CHAPTER 13

Parsimony and models of animal minds

Elliott Sober

1. INTRODUCTION

Dennett's (1971, 1987) ideas about the intentional stance have elicited different reactions from philosophers and scientists. Philosophers have focused on Dennett's alleged anti-realism about the mental while scientists have focused on his three-way distinction among zero-order, first-order, and second-order intentionality. For most philosophers, anti-realism (at least concerning the mental states of human beings) is a *no-no*; for most cognitive scientists, the three-way distinction is useful. This contrasting reaction is typical of a larger pattern: philosophers are more inclined to cite work with which they disagree while scientists are more inclined to cite work on which they wish to build.

Dennett (1987, 1991b) has denied that he is an anti-realist, but that has not stopped philosophers from continuing to affix a scarlet letter "A" to his work. Nor has the specter of anti-realism stopped cognitive ethologists from using the three-way distinction to articulate a central methodological principle, which they call "the principle of conservatism." According to this principle, hypotheses that explain an organism's behavior by attributing lower-order intentionality are preferable to hypotheses that explain the behavior by attributing higher-order intentionality (see, for example, Cheney and Seyfarth [1990, 2007]).

My subject here is the principle of conservatism. What, exactly, does it say and what is its justification? Cognitive scientists often regard the principle as an instance of a more general methodological maxim, namely the principle of parsimony, *a.k.a.* Ockham's razor. They reason as follows: since Ockham's razor is a sound principle of scientific inference, there is no special question about why the principle of conservatism should be used in cognitive science. My main goal in this essay is to trace how the

My thanks to Martin Barrett, Tom Bontly, Simon Fitzpatrick, Chuck Kalish, Robert Lurz, Carolina Sartorio, Armin Schulz, Larry Shapiro, and Shannon Spalding for useful discussion.

general principle is related to the specific one. This tracing suggests that the principle of conservatism needs to be refined. Connecting the principle in cognitive science to more general questions about scientific inference also will allow us to revisit the question of realism versus instrumentalism. Realist philosophers of science often have no problem with the principle of parsimony. Maybe they should not be so sanguine. Finally, connecting the principle of conservatism to more general inferential issues suggests that the principle can be more than a qualitative tie-breaker. If two explanations fit the observations *equally* well, one is told to prefer the explanation that is more parsimonious. The view of parsimony that I'll describe also allows theories to be compared that fit the data *unequally* well. If a complex theory fits the data better than a simpler theory does, which theory is better overall? Many philosophers think there can be no principled answer to this question; they think it is a matter of taste how much weight you put on simplicity compared with goodness-of-fit. The view of parsimony provided by the part of statistics called "model selection theory" suggests that there is room for skepticism about this conventionalist view.

I have argued in earlier publications that invocations of parsimony in science often should be viewed as expressions of subject-matter-specific background theories (Sober [1988, 2005]); it follows that different invocations in different scientific problems may rest on very different foundations (Sober [1990, 2001]). Thus conceived, the way to understand the use of parsimony in a given scientific domain is to uncover the background theory in play. Fitzpatrick (chapter 14) adopts this strategy to assess the principle of conservatism. This is not the strategy I will pursue here. The framework deployed in model selection theory is very general; it is not specific to the subject matter of any one science (which is not to say that there are no assumptions that must be satisfied for the apparatus to apply). How does that general framework help clarify the principle of conservatism?

2. PRELIMINARIES

The usual definition of orders of intentionality is this: a first-order mental state is a mental state that is not about (the presence or absence of) any mental state, a second-order mental state is a mental state that is about (the presence or absence of) a first-order mental state, a third-order mental state is a mental state about (the presence or absence of) a second-order mental state, and so on. Aboutness is judged by the state's propositional content (if it has one). Consider a dog, Fido, and his master, Louise. If Fido believes that a bone is buried in the backyard, this is a first-order

mental state, since the content of Fido's belief (*that there is a bone buried in the backyard*) does not describe anyone's mental state. If Fido believes that Louise *sees* that he is digging in the backyard, this is a second-order state, since the content of Fido's belief adverts to Louise's mental state. And if Louise believes that Fido *realizes* that she is *watching* him, Louise's state is third-order. By convention, a zeroth-order mental state is a limiting case; it denotes a state in which there is no mentation at all.

It is customary in the literature to say that adult human beings have second-order (and higher) intentionality, that apes and monkeys have at least first-order intentionality though it is controversial whether they have second-order intentionality, and that thermostats have only zeroth-order intentionality. This suggests that we should define how orders of intentionality are assigned to organisms (or to "systems") as follows:

Organism O has nth-order intentionality
$=_{def}$ O has at least one nth-order mental state.

Notice that this definition is consistent with a single individual's having multiple orders of intentionality. It usually is assumed that if an organism has second-order mental states, then it also has first- and zeroth-order states. I will adopt this assumption here.

