed model and the hypotheses under test.
14. The principal claim of this paper is not that 'parsimony' is ambiguous. I see no reason to say that the term is like 'bank'. Rather, my locality thesis concerns the *justification* for taking parsimony to be a sign of truth.
15. I have tried to develop this thesis about evidence in Sober (1988a) and in Essay 8.

## REFERENCES

Carnap, R. (1950) *Logical Foundations of Probability* (University of Chicago Press).

Dawkins, R. 1976. *The Selfish Gene* (Oxford University Press).

Edwards, A., and Cavalli-Sforza, L. 1963. 'The Reconstruction of Evolution', *Ann. Human Genetics, 27*, 105.

Edwards, A., and Cavalli-Sforza, L. 1964. 'Reconstruction of Evolutionary Trees' in V. Heywood and J. McNeil (eds.), *Phenetic and Phylogenetic Classification* (New York Systematics Association Publications), *6*, 67–76.

Felsenstein, J. 1983. 'Parsimony in Systematics', *Annual Review of Ecology and Systematics, 14*, 313–33.

Fisher, R. 1930. *The Genetical Theory of Natural Selection* (Reprinted; New York: Dover, 1958).

Haldane, J. 1932. *The Causes of Evolution* (Reprinted New York: Cornell University Press, 1966).

Hempel, C. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (New York: Free Press).

Hennig, W. 1966. *Phylogenetic Systematics* (Urbana: University of Illinois Press).

Jaynes, E. 1968. 'Prior Probabilities', *IEEE Trans. Systems Sci. Cybernetics, 4*, 227–41.

Jeffreys, H. 1957. *Scientific Inference* (Cambridge University Press).

Miller, R. 1987. *Fact and Method* (Princeton University Press).

Popper, K. 1959. *The Logic of Scientific Discovery* (London: Hutchinson).

Quine, W. 1966. 'On Simple Theories of a Complex World', in *Ways of Paradox and Other Essays* (New York: Random House).

Rosenkrantz, R. 1977. *Inference, Method and Decision* (Dordrecht: D. Reidel).

Rosenkrantz, R. 1979. 'Bayesian Theory Appraisal: A Reply to Seidenfeld', *Theory and Decision, 11*, 441–51.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World* (Princeton University Press).

Seidenfeld, T. 1979. 'Why I am not a Bayesian: Reflections Prompted by Rosenkrantz', *Theory and Decision, 11,* 413–40.

Sober, E. 1975. *Simplicity* (Oxford University Press).

Sober, E. 1981. 'The Principle of Parsimony', *British Journal for the Philosophy of Science, 32,* 145–56.

Sober, E. 1984. *The Nature of Selection* (Cambridge, Mass.: MIT Press).

Sober, E. 1987. 'Explanation and Causation: A Review of Salmon's *Scientific Explanation and the Causal Structure of the World*', *British Journal for the Philosophy of Science, 38,* 243–57.

Sober, E. 1988a. 'Confirmation and Law-Likeness', *Philosophical Review, 97,* 617–26.

Sober, E. 1988b. *Reconstructing the Past: Parsimony, Evolution, and Inference* (Cambridge, Mass.: MIT Press).

Sober, E. 1988c. 'What is Evolutionary Altruism?', *Canadian Journal of Philosophy Supplementary Volume, 14,* 75–99.

Van Fraassen, B. 1980. *The Scientific Image* (Oxford University Press).

Wade, M. 1978. 'A Critical Review of Models of Group Selection', *Quarterly Review of Biology, 53,* 101–14.

Williams, G. C. 1966. *Adaptation and Natural Selection* (Princeton University Press).

Wilson, D. 1980. *The Natural Selection of Populations and Communities* (Menlo Park, CA: Benjamin/Cummings).

Wright, S. 1945. 'Tempo and Mode in Evolution: a Critical Review', *Ecology, 26,* 415–19.