In K. Binmore & S. Okasha (Eds.), *Evolution and rationality: Decisions, co-operation and strategic behaviour* (pp. 84-109). Cambridge, United Kingdom: Cambridge University Press. © 2012

CHAPTER 5

Are rational actor models "rational" outside small worlds?

Henry Brighton and Gerd Gigerenzer

CHAPTER 5

Are rational actor models "rational" outside small worlds?

Henry Brighton and Gerd Gigerenzer

5.1 INTRODUCTION

Given a formally well-defined task, a rational actor model defines a rationally justified, optimal response. Rational actor models are desirable goals in the behavioral, cognitive, and social sciences, but in this chapter we use the distinction between small and large worlds to question the cachet associated with the terms "rational" and "optimal." Ideally suited to the analysis of small world problems, both concepts can be counterproductive in the analysis of large world problems. In small worlds, the relevant problem characteristics are certain and uncontroversial in their formalization. For example, a tin can manufacturer seeking to minimize the tin used to package 12 ounces of soup might use solid geometry to determine an optimal can design. In this small world the manufacturer is safe in claiming that no other can design uses less tin. Large worlds are characterized by inherent uncertainty and ignorance, properties which undermine the validity and existence of optimal responses. An aircraft manufacturer designing a flight control system, for instance, faces a large world problem due to the complexity and uncertainty of the operating conditions. Rational actor models may rest on rigorous formal foundations, but they can also signify the questionable use of small world methods to understand large world problems.

With similar concerns, Savage introduced the distinction between small and grand worlds when assessing the limits of Bayesian decision theory (Savage 1954). Savage's decision theory for small worlds – those where an agent has access to a decision matrix defining states of the world, consequences, and actions – shows that the agent will maximize subjective expected utility, providing their preferences satisfy Savage's axioms. Savage saw the problem of casting large world problems, those

involving uncertainty and ignorance, in these terms as "utterly ridiculous" (p. 16). Consequently, we will use the term "Savage's problem" to refer to obstacles and potential dangers in using analytic methods geared for small worlds to theorize, and make statements about, large worlds. Specifically, we consider Savage's problem in the context of inductive inference, where a decision maker is required to generalize from observations and infer statistical properties of the environment. For example, an organism foraging for food may infer regularities in the distribution of food items from observations of previous food items. What distinguishes small from large worlds in inductive inference? Moreover, how significant is Savage's problem to the study of inductive inference, where rational actors and optimal responses play a particularly influential role?

We will examine these questions by first setting out the relationship between inductive inference, uncertainty, and error, using statistical learning theory. Within this setting we consider barriers to the identification of optimal responses. The ubiquity of these barriers in the study of real-world problems places a question mark over the range of applicability of rational actor models, and highlights the need for alternative approaches to studying how they function. We then use the study of simple heuristics to illustrate an approach based on algorithmic modeling and competitive model testing, reflecting a relatively recent movement in statistics and pattern recognition (Breiman 2001). Well-informed theorists know that good models hinge on insightful abstraction, and the assumptions made in order to formalize a model will be breached. These are basic facts about models in general. Using the distinction between small and large worlds, our goal is to understand the limits of a specific class of model, rational actor models, which often discharge uncertainty in order to arrive at an optimal solution.

5.2 UNCERTAINTY IN INDUCTIVE INFERENCE

All organisms face uncertainty. For instance, look closely at any aspect of cognition and it will likely involve the process of inductive inference, the problem of identifying systematic patterns in observations. Everyday examples include inferring the properties of a visual scene, inferring the intentions of a speaker from a single utterance, or deciding if it is quicker to take the bus, bicycle, or train. Because an infinite number of explanations will always be compatible with a finite series of observations, each of these problems involves uncertainty. Two key questions guide the study

of inductive inference in humans and other animals. First, how, in mechanistic terms, do organisms arrive at inductive inferences? Second, what is the relationship between the behavior we observe in organisms, and the optimal behavior, as defined by a rational actor model? We will begin by addressing the second question, and start by taking an idealized perspective on uncertainty, where we adopt the the perspective of an omniscient observer with full knowledge of the task at hand. Relative to this omniscient observer, our first task is to categorize the various forms of uncertainty agents face when making inductive inferences. This will require making the inductive inference problem more precise.

5.2.1 Learning from examples

Consider an agent interacting with a series of partners in some game-theoretic setting. Prior to interacting with a new partner, it would be useful if the agent could accurately categorize this partner as, for example, a likely cooperator or defector. This inference could be based on perceivable features of the partner in question, such as their age, personality, or social status. Supervised learning from labeled examples is the study of algorithms which learn predictive models by generalizing from past observations (Bishop 2006; Hastie et al. 2001). In our game-theoretic example, past observations refer to previously encountered partners, along with a label categorizing their observed behavior. A predictive model accurately categorizes new partners when only the feature values, such as those mentioned above, are known.

Supervised-learning problems are formalized by first defining an input space X of feature vectors used by the agent to encode observations, and an output space Y specifying the structure of the labels being predicted. Observations are drawn from the product space $Z = X \times Y$. Categorization tasks are those where Y is a set of category labels. Regression tasks are those where Y is some range of numeric values. An environment which is unknown to the agent determines the underlying functional relationship between inputs and outputs, and is given by a joint probability distribution $\mu(\mathbf{x},y)$. Given a multiset of r labeled observations $S = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{r} \in Z^r$, the agent uses a learning algorithm to select a hypothesis which represents an informed guess at which systematic pattern best explains the functional relationship between inputs and outputs, not just between the observed examples, but in general.

Hypotheses describe what is systematic in the observations, and allow the agent to make predictions about novel objects, the problem of guessing

y given only $\mathbf{x}$. The hypothesis space $\mathcal{H}$ defines the set of hypotheses the agent can select from, and plays a critical role in its ability to generalize accurately from experience. The learning algorithm is also critical because it determines which hypothesis is selected for a given series of observations. A learning algorithm, L, implements a mapping,

$$L : \bigcup_{r \geq 1} Z^r \mapsto \mathcal{H}, \tag{5.1}$$

from multisets of observations to hypotheses. It will prove useful to view the hypothesis space as a model $f_\theta(\mathbf{x})$, indexed by the parameters θ. Selecting a hypothesis is the process of estimating the parameters θ from observations.

