A Rational Analysis of Rule-based Concept Learning

Noah D. Goodman and Joshua B. Tenenbaum
Massachusetts Institute of Technology

Jacob Feldman
Rutgers University

Thomas L. Griffiths
University of California, Berkeley

We propose a new model of human concept learning that provides a rational analysis of learning feature-based concepts. This model is built upon Bayesian inference for a grammatically structured hypothesis space—a concept language of logical rules. We compare the model predictions to human generalization judgments in several well-known category learning experiments, and find good agreement for both average and individual participants generalizations. We further investigate judgments for a broad set of seven-feature concepts—a more natural setting in several ways—and again find that the model explains human performance.

But what are concepts save formulations and creations of thought, which, instead of giving us the true form of objects, show us rather the forms of thought itself?
—Cassirer (1946, p. 7)

The study of concepts—what they are, how they are used and how they are acquired—has provided one of the most enduring and compelling windows into the structure of the human mind. What we look for in a theory of concepts, and what kinds of concepts we look at, depend on the functions of concepts that interest us. Three intuitions weave throughout the cognitive science literature (see, e.g. Fodor, 1998; Murphy, 2002):

1. *Concepts are mental representations that are used to discriminate between objects, events, relations, or other states of affairs.* Cognitive psychologists have paid particular attention to concepts that identify kinds of things—those that classify or categorize objects—and such concepts will also be our focus here. It is clear how an ability to separate objects according to kind could be critical to survival. To take a classic example, a decision about whether it is appropriate to pursue or to flee an intruder depends on a judgment of kind (prey or predator), and an error in this judgment could have disastrous consequences.

2. *Concepts are learned inductively from the sparse and noisy data of an uncertain world.* Animals make some instinctive discriminations among objects on the basis of kind, but cognition in humans (and probably other species) goes beyond an innate endowment of conceptual discriminations. New kind-concepts can be learned, often effortlessly despite great uncertainty. Even very sparse and noisy evidence, such as a few randomly encountered examples, can be sufficient for a young child to accurately grasp a new concept.

3. *Many concepts are formed by combining simpler concepts, and the meanings of complex concepts are derived in systematic ways from the meanings of their constituents.* Concepts are the constituents of thought, and thought is in principle unbounded, though human thinkers are clearly bounded. The "infinite use of finite means" (Humboldt, 1863) can be explained if concepts are constructed, as linguistic structures are constructed, from simpler elements: for example, morphemes are combined into words, words into phrases, phrases into more complex phrases and then sentences.

In our view, all of these intuitions about concepts are central and fundamentally correct, yet previous accounts have rarely attempted to (or been able to) do justice to all three. Early work in cognitive psychology focused on the first theme, concepts as rules for discriminating among categories of objects (Bruner, Goodnow, & Austin, 1956). Themes two and three were also present, but only in limited ways. Researchers examined the processes of learning concepts from examples, but in a deductive, puzzle-solving mode more than an inductive or statistical mode. The discrimination rules considered were constructed compositionally from simpler concepts or perceptual features. For instance, one might study how people learn a concept for picking out objects as "large and red and round". An important goal of this research program was to characterize which kinds of concepts were harder or easier to learn in terms of syntactic measures of a concept's complexity, when that concept was expressed as a combination of simple perceptual features. This approach

reached its apogee in the work of Shepard, Hovland, and Jenkins (1961) and Feldman (2000), who organized possible Boolean concepts (those that discriminate among objects representable by binary features) into syntactically equivalent families and studied how the syntax was reflected in learnability.

A second wave of research on concept learning, often known as the "statistical view" or "similarity-based approach", emphasized the integration of themes one and two in the form of inductive learning of statistical distributions or statistical discrimination functions. These accounts include prototype theories (Posner & Keele, 1968; Medin & Schaffer, 1978), exemplar theories (Shepard & Chang, 1963; Nosofsky, 1986; Kruschke, 1992), and some theories in between (Anderson, 1990; Love, Gureckis, & Medin, 2004). These theories do not rest on a compositional language for concepts and so have nothing to say about theme three—how simple concepts are combined to form more complex structures (Osherson & Smith, 1981).