The definitions and the assumption just mentioned suggest the following simple probability argument. The statement that a system has second-order intentionality is logically stronger than the statement that it has first-order intentionality, since the former entails the latter, but not conversely. It is a consequence of the axioms of probability that logically stronger statements can't be more probable than logically weaker statements. This means that:

$$\begin{aligned}&\Pr(O \text{ has 2nd-order intentionality})\\ &\leq \Pr(O \text{ has 1st-order intentionality})\\ &\leq \Pr(O \text{ has 0th-order intentionality.}^{1}\end{aligned}$$

The axioms of probability also entail that this ordering of probabilities must remain in place regardless of what new observational evidence is obtained:

$$\begin{aligned}&\text{(1a) } \Pr(O \text{ has 2nd-order intentionality}|E)\\ &\leq \Pr(O \text{ has 1st-order intentionality}|E)\\ &\leq \Pr(O \text{ has 0th-order intentionality}|E), \text{ for any evidence } E.\end{aligned}$$

[1] These inequalities will be strict if $\Pr[O$ has (n+1)th-order intentionality$|O$ has nth-order intentionality$] < 1$.

Is this enough to justify C. Lloyd Morgan's famous "canon"? Morgan (1894) describes his principle as follows:

> 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. (p. 53)

Even if Morgan's "higher" and "lower" are taken to correspond to the orders of intentionality just defined, the answer is *no*. As (1a) makes clear, no evidence can ever make it more probable that an organism has (at least) second-order intentionality than that it has (at least) first-order intentionality. In just the same way, the axioms of probability entail that

$$\begin{aligned}&\Pr(\text{there are at least two apples in the basket})\\ &\quad\leq \Pr(\text{there is at least one apple in the basket})\end{aligned}$$

and

(1b) $\Pr(\text{there are at least two apples in the basket}|E)$
$\leq \Pr(\text{there is at least one apple in the basket}|E)$, for any evidence E.

Proposition (1a) doesn't capture what Morgan is after since Morgan thinks that evidence *can* sometimes justify the attribution of "higher psychical faculties."[2] The same goes for modern cognitive scientists and how they understand their principle of conservatism.

The thought behind these principles is that we should adopt hypotheses that attribute higher-level intentionality to an organism only if the data force us to do so; in the absence of such data, we should assume that the organism has only lower-level intentionality. This proposal is independent of the axioms of probability theory because the hypotheses of interest are *incompatible* with each other. A natural Bayesian representation of the principle of conservatism is to formulate it as a claim about the prior probabilities of three incompatible hypotheses:

(2a) $\Pr(O$ 0th-, 1st-, and 2nd-order intentionality)
$< \Pr(O$ has 0th- and 1st-order intentionality only)
$< \Pr(O$ has 0th-order intentionality only).

[2] In the book's second edition, Morgan (1903) restates the canon and adds: "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" (p. 59).

This inequality is no more a consequence of the axioms of probability than is the following claim about apples:

(2b) Pr(there are exactly two apples in the basket)
< Pr(there is exactly one apple in the basket)
< Pr(there are no apples in the basket).

The ordering of prior probabilities described in proposition (2a) may be revised as new evidence is acquired. The priors embody a "default assumption" that subsequent evidence may displace. In just the same way, proposition (2b) might be a default assumption about a basket – an assumption we make before we have observed it; adopting this assumption does not preclude our obtaining evidence that makes it very probable that the basket contains exactly two apples.

Although (2a) is more in tune with what Morgan and modern cognitive scientists have in mind, there is a catch. What justification does (2a) have? When it comes to apples, we want empirical evidence for a claim like (2b). This evidence might come from frequency data or from a well-confirmed empirical theory. I submit that the same is true for (2a). Morgan attempted to furnish an empirical argument for his canon based on Darwin's theory of evolution, an argument that I think fails (Sober [1998b]). What about frequency data? If we knew that few organisms have second-order intentionality, that more have only first-order, and that still more have only 0th-order intentionality, that would do the trick. But the intent of Morgan's canon and of the principle of conservatism is to provide inferential advice *before* we know any such thing. Nor do we have an empirical theory that provides the needed justification. Perhaps, then, we should regard (2a) as a "primitive postulate." The problem with this approach is the one that Russell (1919, p. 71) described in another context: *it has all the advantages of theft over honest toil.*

3. MODEL SELECTION

Rather than pursue the question of whether Bayesianism is able to explain why the principle of conservatism makes sense, I want to outline some non-Bayesian ideas that have been developed in model selection theory. These, I think, provide an attractive format for characterizing, and for improving upon, what the principle of conservatism says. These ideas are non-Bayesian, in that they do not appeal to the prior or posterior probabilities of the hypotheses considered.

Table 13.1. *Number of individuals with lung cancer out of 1000 in each of four treatment cells*

		Asbestos exposure	
		+	−
Smoking	+	50	30
	−	20	3

Table 13.2. *Probabilities of lung cancer in four treatment cells*

		Asbestos exposure	
		+	−
Smoking	+	$b+a+s+i$	$b+s$
	−	$b+a$	b

I'll begin with a non-psychological example. Suppose you want to model how smoking and asbestos exposure influence lung cancer. Your information is to come, not from some antecedently well-established theory, but from data on how often people in different "treatments" contract the disease. To simplify, let's suppose that smoking and asbestos are dichotomous variables, and that the same is true of lung cancer. Suppose that there are 1,000 people in each "treatment cell" and that the number of individuals in each cell who get the disease is as shown in Table 13.1.