5.2.2 Error-inducing risk and uncertainty

An omniscient observer has full knowledge of the environment, which is assumed to be fully specified by the data-generating distribution $\mu(\mathbf{x},y)$. A hypothesis induced by an agent can also be seen as a joint probability distribution over the observations,[1] and denoted $\sigma(\mathbf{z})$. The discrepancy between these two distributions, μ and σ, will be defined as their Kullback-Leibler divergence, given by

$$D(\mu \,||\, \sigma) = \sum_{z \in Z} \mu(\mathbf{z}) \log_2 (\mu(\mathbf{z}) / \sigma(\mathbf{z})). \tag{5.2}$$

The Kullback-Leibler divergence is perhaps best understood from a coding perspective, where a probability distribution over the observation space implies an optimal coding scheme which assigns a code of length $log_2(1/p)$ bits to an observation occurring with probability p. The greater the probability of the observation, the shorter its code. The Kullback-Leibler divergence measures the number of additional bits required to code events governed by σ when using the code for μ (Cover and Thomas 1991). When the distributions are identical, $D(\mu||\sigma) = 0$. We will use this measure of divergence to arrive at an abstract categorization of three basic forms of discrepancy, referred to as stochasticity, underspecification, and misspecification. These discrepancies can exist independently, and their potential combinations are depicted in Figure 5.1. Both the organism and the theorist attempting to understand the organism face these uncertainties.

[1] Here we assume that the agent learns a generative model from the observations, one which models both $\Pr(y|\mathbf{x})$ and $\Pr(\mathbf{x})$ (see Bishop 2006, for further discussion).

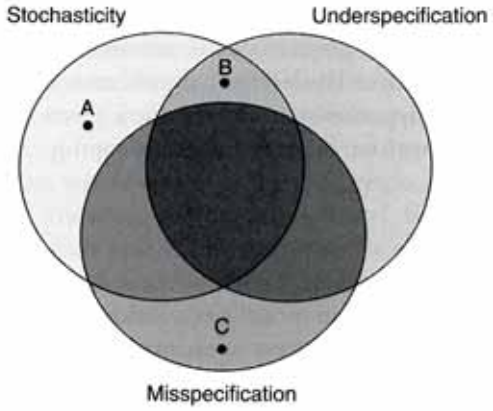

Figure 5.1. Possible combinations of the three basic forms of discrepancy: stochasticity, underspecification, and misspecification. The discrepancies are independent of each other, and all combinations are possible. Point A corresponds to the problem of predicting the outcomes of a fair die. Point B corresponds to the problem of predicting the relative frequency of red and black balls in an urn. Point C corresponds to the problem of predicting the toss of a loaded die which always falls on 4, using a misspecified hypothesis space which allows predicting only 1 or 6.

Stochasticity

Consider an agent with certain knowledge of μ, such that $\mathcal{H} = \{\sigma\}$ and $D(\mu\|\sigma)=0$. This agent will make error-free predictions if outputs are always a deterministic function of the inputs. If the predictive distribution $\Pr(y|\mathbf{x})$ is stochastic, such that for at least one input $\mathbf{x}$, the output is nonunique, then error will result. Stochasticity is the most basic form of discrepancy between agent and environment. It arises due to external randomness and cannot be eliminated through further observation, or by designing the organism differently. For example, even though an agent knows that a die is fair, it will make errors when predicting rolls of the die (Figure 5.1, point A). In short, even when granted full causal knowledge of the process governing observations – such as knowledge of physical probabilities (Giere 1999), propensities (Popper 1959), or a priori probabilities (Knight 1921) – the agent will still make errors when predicting events under conditions of stochasticity.

Underspecification

Discrepancies due to underspecification exist when the number of observations underdetermines the choice of hypothesis, and is too small to reliably converge on the best model σ^* in some $|\mathcal{H}| > 1$, where

$$\sigma^* = \min_{\sigma \in \mathcal{H}} D(\mu \| \sigma). \tag{5.3}$$

At one extreme, in the complete absence of observations, the agent faces what Knight (1921) refers to as uncertainty. For Knight, an agent may know that an urn contains red and black balls (knowledge coded implicitly in the hypothesis space) but the probability of choosing a black ball from the urn is uncertain if no other information is available (Figure 5.1, point B). As the number of observations increases from zero, the agent moves from an uncertainty situation to a risk situation, and can begin measuring what Knight terms statistical probabilities. Underspecification, in Knightian terms, will therefore include situations of both risk and uncertainty.

Misspecification

Discrepancies due to misspecification result from the agent's inability to model the data-generating distribution exactly, such that $D(\mu \| \sigma^*) > 0$. Misspecification can occur in the absence of stochasticity and underspecification. To take a simple example, an agent entertaining two hypotheses, $\mathcal{H}$ = {"die always falls on 1", "die always falls on 6"}, will make errors when predicting a loaded die which always falls on 4 (Figure 5.1, point C). Such errors are unrelated to stochasticity and underspecification.

Nonstationarity

So far we have implicitly assumed that the probability distribution on the observation space – which determines what we observe – is the same as the distribution which determines the accuracy of our inferences. Situations involving nonstationarity are those where these two distributions differ. We will assume that nonstationary problems represent a special case of misspecification, since any temporal dependency on the distribution generating observations, $\mu_1(\mathbf{x},y)$, and the distribution which determines the accuracy of predictions, $\mu_2(\mathbf{x},y)$, can be reformalized by considering a single distribution, $\mu_t(\mathbf{x},y)$, parameterized by time t. The class of nonstationary problems nevertheless includes forms of discrepancy worthy of study in their own right, and poses problems of great practical significance (e.g. Quiñonero-Candela et al. 2009; Hand 2006).

The generating distribution $\mu(\mathbf{x},y)$ when rewritten $\Pr(y|\mathbf{x})\Pr(\mathbf{x})$ highlights two broad categories of nonstationarity. First, covariate shift occurs when $\Pr(\mathbf{x})$ changes (Shimodaira 2000). Second, what we term predictive shift occurs when $\Pr(y|\mathbf{x})$ changes. Figure 5.2 illustrates how the two forms of

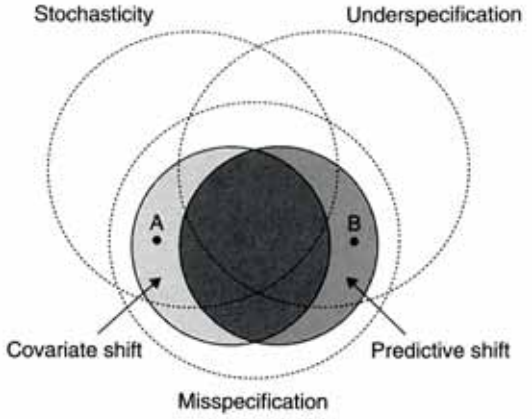

Figure 5.2. Nonstationarity as a form of misspecification. Covariate shift refers to a change in the probability distribution over inputs. Predictive shift refers to a change in conditional probability of the outputs given the inputs. These two forms of nonstationarity can occur independently. Point A, discussed in the main text, refers to an example of sample selection bias, where partners in a game-theoretic setting may be of similar age. Point B is an example of predictive shift, also discussed in the main text, where potential partners may change their strategy.

nonstationarity can exist independently and always represent a form of misspecification. An example of covariate shift is sample selection bias, which occurs when a potentially unknown process probabilistically rejects observations (Heckman 1979). For example, in our game-theoretical setting, the partners an agent experiences may be restricted to be those of a similar age (Figure 5.2, point A). Here, the predictive distribution remains constant but our observations of it will be skewed. Predictive shift could occur if potential partners became aware of an agent's decision process, and adjusted their behavior to exploit it. Here, the functional relationship between observable features and outcomes may change (Figure 5.2, point B).

5.2.3 Further uncertainties

These basic uncertainty categories refer to discrepancies between an environment, experienced through a series of observations, and abstract properties of the hypothesis space. A full categorization of real-world uncertainties is infeasible and unbounded, and some basis for separating exogenous from endogenous uncertainty is always necessary. For example, all biological systems operating at temperatures above 0 kelvin

must be robust to the uncertainties arising from thermal noise, but we will abstract from this problem even though it imposes significant constraints on functional design (Wagner 2005). Perhaps the most significant source of error-inducing discrepancy to be sidestepped in the coming discussion will be the costs associated with errors. In particular, we assume that the prices paid for incorrect predictions are given by standard loss functions, such as zero–one loss for classification problems, and squared difference for regression problems. Organisms in natural environments must contend with events incurring potentially very different costs, such as the devastating consequences of the highly improbable (Taleb 2009; Bookstaber and Langsam 1985).

In addition to the basic uncertainty categories defined above, one must also consider uncertainty surrounding the algorithm L used to select hypotheses. What we will term computational uncertainty arises from both fundamental and idiosyncratic barriers which constrain the processing of observations. Such constraints introduce error-inducing discrepancies, just like the basic categories discussed above. Misspecification, for example, arises because finite observations will tend to undermine the selection of the most predictive hypothesis. In the same way, resource restrictions on L can limit the ability of the organism to arrive at a predictive model. Constraints on L range from primitive considerations of computability and intractability, to additional, organism-specific biological and computational resource limitations. The former apply to all computational agents (Hopcroft and Ullman 1979; Papadimitriou 1995). The precise nature of the latter will depend on the agent, and include constraints such as limitations on working memory. Unlike many sources of uncertainty, computational uncertainty poses a fundamental problem for all agents in all contexts, with the only exception being omnipotent, and therefore fictitious, agents.

5.3 FROM ERROR TO OPTIMALITY

An omniscient observer knows the true state of nature. A rational actor achieves the best possible response relative to a set of assumptions about the true state of nature. Consequently, our degree of faith in a rational actor model should be in proportion to our knowledge of the problem. When facing large worlds, we should question both the validity of our models, and our use of the terms "rational" and "optimal." These issues focus the remainder of discussion onto the following question: Under conditions of stochasticity, underspecification, and misspecification,

what does it mean for an organism to be rational, and respond optimally? Answering this question will first require taking a closer look at the relationship between the organism, the environment, and error.

5.3.1 Analyzing and categorizing error

Relative to the generating distribution μ, the hypothesis $f(\mathbf{x})$ chosen by the agent after observing a sample of observations S is likely to incur some degree of error. To simplify matters, we will focus on the problem of regression where error is commonly defined as the squared difference between the true and predicted value. The lowest achievable error is incurred by the function $g(\mathbf{x})$, defined as the conditional expectation $\mathrm{E}[y|\mathbf{x}] = \int yp(y|\mathbf{x})dy$ of the predictive distribution $\Pr(y|\mathbf{x})$. Relative to the generating distribution, the total error incurred by $f(\mathbf{x})$ is referred to as the expected loss,

$$\text{expected loss} = \int\{f(\mathbf{x}) - g(\mathbf{x})\}^2 p(\mathbf{x})\mathrm{d}\mathbf{x} + \int\{g(\mathbf{x}) - y\}^2 \mu(y,\mathbf{x})\mathrm{d}\mathbf{x}\,\mathrm{d}y. \quad (5.4)$$

Notice that the first term of this expression is dependent on the chosen hypothesis, $f(\mathbf{x})$, but the second term is not, and therefore it cannot be reduced, whatever our choice of $f(\mathbf{x})$. This second term expresses the error incurred by the agent when always selecting $f(\mathbf{x}) = g(\mathbf{x})$. This part of the error corresponds directly to error arising from stochasticity, which is commonly referred to as irreducible error or noise.

Controllable components of error

The first term of Equation 5.4 measures that part of the error we can reduce through the appropriate design of the agent, since it depends on the policy for selecting $f(\mathbf{x})$ from observations. Adopting a frequentist perspective, and imagining that the tape of experience were replayed several times, we will use a standard technique in statistical learning theory for decomposing the controllable error into two components, referred to as bias and variance (O'Sullivan 1986; Geman et al. 1992; Bishop 2006; Hastie et al. 2001). For each replay of the tape of experience, a potentially different sample $S = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{r}$ will be observed, where S is a multiset of r observations sampled from $\mu(y,\mathbf{x})$. Now, replaying the tape of experience k times yields k of these multisets, given by the ensemble $\mathcal{S} = \{S^{(1)}, S^{(2)}, \ldots, S^{(k)}\}$.