The task is to model how smoking and asbestos affect the probability of lung cancer. The probability in each treatment cell of contracting the disease is represented in Table 13.2: "b" represents the "baseline" probability of lung cancer when an individual does not smoke and is not exposed to asbestos; "s" is the effect of smoking alone; "a" is the effect of asbestos exposure alone; and "i" is an interaction term. Here are some causal models to consider:

(Null) $s = a = i = 0$. The value of b is left open.
(Only smoking) $a = i = 0$. The values of b and s are left open.
(Only asbestos) $s = i = 0$. The values of b and a are left open.
(Two additive causes) $i = 0$. The values of b, a, and s are left open.
(Two interacting causes) The values of b, a, s, and i are left open.[3]

[3] It wouldn't matter to the framework of model selection theory if the parameters in a model that are not set equal to zero were constrained to be non-zero (rather than having their values left entirely open, as above).

The Null model is the simplest one listed; it has a single adjustable parameter (b) and says that smoking and asbestos make no difference to one's risk of lung cancer. The models of Only Smoking and Only Asbestos each have two adjustable parameters, the model of Two Additive Causes has three adjustable parameters, and the model of Two Interactive Causes has four.

How might these models be tested against each other? The fact that all contain adjustable parameters makes it difficult to see what they predict about the data. Model selection theory solves this problem by shifting attention from a model M to the instantiation of the model, $L(M)$, that assigns the free parameters in M the values that maximize the probability of the data. For example, consider the model of Two Interacting Causes and the data in Table 13.1. This model maximizes the probability it assigns to the data if its adjustable parameters are set at $b = 3/1000$, $s = 27/1000$, $a = 17/1000$, and $i = 3/1000$. These are the maximum likelihood estimates of the parameters. This model is able to fit the data perfectly; the maximum likelihood estimates exactly match the frequencies in the data. The other models cannot do this; however, they are simpler.

H. Akaike, the father of model selection theory, introduced two innovations into statistics (Forster and Sober [1994]). The first was the description of a goal that models might be asked to attain; the second was a theorem that throws light on how one might estimate a model's ability to attain that goal. The new goal was predictive accuracy. Rather than asking whether a model is true or probably true, one asks whether it will accurately predict new data when its parameters are fitted to old data. We have seen that the model of Two Interacting Causes can fit the old data perfectly. If four new groups of people are sampled from the same population from which the people in the old data set were drawn, how well will the fitted model predict their frequencies of lung cancer? Akaike's second contribution was a theorem. Akaike (1973) proved, from some surprisingly general assumptions, that

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

Here k is the number of adjustable parameters in M; it measures the model's complexity. This theorem is the basis for a proposed criterion for estimating the predictive accuracy of models. The Akaike Information Criterion (AIC) scores a model by calculating its value for the quantity $\log\{\Pr[data|L(M)]\} - k$. More complex models will have higher values for $\log\{\Pr[data|L(M)]\}$, but they will incur a larger penalty by virtue of

having a larger value for *k*. The point of AIC is to compare models with each other. What matters is not a model's absolute AIC score, but how its score compares with those of other, competing, models. Whether a simpler model has a better score than a complex model depends on the data.

Akaike's theorem shows why the complexity of a model (as measured by the number of adjustable parameters it contains) is not an aesthetic frill; it is relevant to estimating predictive accuracy. The framework that Akaike proposed is interesting for the additional reason that it explains why it makes sense in science to test models that everyone knows are false. Null models frequently have this feature. If the goal were simply to discover which models are probably true, idealized models could be dismissed summarily. However, if the goal is predictive accuracy, it makes sense to test idealized models against each other. Surprisingly, the Akaike framework shows that a model known to be false can sometimes be expected to be more predictively accurate than a model known to be true. AIC embodies an instrumentalist epistemology.

Although philosophers often describe the principle of parsimony as a tie-breaker, AIC and the other model selection criteria discussed in statistics are more than that. They not only entail that the more parsimonious of two models is better when they fit the data equally well; they also indicate how models should be compared when they differ in both simplicity and goodness-of-fit. This resource may come in handy for cognitive scientists who want parsimony to be more than a qualitative and informal criterion.

4. ALTRUISM AND SPITE IN CHIMPANZEES

Silk *et al.* (2005) and Jensen *et al.* (2006) conducted experiments that were designed to discover whether chimps have other-directed preferences or are indifferent to the welfare of others. There are differences between the two studies, but the conclusions are on the same page: the former concludes (p. 1357) that "chimpanzees do not take advantage of opportunities to deliver benefits to familiar individuals at no material cost to themselves," the latter (p. 1013) that "chimps made their choices based solely on personal gain."

All the experiments place a chimp in a situation in which it must choose among actions. Silk *et al.* (2005) studied whether chimps choose to send food to both their own cage and to another cage more often when there is another chimp in the other cage or when the other cage is empty. They found that the frequencies of provisioning both cages in these two settings are not significantly different. Evidently, the chimps care only about

Table 13.3. *Frequencies of four types of behavior*

		Benefit to other	
		+	−
Benefit to self	+	49%	46%
	−	3%	2%

Table 13.4. *Probabilities of four types of behavior*

		Benefit to other	
		+	−
Benefit to self	+	$b+s+a+i$	$b+s$
	−	$b+a$	b

getting food for themselves; the presence or absence of another chimp – a potential recipient of their donation – does not matter. Jensen *et al.* (2006) tested how often actors choose to provide food to both self and other as opposed to providing food only for self. The amount of food that actors obtain for themselves is the same in both cases and there is no more effort involved in choosing "both" rather than "just me." Jensen *et al.* also studied what chimps do when they cannot benefit themselves. Will they provision another chimp? Here again, it appears to be a matter of indifference to the actor what happens to the would-be recipient.