Rather than the error incurred by the agent after observing a specific sample, we would like to know about its mean error, relative to the ensemble

of hypotheses selected by the agent's learning algorithm for each member of $\mathcal{S}$. For a given $\mathbf{x}$, the expectation of the first term of Equation 5.4 for an ensemble of samples of size k is $\mathbb{E}_S\left[\left\{f(\mathbf{x})-g(\mathbf{x})\right\}^2\right]$. Integrating over $\mathbf{x}$, this expectation is the sum of two terms, the first of which is bias,

$$(\text{bias})^2 = \int\left\{\mathbb{E}_S\left[f(\mathbf{x})\right]-g(\mathbf{x})\right\}^2 p(\mathbf{x})d\mathbf{x}, \tag{5.5}$$

which is the squared difference between mean prediction made by the functions induced from the ensemble, and the true function $g(\mathbf{x})$, over all inputs. The second term is the variance, given by

$$\text{variance} = \int \mathbb{E}_S\left[\left\{f(\mathbf{x})-\mathbb{E}_S\left[f(\mathbf{x})\right]\right\}^2\right]p(\mathbf{x})\,\mathrm{d}\mathbf{x}, \tag{5.6}$$

which is the squared difference between the mean prediction of the ensemble, and the predictions of the individual functions induced for each member of the ensemble. Collecting the terms for bias, variance, and irreducible error, the expected loss given by Equation 5.4 can be summarized as

$$\text{Total error} = (\text{bias})^2 + \text{variance} + \text{irreducible error}. \tag{5.7}$$

By decomposing the error in this way, we have a more precise understanding of the effects of misspecification, underspecification, and stochasticity. The bias and variance incurred by the algorithm will always be a function of the sample size, r, and the properties of the generating distribution. In general, variance decreases as a function of r, the size of the observed sample. Variance occurs when changes to the sample lead to changes in the error of the induced hypothesis. Limiting variance requires reducing the sensitivity of the learning algorithm to the effects of resampling. Generally speaking, bias occurs due to an inability of the algorithm's hypothesis space to model the generating distribution.

The bias/variance dilemma

Keeping in mind that the generating distribution is unknown to the agent, considerations of bias and variance highlight a fundamental problem in inductive inference known as the bias/variance dilemma. At one extreme, the agent's learning algorithm could express a wild guess by ignoring the observations altogether, by always selecting the same hypothesis. This approach guarantees zero variance, but can lead to high

bias unless the guess turns out to be correct, or close to correct. At the other extreme, the agent's learning algorithm could hedge its bets, let the observations speak for themselves, and select a hypothesis from a highly flexible model space capable of approximating any function. This policy could in principle guarantee zero bias, but usually at the expense of high variance since the flexibility of the hypothesis space is likely to lead to an oversensitivity to the vagaries of particular samples. The bias/variance dilemma arises because methods for minimizing variance tend to increase bias, and methods for minimizing bias tend to increase variance. The two need to be balanced, a process which should be guided by knowledge of the task at hand.

5.3.2 The relationship between bias, variance, and optimality

What is the optimal response to the bias/variance dilemma? Zero variance and zero bias will be achieved by a learning algorithm which always induces the same hypothesis, the conditional expectation of the predictive distribution $g(\mathbf{x})$. Recall, though, that we are interested in cases where the generating distribution is not known with certainty. Taking a Bayesian perspective, the assumptions made about the generating distribution, which are required to generalize from data in any way, are coded in the structure of the hypothesis space and the prior. Optimality is always defined relative to these assumptions. If we knew the generating distribution, then the hypothesis space should contain a single hypothesis, $g(\mathbf{x})$. If we knew the functional form of the generating distribution was, for example, linear, then the inference problem reduces to estimating the parameters of the linear model. Optimality is not about the presence or absence of error, but the degree of error relative to a set of assumptions. The optimal algorithm incurs bias and variance, like any other algorithm. The label "optimal" simply marks out a particular algorithm as incurring the least error among all algorithms sharing the same hypothesis space and prior.

5.4 FROM SMALL WORLDS TO LARGE WORLDS

What is a small world problem? The most restrictive definition of a small world problem is one where the optimal response is certain, and uncontroversial. For problems of inductive inference, optimal responses are usually framed in terms of Bayesian statistics. These optimality results rely on two conditions being met. First, the structure and properties of

the problem are known a priori, such that the appropriate hypothesis space and prior are known with certainty, rather than inferred from observations. Second, the structure of the hypothesis space and prior permit calculation of the posterior probability using exact methods. The first condition settles any dispute over the external validity of the model – the degree to which the properties of the model match those of the system being modeled. The second condition settles any dispute over the internal validity of the model – the degree to which the inferences made using the model are rationally justified, assuming that the model is correct, and has external validity. Obviously, very few problems of interest to the cognitive, behavioral, and social sciences can be said to satisfy the first condition. Similarly, many problems, when formalized in these terms, require use of inexact, approximate Bayesian methods. However, this definition of a small world guarantees that statements of rationality and optimality provide exactly what they advertise. This guarantee is possible because all relevant aspects of the problem, and solution, are certain. In practice, though, rational actor models are not absolute statements but statements about optimality relative to some set of assumptions.

The notion of a large world becomes relevant as soon as the structure and properties of the problem are inferred from observations, and issues of misspecification, underspecification, nonstationarity, and all the other categories of uncertainty we discussed begin to impact on the external validity of the model. In addition, issues of analytic tractability impacting on the process of determining the posterior distribution can also introduce uncertainty. Thus, even in cases where we can guarantee an appropriate formalization of the problem, the analytic and computational demands of performing rational calculation also need to be considered, as they can lead to additional error. Now, all good theorists know that "a map *is not* the territory" (Korzybski 1958, p. 750) and all models are "wrong." Given this, why question the objective of rational actor models? After all, if your underlying assumptions are openly stated, then everyone is free to assess the implications of your model as they see fit. Our point is that increasing uncertainty and ignorance should at some point lead us to question the notions of rationality and optimality. When factors known to compromise the internal and external validity of the model are at play, declarations of rationality and optimality become less and less meaningful. Next, we show that the use of rational actor models and optimality results in the analysis of large world problems is a choice, not a necessity (Brighton and Olsson 2009).

5.4.1 Distinguishing relative from absolute function

In developing a rational actor model one is attempting to answer the question of how to make optimal inferences, predictions, given assumptions about the properties of the task at hand. When these properties are uncertain, how else can functioning be assessed? We will contrast the pursuit of rational actor models with an alternative. Borrowing the terminology of Breiman (2001), rational actor models rely on data modeling, an approach used routinely in statistics where one assumes an underlying statistical model, and then uses observations to estimate the model parameters. We will adopt an approach referred to by Breiman as algorithmic modeling, where the underlying statistical model is treated as unknown (or nonexistent), and the estimated predictive accuracies of alternative statistical models, or learning algorithms, are then compared in an attempt to learn about the problem. These two approaches are both mute on the problem of assessing the rationality of organisms. To illustrate their use in examining the ability of organisms to function in uncertain environments we will use two fictional modelers, named Jones and Smith.

Modeler Jones follows a methodology known as the rational analysis of cognition, an influential approach inspired by optimality modeling in biology (Anderson 1991; Oaksford and Chater 1998, 2007). A rational analysis typically proceeds by inferring from observations of the task environment an appropriate observation space, hypothesis space, and prior over the hypothesis space. These components define a probabilistic model of the task environment. Jones then derives an optimal response function which specifies the inferences that a rational, usually Bayesian observer would make in response to further observations of the environment. The rationality of the agent in question is then assessed by comparing its behavioral responses with those of the rational model. If the rational model provides a close fit to the agent's responses, Jones argues that the agent makes optimal inferences.

Modeler Smith is interested in exactly the same problem, begins by observing the task environment, and also proposes an appropriate observation space. Crucially though, Smith *refrains* from making an inference about an optimal response, an inference which requires postulating an hypothesis space and generating distribution. Instead, Smith estimates the predictive accuracy of potentially several statistical models *relative to observations* of the environment. Predictive accuracy is often measured using cross-validation, a technique we illustrate below (see Rissanen

1986, for an alternative relativistic approach which stresses compression rather than prediction). This comparison informs the question of which mechanistic design features prove functional, rather than optimal, in the task environment. The ability of each mechanism to describe the agent's behavior is an additional question, guided by functional analysis, but assessed by experimentation (Brighton and Gigerenzer 2011). In contrast to the approach of Jones, Smith refrains from making an inference about the generating distribution, conducts a competitive test of rival process-level theories instead of seeking a single rational actor model, and therefore sees success as something always relative to alternative models, rather than being defined by the benchmark of a single rational actor model.

5.4.2 Relative functioning in large worlds

What insights can Smith's relativist approach offer? As an illustration, consider the problem of how a retail marketing executive might distinguish between active and inactive customers based on their purchasing history. Current thinking on how to best address this problem centers on stochastic customer-base-analysis models, such as the fairly sophisticated Pareto/negative binomial distribution model. Following modeler Jones' approach, this model defines rationally justified inferences about customers, given an assumed underlying probability distribution. In the spirit of modeler Smith's approach, Wübben and Wangenheim (2008) compared the performance of several models, including the Pareto/negative binomial distribution model, and found that the best-performing model was a simple hiatus rule which predicts inactivity if a customer has not made a purchase in the preceding nine months (and for one problem, six months). To take another example, consider the problem of searching literature databases, where the task is to order a large number of articles so that the most relevant ones appear at the top of the list. When examining this problem, Lee et al. (2002) constructed a "rational" Bayesian model and found that in comparison to several competing models, a simple one-reason heuristic proved superior.

What can these, and similar findings, tell us? Models implementing a "rational" response to an underlying distribution which is assumed, but inaccurate, can result in performance inferior to that of a model which makes no explicit attempt to model an underlying distribution, or conduct any form of rational calculation. Many people find results like these surprising, but modeler Jones may argue that they say nothing about rational

actor models in general, they merely supply an argument against poorly designed rational actor models. Jones is correct, but misses the point. In a large world, one cannot identify a rational model, or an optimal response, with any confidence. Rather, the "rational" response is to search for better and better models. In this sense, the pursuit of rational actor models for large worlds can be counterproductive. First, one cannot assume that rational calculation brings us closer and closer to effective responses to uncertainty. Second, the alternative is not silence, but an examination of alternative models, particularly those which have been successful in the past. Perhaps the worst case scenario is that Jones' approach focuses functional analyses onto small world problems which bear little relation to the large world problems of interest. Next, we focus on the first issue, which concerns common intuitions about rational calculation.

5.5 HOW TO CONFRONT LARGE WORLDS

Mathematical objects such as generating distributions and rational actors could be viewed as misnomers when set against the uncertainty surrounding natural environments and the constraints impacting on computational actors. From a rational actor perspective, these complicating factors limit the opportunities for rigorous functional analysis. But rather than defining limits on rational analysis, these complexities highlight a need for alternative approaches (e.g. McNamara and Houston 2009). The relativist approach taken by our fictional modeler Smith is standard practice when comparing the predictive accuracy of learning algorithms and cognitive models on real-world datasets (e.g. Perlich et al. 2003; Brighton 2006; Chater et al. 2003; Czerlinski et al. 1999; Gigerenzer and Goldstein 1996). Datasets are simply collections of observations relating covariates to a dependent variable. For example, the daily stock prices over the last year, test results for a population of patients undergoing treatment, or a series of utterances and their associated meanings. Datasets are usually samples from an unknown and potentially unknowable distribution, often suffer from uncertainty arising from noisy or missing information, and may represent a snapshot of a nonstationary problem. In much the same way, an organism functioning in an environment, and a scientist attempting to explain this organism's functioning, both face problems shot through with uncertainty. The question of which strategy is optimal under these circumstances cannot be answered. But we can study why some strategies repeatedly prove more effective than others.

5.5.1 Making inferences with simple heuristics

The study of simple heuristics illustrates several crucial issues. Simple heuristics are cognitive process models which ignore information (Gigerenzer et al. 1999; Gigerenzer and Brighton 2009; Gigerenzer and Gaissmaier 2011). A central question in the study of heuristics is understanding when and why they outperform alternative cognitive models. For many, the idea that ignoring information can prove functional is counterintuitive. Consider, for instance, the simple heuristic, take-the-best (Gigerenzer and Goldstein 1996). Take-the-best is a cognitive process model describing how people decide which of two objects has a greater value on some criterion of interest, such as the price of two houses, or the maximum speed of two vehicles. If the criterion values are unknown, environmental cues such as the presence of a swimming pool or whether the vehicle has a jet engine or not, can be used to make an inference. More precisely, take-the-best has three steps:

(1) Search rule: Search through cues in order of their validity.
(2) Stopping rule: Stop on finding the first cue that discriminates between the objects (i.e. cue values are 1 and 0).
(3) Decision rule: Infer that the object with the positive cue value (1) has the higher criterion value.