In order to mimic the structure of the lung cancer modeling problem described before, I want to consider the following experiment, which is inspired by Silk *et al.* (2005) and by Jensen *et al.* (2006). The point is not that this is a good experimental design; rather, I want to start exploring how model selection ideas apply in intentional psychology. When it is time for a meal, an actor and a second chimp ("the other") are in facing cages. The actor has the option of producing the four outcomes shown in Table 13.3. Actors can cause food to be provided to both self and other, just to self, just to other, or to neither. The four outcomes are equally easy for actors to achieve and donating to the other chimp does not affect the amount of food that actors obtain for themselves. Suppose the frequencies of these four types of behavior are those given in Table 13.3.

Now let's consider some models of the chimps' behaviors that are expressed in terms of the probabilistic parameters shown in Table 13.4: "b" represents the probability of performing an action that provides food

to neither self nor other; "s" is the probability of performing an action that provides benefit to self; "a" is the probability of performing an action that provides benefit to others; "i" is an interaction term.

(Null)	$s = a = i = 0$. The value of b is left open.
(Pure selfishness)	$a = i = 0$. The values of b and s are left open.
(Pure altruism)	$s = i = 0$. The values of b and a are left open.
(Additive motivational pluralism)	$i = 0$. The values of b, s, and a are left open.
(Interactive motivational pluralism)	The values of b, s, a, and i are left open.

The Null model says that chimps are "nihilists": they care neither about self nor other. The next two models, Pure Selfishness and Pure Altruism, are both monistic models: they say that chimps care only about self or only about others. The last two models are pluralistic: they say that chimps care about both self and other (Sober and Wilson [1998]). As before, the most complex model can achieve perfect fit-to-data. Nonetheless, depending on the frequency data and the sample size, it may turn out that Pure Selfishness is the model that receives the best AIC score.

5. MODEL EVALUATION VERSUS HYPOTHESIS TESTING

Jensen *et al.* (2006) touch on a possibility that can arise in any study, one that I think shows that the comparative and instrumentalist framework of model selection theory is superior to the accept/reject framework of conventional Neyman–Pearson hypothesis testing. Their experiments led them to conclude that chimps do not have other-directed preferences (at least when it comes to food sharing in the kind of circumstance they investigated). This conclusion was based on pooling data from all the chimps in the study. This leaves it open that when chimps are considered one by one (each participated in multiple experiments that each involved a number of trials), the evidence may indicate that some of them have other-directed preferences. Indeed, Jensen *et al.* (2006, p. 1019) say that "two of the six actors showed some possible signs of altruism."[4]

[4] Jensen *et al.* (2006) note that "these individuals were also the only two individuals who begged from, or harassed, the recipients" and speculate that the two chimps who provisioned others may have

This illustrates a paradoxical possibility that can arise in Neyman–Pearson hypothesis testing. You are testing Pure Selfishness against Motivational Pluralism. If you pool your data, you conclude that the chimps are egoists rather than pluralists. But if you consider the chimps one by one, constructing a different pair of models for each, you conclude that some are egoists while others are motivational pluralists. Neyman–Pearson theory sanctions both conclusions. This is odd: how can it make sense to accept egoism and reject pluralism for *all* the chimps but to do the reverse for *some* of them? Shifting to a model selection framework provides a solution to this puzzle. There are two prediction problems you might contemplate. One is predicting a new set of pooled data from the six chimps (or from six new chimps drawn from the same population); the other is predicting the separate outcomes on new experiments on each of six chimps. It is not paradoxical that different models might be better in different prediction tasks.

6. HIGHER AND LOWER

The principle of conservatism, like Morgan's canon, describes a preference concerning *kinds* of parameters, not *numbers* of parameters. It says that a model that postulates only lower-level intentionality is preferable to one that postulates higher-level intentionality if both fit the data equally well. This principle does not care if the lower-level model has a very large number of adjustable parameters while the higher-level model has only a few. It is hard to see how this principle can make sense from the point of view of model selection theory. Consider the lung cancer example. Smoking is one possible cause and asbestos exposure is another; the model that says that only smoking makes a difference is more parsimonious than a model that says that both do; and the two monistic models (Only Smoking and Only Asbestos) are equally parsimonious. Higher and lower kinds of causes don't matter.

Not only does AIC not care about the number of kinds of causes; it also doesn't care about numbers of causes (unless this count is mirrored in the number of adjustable parameters). I suppose that the additive and the interactive models of the causes of lung cancer both postulate two

expected the recipients of their largesse to have given them food. The point the authors are making here pertains to whether these two chimps have ultimate or merely instrumental other-directed preferences (Sober and Wilson [1998]). It does not undercut the conclusion that these chimps have other-directed preferences. The experiments addressed the latter issue and did not address the question of ultimate versus instrumental.

causes, smoking and asbestos exposure. The point is that they differ in their number of adjustable parameters, and that is what matters.

Increasing the number of causes in a model need not increase the number of adjustable parameters; everything depends on how those new causes are modeled. Consider the following example from genetics. Suppose you suspect that the three genotypes (*AA*, *Aa*, *aa*) found at a locus may influence an organism's probability of surviving to adulthood. One of the models you consider has two parameters (b and d); it says that $\Pr(\text{surviving}|aa) = b$, $\Pr(\text{surviving}|Aa) = b + d$, and $\Pr(\text{surviving}|AA) = b + 2d$. You then wish to consider the possibility that n loci, each with two alleles, affect survivorship. It isn't true that every n-locus model must have more than two parameters. Consider a model that says that at each locus, the organism has zero, one, or two "plus" alleles and that the probability of surviving is an additive function of the number of plus alleles: $\Pr(\text{surviving}/i \text{ plus alleles}) = b + id$. There may be n causes of survivorship (the n loci), but there are just two parameters.

If we drop the fixation on "higher" and "lower," a better formulation of the principle of conservatism becomes available: a model that postulates only lower-order intentionality (using n parameters to do so) is better than a model that postulates *both* lower-order intentionality (using n parameters) *and* higher-order intentionality (using m additional parameters) if the two models fit the data equally well. However, if introducing higher-level intentionality permits one to have *fewer* parameters overall while still fitting the data equally well, parsimony will speak in favor of introducing higher-level intentionality. This possibility will be discussed soon.

7. IDENTIFIABILITY

To apply AIC to a model M, there must be a unique maximum likelihood estimate for each of the parameters in M. When this fails to be true, the model is said to be *unidentifiable*. Here's a simple example. Suppose you heat a kettle on your stove to different temperatures and measure how much pressure there is in the kettle at each temperature. You do this n times and display your n observations as n data points in Cartesian coordinates, the x-axis representing temperature, the y-axis representing pressure. You now face a curve-fitting problem. What is the general relationship between temperature and pressure in your kettle? You want to draw a line in the x-y plane. Which line should you draw?

You should consider various models. One of them might be the linear model LIN, which says that $y = mx + b + e$. This model has three

adjustable parameters, the last one being an error term that allows you to represent the possibility that your thermometer and pressure gauge may be subject to error. If you have a large number of data points, there is a single straight line that fits the data best; this is L(LIN). But suppose you have just one data point. There are infinitely many straight lines that pass exactly through this point. They make different predictions about new data. Since there is no such thing as *the* best-fitting straight line, the problem of estimating how accurately LIN will predict new data when fitted to old cannot be addressed. LIN is not identifiable. In general, for a model with n adjustable parameters to be identifiable, you need more than n data points. In practice, scientists recommend that you restrict your evaluation to models that have far fewer parameters than the number of observations you have (see, for example, Burnham and Anderson [2002]).

In the lung cancer example and also in the present example about pressure and temperature, you observe the values of candidate causal variables and also the values of the effect term. However, in cognitive science, you can't observe the beliefs and desires and other mental states that individuals have, though you can observe their behavior. How, then, is model selection theory applicable in this science?

In the experiment I invented on food sharing, you observe the frequencies of four types of meal-time behavior. You don't observe the chimp's preferences. However, this isn't necessary. Rather, for each model M, you need to find a quantitative representation of the preferences allowed by M that renders the observations maximally probable, thereby finding $L(M)$. For example, the model of Pure Selfishness makes the observations maximally probable when it sets $b = 0.02$ and $s = 0.45$. The second of these parameters represents how much chimps prefer receiving food themselves rather than going without. To find L(Pure Selfishness), what is required is not that

$$\Pr(\text{Pure Selfishness } \& \text{ s } = 0.45|\text{data}) \text{ is high}$$

or that

$$\Pr(s = 0.45|\text{Pure Selfishness } \& \text{ data}) \text{ is high,}$$

but only the more modest thesis that

$$\Pr(\text{data}|\text{Pure Selfishness } \& \, s = 0.45) > \Pr(\text{data}|\text{Pure Selfishness } \& \, s = x), \text{ for any } x \neq 0.45.$$

Seeing that this inequality is true does not require that you find the model of Pure Selfishness plausible.

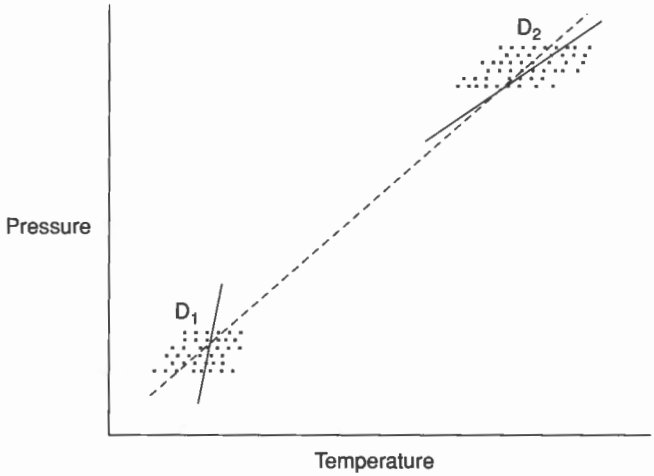

Figure 13.1 The disunified model DIS fits the two data sets D_1 and D_2 better than the unified model UNI does. L(DIS) is depicted by two solid lines, L(UNI) by a single dashed line, and data points by dots.

8. PARSIMONY, UNIFICATION, AND ORDERS OF INTENTIONALITY

The model selection approach to parsimony helps explain why unification is a theoretical virtue. Consider a unified model that applies the same n parameters to multiple data sets and a disunified model that applies a different set of n parameters to each data set. The unified model is more parsimonious, so if the two models fit the data about equally well, a model selection criterion such as AIC will estimate that the unified model can be expected to have greater predictive accuracy (Forster and Sober [1994]).