In short, take-the-best searches for the first cue which discriminates between the objects, and then makes an inference using this cue alone. For example, a house with a swimming pool is likely to be more expensive than a house without. All other cues are ignored. Specifically, take-the-best searches for cues in order of their validity, which is a measure of how accurately each cue has made inferences on previous comparisons. This measure is simple in comparison with the measures used by most models, as it ignores any dependencies between cues. That is, take-the-best orders cues by assuming that the predictive ability of a cue can be determined independently of the value of the other cues.

Take-the-best is a linear model.[2] Furthermore, it deviates from common statistical intuition, and the methods employed by commonly assumed models of cognitive processing. For example, the Gauss-Markov theorem proves that among all unbiased linear models, the least-squares estimate incurs the lowest variance (e.g. Fox 1997). Assuming that the

[2] This is true at an outcome level, but take-the-best differs at a process level from the standard linear models, such as linear regression.

problem in question is linear, this tells us that a logistic regression model, for example, should incur lower error than take-the-best. How can we put this statement to the test? A classic task in the study of simple heuristics is the problem of inferring which of two German cities has a greater population. The dataset in question describes the eighty-three largest German cities using nine binary cues which detail properties of the cities, such as whether or not a city has a university or an intercity train station, or is located in Germany's industrial belt. After observing some sample of these cities, and estimating the model parameters from this sample, the accuracy of competing algorithms, like take-the-best and logistic regression, can be measured. Accuracy here refers to whether or not larger cities are correctly inferred as being larger. Crucially, the accuracy of these inferences is measured by considering only novel pairs of cities, comparisons between cities which were *not* used to estimate the model parameters. This process is repeated for different samples, and is termed cross-validation (Stone 1974).

Figure 5.3 plots the mean predictive accuracy of take-the-best as a function of the size of sample used to estimate the cue validities. In the same plot, we report the predictive accuracy of logistic regression. Means are reported, which are taken with respect to 5,000 random partitions of the dataset into training and testing sets. Now, why, in contrast to our interpretation of the Gauss-Markov theorem, does take-the-best outperform logistic regression so significantly? The Gauss-Markov theorem, despite often being interpreted as statistical justification for the widespread use of the least-squares linear model, is a small world theorem. The theorem holds when fitting, rather than predicting, data. As soon as one makes predictions about unseen data, the effects of underspecification and misspecification can change matters significantly.

A common response to this result, and similar findings for many other datasets (e.g. Brighton 2006; Czerlinski et al. 1999), is that logistic regression is a weak competitor, and more sophisticated methods designed specifically for avoiding the problem of overfitting will outperform take-the-best (e.g. Chater et al. 2003). In response, Figure 5.3 also compares the performance of take-the-best with a support vector machine, one of the most advanced methods in statistical pattern recognition (Vapnik 1995). Again, take-the-best outperforms the support vector machine over a significant portion of the learning curve. Notice that we have compared three models without postulating, or making assumptions about, a generating distribution determining the relationship between properties of German cities and their populations. Instead, we have respected the fact

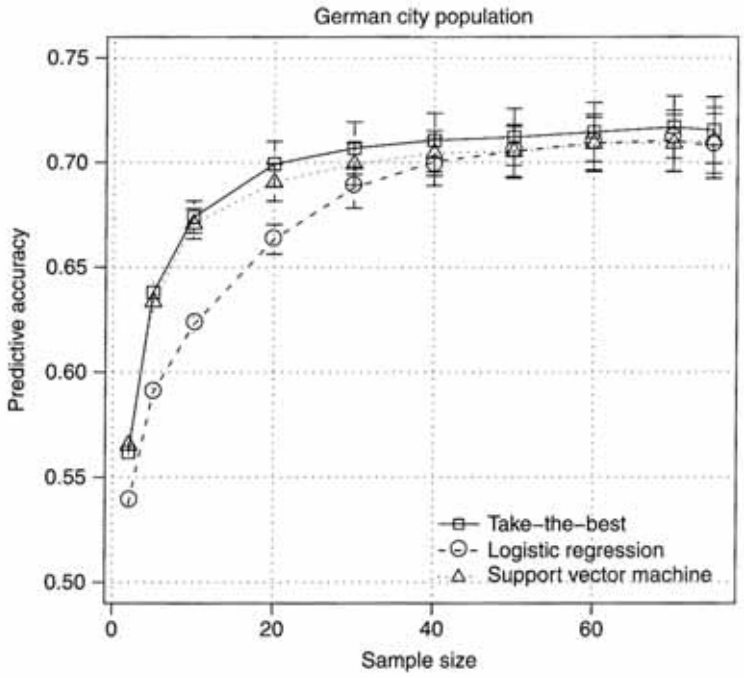

Figure 5.3. Can ignoring information be beneficial? For the German city population task, the predictive accuracy of the simple heuristic take-the-best is plotted as a function of the size of the sample used to estimate the model parameters. Take-the-best, despite ignoring conditional dependencies between cues and using only one cue to make an inference, outperforms both logistic regression and a support vector machine. Mean predictive accuracies are reported for 5,000 random partitions of the dataset. Error bars show the variance.

that "a set of data is one thing and a mathematical object, such as a distribution, is quite another, not only different in degree but different in kind" (Rissanen 1986, p. 395), and measured the relative ability of the models to make accurate inferences relative to the observations. How, though, are results like these related to the question of decision making in large worlds?

5.5.2 *The bias/variance dilemma in large worlds*

Recall from our discussion of the bias/variance dilemma that the problem of making accurate predictions can be decomposed into two subproblems: the problem of reducing variance, and the problem of reducing

bias. Variance arises due to underspecification, a lack of observations, while bias tends to arise as result of misspecification, an inherent inability on the part of the model to capture predictive patterns in the data. In their influential presentation of the bias/variance dilemma, Geman et al. (1992) discussed how restricting the hypothesis space can reduce variance. If these restrictions do not introduce additional bias with respect to the problem at hand, prediction error is likely to decrease. Another approach, taken by several statistical models, is to introduce bias which *is* relevant to the problem at hand, under the assumption that it will be outweighed by a greater reduction in variance. Recently, we showed that take-the-best succeeds by exploiting this trick; take-the-best's performance advantage arises exclusively through a reduction in variance by ignoring conditional dependencies between cues (Brighton and Gigerenzer 2007; Gigerenzer and Brighton 2009). The success of ridge regression and the naïve Bayes classifier can also be explained in these terms (e.g. Hoerl and Kennard 2000; Domingos and Pazzani 1997; Friedman 1997; Hastie et al. 2001).