Figure 13.1 represents a simple example. Suppose the experiment you run on the kettle in your kitchen produces the two data sets D_1 and D_2. Consider the following two models:

(UNI) $y = mx + b + e$.

(DIS) $y = m_1x + b_1 + e_1$ for the first data set.

$y = m_2x + b_2 + e_2$ for the second data set.

The unified model has three adjustable parameters; the disunified model has six. Depending on the models' fit to data, UNI may receive the better AIC score.

This difference between unified and disunified models is the key, I believe, to understanding how the model selection framework applies to

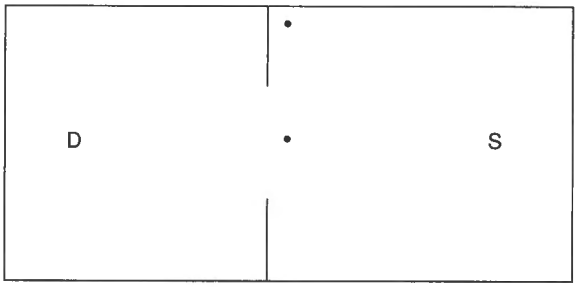

Figure 13.2 The subordinate chimp S can see both food items (represented by dots); the dominant chimp D can see just one. From Hare *et al.* (2000).

Table 13.5. *Frequencies of events in four experiments (Hare et al. [2000, 2001])*

1	$f(S \text{ takes } x \mid \text{an opaque barrier is between } D \text{ and } x) = f_{1a}$ $f(S \text{ takes } x \mid x \text{ is out in the open}) = f_{1b}$	$f_{1a} > f_{1b}$
2	$f(S \text{ takes } x \mid \text{a transparent barrier is between } D \text{ and } x) = f_{2a}$ $f(S \text{ takes } x \mid x \text{ is out in the open}) = f_{2b}$	$f_{2a} \leq f_{2b}$
3	$f(S \text{ takes } x \mid \text{only } S \text{ was present during food placement}) = f_{3a}$ $f(S \text{ takes } x \mid S \text{ and } D \text{ both present during food placement}) = f_{3b}$	$f_{3a} > f_{3b}$
4	$f(S \text{ takes } x \mid S \text{ was present during food placement, } D \text{ was not present though another dominant individual was}) = f_{4a}$ $f(S \text{ takes } x \mid S \textit{ and } D \text{ are both present during food placement}) = f_{4b}$	$f_{4a} > f_{4b}$

S is a subordinate chimp, *D* is a dominant chimp, and *x* is a food item. In all the experiments, *S* and *D* both have the opportunity to try to grab food.

issues about orders of intentionality. Consider, for example, the experiments conducted by Hare *et al.* (2000, 2001) in which a subordinate chimp chooses which food items it will attempt to grab while a dominant chimp is present. In the first experiment (depicted in Figure 13.2), one food item is out in the open where both individuals can see it while the other food item is visible only to the subordinate (because there is an opaque barrier between the dominant and that food item). The result is that subordinates tend to go for the food item that the dominant individual cannot see. Hare *et al.* ran other experiments, accumulating the kind of frequency data represented in Table 13.5. For example, they used a transparent barrier instead of one that is opaque, and found that subordinates do not preferentially go for the object in front of that barrier. They also dispensed with barriers entirely and compared what subordinates do when they alone have watched where food is hidden and what subordinates do when they and a dominant both

watch. Hare *et al.* defend the hypothesis that subordinates decide what to do by forming beliefs about what the dominant chimp has and has not seen. (See Fitzpatrick [chapter 14] for further discussion.)

If you look at these experiments one by one, it isn't hard to invent a first-order explanation for *each*; what is more difficult is inventing a single first-order explanation that works for *all* (Tomasello and Call [2006], p. 371).[5] In contrast, a unified explanation is easy to achieve if you resort to a second-order hypothesis. But so what? The two interpretations *seem* to fit the data equally well and they *seem* not to disagree with each other about any possible observation. And why is the fact that one explanation is unified while the other is disunified epistemologically significant (Heyes [1998])? Things look decidedly different when we view this problem through the lens of model selection theory. In fact, the two models fit the data *unequally* well, they assign *different* probabilities to what will happen in new experiments, and the difference in parsimony is relevant to estimating which will be more predictively accurate.

The simple point with which to begin is that the investigators pooled the behaviors of the different chimps that participated in each experiment, yielding a pair of frequencies for each experiment, as shown in Table 13.5. We must use these frequencies to estimate the values of the parameters used in two models. Here is a second-order model that has two adjustable parameters:

(Second) $\Pr(S \text{ takes } x \mid S \text{ believes that } D \text{ did not see } x) = p$
$\Pr(S \text{ takes } x \mid S \text{ believes that } D \text{ saw } x) = q$

The frequencies f_{1a}, f_{3a}, and f_{4a} help one estimate the first probability while f_{1b}, f_{3b}, f_{4b} as well as f_{2a} and f_{2b} bear on the second.

How should a first-order model be formulated? Here's an example to consider:

(First) $Pr(S \text{ takes } x \mid S \text{ believes that an opaque barrier is between } D \text{ and } x) = p_1$
$Pr(S \text{ takes } x \mid S \text{ believes that no opaque barrier is between } D \text{ and } x) = q_1$
$Pr(S \text{ takes } x \mid S \text{ believes that a transparent barrier is between } D \text{ and } x) = p_2$

[5] Lurz (2009) proposes such a unified first-order account and describes as an alternative experimental protocol for distinguishing second-order accounts from their complementary first-order rivals.

$Pr(S \text{ takes } x \mid S \text{ believes that no transparent barrier is between } D \text{ and } x) = q_2$

$Pr(S \text{ takes } x \mid S \text{ believes that } D \text{ was not present during food placement}) = p_3$

$Pr(S \text{ takes } x \mid S \text{ believes that } D \text{ was present during food placement}) = q_3$

Data from the first experiment is relevant to estimating the first pair of parameters, data from the second experiment is relevant to estimating the second pair, and data from the third and fourth experiments is relevant to estimating the third. Since First has more adjustable parameters, it will fit the data better than Second will. However, it may turn out that the simpler model receives the better AIC score.[6]

In applying the model selection framework to First and Second, you need to find maximum likelihood estimates of parameters that represent $Pr(A|B)$ by attending to frequency data concerning $f(A|P)$, where A is an action, B is a belief state, and P is a physical property of the experiment (e.g., where food items are located). How can these data be used to estimate these probabilities? The axioms of probability entail that

$$Pr(A|P) = Pr(A|B\&P)Pr(B|P) + Pr(A|-B\&P)Pr(-B|P).$$

If P, B, and A form a causal chain, with B screening off P from A, which means $Pr(A|P\&B) = Pr(A|B)$ and $Pr(A|P\&-B) = Pr(A|-B)$, this equality simplifies to

$$Pr(A|P) = Pr(A|B)Pr(B|P) + Pr(A|-B)Pr(-B|P).$$

If we adopt the assumption that $Pr(B|P) = 1$, which entails that $Pr(-B|P) = 0$, we obtain

$$Pr(A|P) = Pr(A|B),$$

which allows $Pr(A|B)$ to be estimated from $f(A|P)$. The assumption that $Pr(B|P) = 1$ means that the physical circumstances of the experiment, along with the chimp's other mental states,[7] *determine* what its belief state will be. Perhaps there is a way to secure identifiability without making this assumption, but I don't see what it would be. Notice that the assumption that $Pr(B|P) = 1$ reduces the number of *adjustable* parameters in both

[6] This kind of argument also applies to the comparison of zeroth- and first-order models of intentionality.

[7] It is assumed that subordinates want food and don't want to be punished by dominants.

models. The data are no longer asked to supply estimates for parameters that describe $Pr(B|P)$ but need only do so for parameters of the form $Pr(A|B)$.

I have no stake in claiming that First is the best first-order model nor that Second is the best model of second-order. First and Second are just the examples I have used to illustrate how model selection applies to the problem at hand. These models differ by four parameters. If there were more qualitatively different data sets from additional experiments, the difference in parsimony between the first- and second-order models might increase. In model selection, the difference in parsimony defines a threshold: it indicates how much better the more complex model must fit the data for it to have the better AIC score. The larger the difference in parsimony, the higher the bar is set.

How does this comparison of First and Second connect with the assumption, mentioned earlier, that an organism that has second-order beliefs also must have first-order beliefs? This assumption does not mean that a second-order model must have *parameters* that represent the impact of first-order beliefs. By the same token, even if an organism with a psychology must extract energy from its environment, it does not follow that a psychological model must contain parameters that represent those energetic processes. The Second model does not contain parameters that represent any first-order beliefs, though the model is perfectly consistent with the thought that second-order beliefs occur only when they are caused by first-order beliefs. The two models I have considered are shown in Figure 13.3.[8]

Does it make sense to insist that we should not compare First and Second but should instead compare First with a new model that has parameters that represent both first- and second-order beliefs? This new model, which I'll call First+Second, is represented in Figure 13.4.

To consider the competition between First and First+Second in a model selection framework, we need to figure out how each can be rendered identifiable. If we pursue a strategy similar to the one I described in connection with the competition between First and Second, the result will be that there is no real difference between First+Second and Second: if we assume $Pr(B_i|P) = 1$ and that $Pr(B|B_i) = 1$, the only adjustable parameters that remain in First+Second are of the form $Pr(A|B)$. Understood in this way, First+Second is in fact more parsimonious than First. That may seem strange, but the question remains of how the two models can be identified without our being driven to that conclusion. There is another problem with

[8] See the discussion of thirst in Whiten (1996, p. 284); see also Sober (1998a).

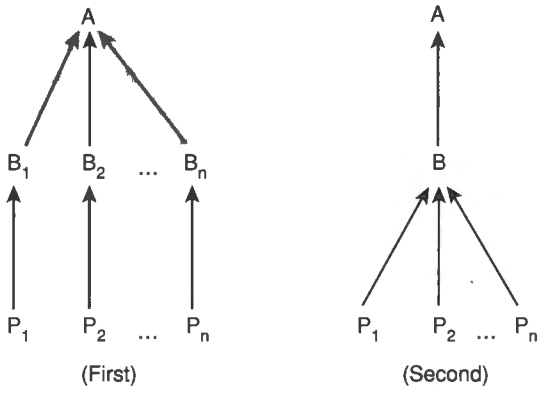

Figure 13.3 The second-order model says that the diverse physical circumstances P_1, $P_2, \ldots, P_n$ all cause the same second-order belief state B, which in turn causes action A. The first-order model says that different physical circumstances cause different first-order belief states $B_1, B_2, \ldots, B_n$, which each cause A.