From a rational actor perspective, where one conducts rational calculation with respect to an assumed hypothesis space and prior probability distribution, how can the success of alternative algorithms, like the naïve methods mentioned above, be explained? After all, why should deviations from a full Bayesian calculation prove functional? Advocates of rational actor models will likely argue that the success of any model can be explained using rational principles. For example, one can ask under which conditions take-the-best is optimal, and thereby frame take-the-best as a rational actor model under these conditions. Thus, it is trivially true that the success of heuristics, or any other model, in no way challenges the small world study of rationality. Two points are crucial here. First, while seeking a rational explanation for apparently nonrational processes can certainly yield great insight, this task is far from being a trivial exercise, and will often achieve only a limited, approximate explanation (e.g. Sanborn et al. 2010). The optimality conditions of the naïve Bayes classifier, for example, are only partially known, despite sustained study and a clear Bayesian interpretation (e.g. Domingos and Pazzani 1997; Kuncheva, 2006). Nevertheless, the naïve Bayes classifier appears in the top-ten list of data mining algorithms, and is used routinely in uncertain contexts, those where the modeler is largely ignorant of the generating distribution (Wu et al. 2007). Second, this issue has little or no impact on the question of what strategies prove functional in a large world. If the underlying structure of the problem is unknown, then optimality results offer little help. In large worlds, all we can do is estimate the predictive

accuracy of a range of methods, such as those algorithms which have performed well in the past. The outcome of such an exploration will never be a proof of optimality, nor a state of understanding which tells us what is "rational." In large worlds, we have no choice but to abandon the objective of an optimal response, and instead search for improved understanding. Possessing a rational interpretation of the algorithms used to conduct this search changes nothing.

5.5.3 Understanding large worlds

We defined small worlds as presenting problems where all relevant factors are certain and their formalization yields to feasible rational calculation. Uncertainty enters the picture when the characteristics of the problem must be inferred from observations, or when the complexity of the problem, once formalized, renders rational calculation infeasible. Faced with some degree of ignorance, we have two options. First, we can develop a rational actor model by making the assumptions necessary to recast the problem as a certain, small world problem. Second, we can acknowledge that an optimal response to an inherently uncertain problem is a fiction, and likely to obscure our understanding (Klein 2001). Instead, we should aim to examine the relative ability of competing models to explain the observations. The best model we can find is neither optimal nor rational, merely functional.

Put in these terms, at least two factors need to be considered in any proposed definition of a large world. First, one must gauge the impact of the basic uncertainty categories we introduced earlier, such as underspecification, misspecification, and nonstationarity. From our perspective, real-world problems of theoretical significance tend not to be small world problems. We assume a large world by default, which is why we study heuristics rather than rational actor models. Our goal is to understand how, in mechanistic terms, organisms cope with uncertainty in the large worlds to which they are adapted. The second factor to consider is the analytic tractability of performing rational calculation. Thus, it would be an oversight not to mention here that exact computation of the posterior is only possible for certain priors; brute force calculation is only tractable for very restricted hypothesis spaces; and computationally tractable algorithms rely almost exclusively on inexact methods which have optimality guarantees which are asymptotic in nature (e.g. Bishop 2006; MacKay 2003).

Rational principles themselves can be disputed, even in small worlds. For classification tasks, the problem of assigning a class $y \in Y$ to a novel

observation $\mathbf{x}$, the gold standard is the Bayes optimal classifier, which given a sample of observations S, a hypothesis space $\mathcal{H}$, and prior $\Pr(\mathcal{H})$, assigns a class y to an observation by

$$y = \max_{y_i \in Y} \sum_{H_j \in \mathcal{H}} \Pr(y_i \mid \mathbf{x}, H_j) \Pr(\mathcal{H}_j \mid S). \tag{5.8}$$

The class y assigned to $\mathbf{x}$ is calculated by considering the predictions of all hypotheses in $\mathcal{H}$, with each prediction weighted by the posterior probability of the hypothesis. Despite Equation 5.8 being widely regarded as defining an optimal response for classification problems, Grünwald and Langford (2007) have proven its suboptimality under certain forms of misspecification (see also Diaconis and Freedman 1986). More broadly, rationality principles used in the study of inductive inference are necessarily based on statistical assumptions. For instance, Bayesian rationality and algorithmic information theory are closely related due to the relationship between probability and coding (which we drew on when explaining Kullback-Leibler divergence), but they can differ both in theory (Vitányi and Li 2000; Grünwald 2005) and practice (Kearns et al. 1997; Pitt et al. 2002). Subtleties such as these highlight that "rationality principles are invented rather than discovered" (Binmore 2009, p. 2) and we should question the view that there is one rationality that follows directly from the laws of probability theory, logic, or some other calculus.

In our introduction, we used the example of the aircraft manufacturer, and argued that no engineer can hope to arrive at an accurate model of relevant mechanical, atmospheric, and human factors which determine the safety of a flight control system. Furthermore, even if this were possible, the analytic task of finding an optimal response to an appropriate criterion, even if such a criterion exists in any meaningful sense, is beyond the abilities of any expert. We view the significant questions facing the study of inductive inference as having more in common with this large world engineering problem than with the development of rational actor models for small world problems. The study of large world inference problems, like the study of flight control systems, is the search for mechanisms which are more and more robust to uncertainty, but also the study of when and why these mechanisms prove robust (Kitano 2004; Wagner 2005; Hammerstein et al. 2006). This second task can certainly profit from the study of rational actor models and optimality modeling. But this fact in no way questions or undermines the substantive, empirical issue of understanding robust responses to large worlds. Rational inquiry

should not be confused with rational actor models. The former is a commitment to sound science, but the latter is a commitment to a particular approach to modeling.

5.6 ARE RATIONAL ACTOR MODELS "RATIONAL" OUTSIDE SMALL WORLDS?

Savage introduced the notion of a small world when questioning the validity of Bayesian decision theory under conditions of uncertainty. We see Savage's question as applying more broadly, beyond decision theory, and impacting directly on the two key questions which guide the study of inductive inference in humans and other animals. First, how, in mechanistic terms, do organisms arrive at inductive inferences? Second, what is the relationship between the behavior we observe in organisms, and the optimal behavior as defined by a rational actor model? Now, in attempting to answer the second question, do we run the danger of committing a Type III error, and finding the right answer to the wrong question? In contrast to the idea that humans are biased, error-prone users of dumb heuristics, there is a growing tendency to view humans as astonishingly well-adapted to an uncertain world, as evidenced by the ability of humans to handle uncertainty in ways which exceed even the most sophisticated human-engineered machinery (Geman et al. 1992; Poggio and Smale 2003; Tenenbaum et al. 2006). In this sense, evolved biological organisms can be seen as existence proofs of adaptive responses to large, uncertain worlds. This perspective makes the first question, the question of uncovering the underlying cognitive and perceptual mechanisms of humans and other animals, the key question.