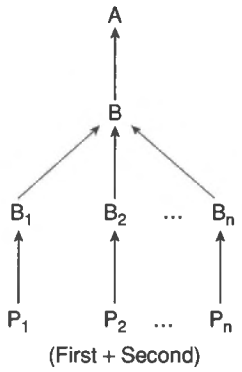

Figure 13.4 The Model First+Second says that different physical circumstances P_1, $P_2, \ldots, P_n$ cause different first-order beliefs $B_1, B_2, \ldots, B_n$, which each cause the second-order belief B, which in turn causes the action A.

rejecting the comparison of First and Second and insisting that the only relevant comparison is of First with First+Second, where First+Second is required to have adjustable parameters for all the beliefs it mentions. To do so is to refuse to consider the possibility that a model that introduces second-order intentionality might in fact be more parsimonious than one that restricts itself to first-order intentionality. Think how unmotivated it would be to take that stance in connection with the competition between a

zeroth- and a first-order model. Do we really want to say that introducing intentionality cannot simplify the resulting model? If it can, some of the parameters representing zeroth-order states must be dropped.[9]

9. ANTI-REALISM

Responding to philosophers who interpret him as an instrumentalist, Dennett (1991b, p. 29) says that he advocates "a mild and intermediate form of realism" about intentional psychology and that he thinks it isn't useful to try to locate his position in the dichotomous choice between realism and instrumentalism (p. 51). Even so, there are instrumentalist themes in his essay that connect it with ideas from model selection theory. Dennett (p. 36) emphasizes the role of simplified idealizations in science and sees this as the right context in which to understand folk and scientific psychology. Even though idealized models are false, they can be useful in making predictions. Another point of contact occurs in Dennett's discussion of "real patterns." What is the difference between a finite data set that has a pattern and one that is random? Dennett (p. 32) employs an idea from computer science according to which an n-member sequence is random if it cannot be compressed – if it cannot be generated by a rule that requires fewer than n symbols to state. If the sequence is random, the most succinct way to generate it is by brute enumeration. Minimum description length has developed into a criterion for model selection; it has been given an instrumentalist interpretation and it has been related to model selection criteria like AIC (Grünwald [2007]).

In spite of these links, there are some differences between the way I've understood instrumentalism in connection with model selection and Dennett's (1991b) discussion of realism and instrumentalism in intentional psychology. Dennett frames the philosophical question as follows. It is agreed that folk psychology is predictively successful. The question is how one should explain why this is so. Is the best explanation that beliefs and desires exist, or that they are useful fictions, or is there a coherent third alternative to consider? Whatever the best answer is to this question, it is not one that model selection theory addresses. This is because the Akaike's framework allows one to be an instrumentalist about *models*; it doesn't support instrumentalism about other propositions. A "model" (in the sense of that term used in statistics) is a special kind of beast. It contains

[9] This bears on the criticisms that Povinelli and Vonk (2006) make of Hare *et al.*'s experiments. See Fitzpatrick (chapter 14) for discussion.

adjustable parameters that can be fitted to data. Not every statement is a model in this sense. The statement that physical objects exist is not a model, nor is the statement that beliefs and desires exist. Not every consequence of a model is a model. Consider the models of lung cancer discussed earlier. One may want to be an instrumentalist about these models while at the same time being a realist about smoking, asbestos exposure, and lung cancer. Surely there is no reason to regard smoking, asbestos, and cancer as useful fictions. By parity of reasoning, instrumentalism about models that postulate beliefs and desires does not entail instrumentalism about beliefs and desires.

The thought that instrumentalism is appropriate for some propositions but not for others finds another application in the distinction between models and fitted models. I have described AIC as a device for estimating the predictive accuracy of a model M, but it is equally true that it estimates the closeness to the truth of $L(M)$, when closeness is measured by Kulback-Leibler distance. This is why I have defended the mixed philosophy of *instrumentalism for models, realism for fitted models* (Sober [2008]).

If AIC is a device for estimating a model's predictive accuracy, where does that leave the issue of explanation? Philosophers often object to Dennett's intentional stance (instrumentalistically understood) because they think it robs intentional psychology of the power to explain behavior. But surely an idealized model can be explanatory even though it is false. This doesn't just mean that it would be explanatory if it were true. That faint praise applies to models that are wildly wrong. Good idealizations can help us understand even when we know they are false. Those who think that explanations must be true may insist that what does the explaining is not the false idealization I but the true statement that I is a good idealization. Even so, instrumentalism about models does not mean that models cannot help explain.

The connection between the Akaike framework and Dennett's "mild and intermediate realism" is worth exploring further, but I will not attempt to do so here. The observation I would make in conclusion is that the framework offers something to both instrumentalists and realists. Instrumentalists can see the point of false models as predictive and explanatory devices and realists can see the point of comparing fitted models in order to judge which are closest to the truth. And both can see why parsimony matters. This last dividend may be the one most relevant to cognitive scientists who puzzle over what the principle of conservatism means and why it should be taken seriously.