Rather than hallmarks of advanced understanding, we have argued that the concepts of rationality and optimality have the potential to be counterproductive when examining large worlds. Without doubt, Bayesian decision theory and Bayesian statistics, to take two examples, inform the question of what is rational in a small world. The crux of the issue is then to assess the dangers of using the same concepts to understand large worlds. While it is possible to analyze uncertain worlds by introducing assumptions which discharge uncertainty, and recast the problem as a small world problem, we have argued for an alternative. The alternative is to follow a relativist approach which makes no attempt to identify an optimal, rational response, but instead aims to incrementally improve understanding by identifying models with greater and greater predictive accuracy. Crucially, this approach dispenses with the need to

make the assumptions necessary to recast the problem as a small world problem. For some problems, we have to accept that our uncertainty and ignorance make finding optimal, rational responses a meaningless endeavor. This fact, though, does not make the study of uncertainty meaningless.

REFERENCES

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14: 471–517.

Binmore, K. (2009). *Rational Decisions*. Princeton University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Bookstaber, R. and Langsam, J. (1985). On the optimality of coarse behavior rules. *Journal of Theoretical Biology*, 116: 161–193.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–231.

Brighton, H. (2006). Robust inference with simple cognitive models. In Lebiere, C. and Wray, R., editors, *Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems* (AAAI Technical Report SS-02-06), pages 189–211. AAAI Press, Menlo Park, CA.

Brighton, H. and Gigerenzer, G. (2007). Bayesian brains and cognitive mechanisms: Harmony or dissonance? In Chater, N. and Oaksford, M., editors, *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, pages 1179–1191. Cambridge University Press.

(2011). Towards competitive instead of biased testing of heuristics: A reply to Hilbig and Richter (2011). *Topics in Cognitive Science*, 3: 197–205.

Brighton, H. and Olsson, H. (2009). Identifying the optimal response is not a necessary step toward explaining function. *Behavioral and Brain Sciences*, 32: 85–86.

Chater, N., Oaksford, M., Nakisa, R., and Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes*, 90: 63–86.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York, NY.

Czerlinski, J., Gigerenzer, G., and Goldstein, D. G. (1999). How good are simple heuristics? In Gigerenzer, G., Todd, P. M., and ABC Research Group, *Simple Heuristics That Make Us Smart*, pages 119–140. Oxford University Press.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1): 1–26.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29: 103–130.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications, Thousand Oaks, CA.

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1: 55–77.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4: 1–58.

Giere, R. G. (1999). *Science without Laws*. University of Chicago Press.

Gigerenzer, G. and Brighton, H. (2009). *Homo heuristicus:* Why biased minds make better inferences. *Topics in Cognitive Science*, 1: 107–143.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62: 451–482.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4): 650–669.

Gigerenzer, G., Todd, P. M., and ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press.

Grünwald, P. (2005). Minimum description length tutorial. In Grünwald, P., Myung, I. J., and Pitt, M. A., editors, *Advances in Minimum Description Length*, pages 23–79. MIT Press, Cambridge, MA.

Grünwald, P. and Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66: 119–149.

Hammerstein, P., Hagen, E. H., Herz, A. V. M., and Herzel, H. (2006). Robustness: A key to evolutionary design. *Biological Theory*, 1(1): 90–93.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21: 1–14.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1): 153–161.

Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42: 80–86.

Hopcroft, J. E. and Ullman, J. D. (1979). *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, MA.

Kearns, M., Mansour, Y., Ng, A. Y., and Ron, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27: 7–50.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5: 826–837.

Klein, G. (2001). The fiction of optimization. In Gigerenzer, G. and Selten, R., editors, *Bounded Rationality: The Adaptive Toolbox*, pages 103–121. MIT Press, Cambridge, MA.

Knight, F. H. (1921). *Risk, Uncertainty, and Profit*. Houghton Mifflin Co., Boston, MA.

Korzybski, A. (1958). *Science and Sanity*, 4th edition. International Non-Aristotelian Library Publishing Co., Lakeville, CT.

Kuncheva, L. I. (2006). On the optimality of Naïve Bayes with dependent binary features. *Pattern Recognition Letters*, 27: 830–837.

Lee, M. D., Loughlin, N., and Lundberg, I. B. (2002). Applying one reason decision-making: The prioritization of literature searches. *Australian Journal of Psychology*, 54: 137–143.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press.

McNamara, J. M. and Houston, A. I. (2009). Integrating function and mechanism. *Trends in Ecology and Evolution*, 24: 670–674.

Oaksford, M. and Chater, N., editors (1998). *Rational Models of Cognition.* Oxford University Press.

(2007). *Bayesian Rationality.* Oxford University Press.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1: 502–518.

Papadimitriou, C. H. (1995). *Computational Complexity.* Addison-Wesley, Menlo Park, CA.

Perlich, C., Provost, F., and Simonoff, J. S. (2003). Tree induction vs. logistic regression: A learning curve analysis. *Journal of Machine Learning Research*, 4: 211–255.

Pitt, M. A., Myung, I. J., and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109(3): 472–491.

Poggio, T. and Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society*, 50: 537–544.

Popper, K. R. (1959). The propensity interpretation of probability. *British Journal of the Philosophy of Science*, 10: 25–42.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., editors (2009). *Dataset Shift in Machine Learning.* MIT Press, Cambridge, MA.

Rissanen, J. (1986). Stochastic complexity and statistical inference. *Lecture Notes in Control and Information Sciences*, 83: 393–407.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4): 1144–1167.

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90: 227–244.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36: 111–147.

Taleb, N. N. (2009). Errors, robustness, and the fourth quadrant. *International Journal of Forecasting*, 25: 744–759.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7): 309–318.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

Vitányi, P. M. B. and Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, 46(2): 446–464.

Wagner, A. (2005). *Robustness and Evolvability in Living Systems*. Princeton University Press.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14: 1–37.

Wübben, M. and Wangenheim, F. von (2008). Instant customer base analysis: Managerial heuristics often get it right. *Journal of Marketing*, 72: 82–93.