An important recent development in the statistical tradition has been the rational analysis of concept learning in terms of Bayesian inference (Shepard, 1987; Anderson, 1990; Tenenbaum & Griffiths, 2001). These analyses show how important aspects of concept learning—such as the exponential-decay gradient of generalization from exemplars (Shepard, 1987) or the transitions between exemplar and prototype behavior (Anderson, 1990)—can be explained as approximately optimal statistical inference given limited examples. However, these rational analyses have typically been limited by the need to assume a fixed hypothesis space of simple candidate concepts—such as Shepard's (1987) "consequential regions": connected regions in a low-dimensional continuous representation of stimuli. The standard Bayesian framework shows how to do rational inductive learning given such a hypothesis space, but not where this hypothesis space comes from nor how learners can go beyond the simple concepts it contains when required to do so by the complex patterns of their experience.

The last decade has also seen renewed interest in the idea that concepts are constructed by combination, as a way of explaining apparently unbounded human conceptual resources. This is especially apparent in accounts of concepts and concept learning that place compositionality at center stage (Fodor, 1998; Murphy, 2002). Logical or rule-based representations are often invoked. Most relevant to our work here is the rules-plus-exceptions (RULEX) model of Nosofsky, Palmeri, and McKinley (1994), which is based on a set of simple heuristics for constructing classification rules, in the form of a conjunction of features that identify the concept plus a conjunction of features which identify exceptions. The RULEX model has achieved strikingly good fits to classic human concept-learning data, including some of the data sets that motivated statistical accounts, but it too has limitations. Fitting RULEX typically involves adjusting a number of free parameters, and the model has no clear interpretation in terms of rational statistical approaches to inductive learning. Perhaps most importantly, it is unclear how to extend the rule-learning heuristics of RULEX to more complex rep-

resentations. Therefore, while RULEX uses a compositional representation, it is unable to fully leverage compositionality.

Our goal here is a model that integrates all three of these major themes from the literature on concepts and concept learning. Our *Rational Rules* model combines the inferential power of Bayesian induction with the representational power of mathematical logic and generative grammar; the former accounts for how concepts are learned under uncertainty while the latter provides a compositional hypothesis space of candidate concepts. This paper is only a first step towards an admittedly ambitious goal, so we restrict our attention to some of the simplest and best-studied cases of concept learning from the cognitive psychology literature. We hope readers will judge our contribution not by these limits, but by the insights we develop for how to build a theory that captures several deep functions of concepts that are rarely integrated and often held to be incompatible. We think these insights have considerable generality.

Our approach can best be understood in the tradition of rational modeling (Anderson, 1990; Oaksford & Chater, 1998), and specifically rational models of generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001). The main difference with earlier work is the adoption of a qualitatively richer and more interesting form for the learner's hypothesis space. Instead of defining a hypothesis space directly as a set of subsets of objects, with no intrinsic structure or constructive relations between more complex hypotheses and simpler ones, we work with a compositional hypothesis space generated by a probabilistic grammar. The grammar yields a "concept langauge" which describes a range of concepts varying greatly in complexity, all generated from a small basis set of features or atomic concepts. Hypotheses range from the simplest, single-feature, rules to complex combinations of rules needed to describe an arbitrary discrimination boundary. The prior probability of each hypothesis is not specified directly, by hand, but rather is generated automatically by the grammar in a way that naturally favors the simpler hypotheses. By performing Bayesian inference over this concept language, a learner can make rational decisions about how to generalize a novel concept to unseen objects given only a few, potentially noisy, examples of that concept. The resulting model is successful at predicting human judgments in a range of concept learning tasks at both the group and individual level, using a minimum of free parameters and arbitrary processing assumptions.

This analysis is an advance for several viewpoints on concepts and concept learning. From the vantage of rule-based models of categorization (e.g. RULEX) the Rational Rules model provides a rational inductive logic explaining *why* certain rules should be extracted. This naturally complements the insights of existing process-level accounts by tying the rule-learning competency to general principles of learning and representation. To rational statistical accounts of concept learning, the grammar-based approach of the Rational Rules model contributes a more satisfying account of the hypothesis space and prior: from simple components the grammar compactly specifies an infinite, flexible hypothesis space of structured rules and a prior that controls complexity.

For compositionality advocates, we provide a way to use the "language of thought" (Fodor, 1975) (reified in the concept language of logical rules) to do categorization under uncertainty. That is, the grammar-based induction approach suggests a compelling way to quantitatively relate the language of thought, already an elegant and rational view of the mind, to human behavior, even at the level of predicting individual participants. Finally, by showing that the rational statistical approach to concepts is compatible with the rule-based, combinatorial, approach, and that the union can accurately predict human behavior, we hope to cut one of the gordian knots of of modern psychology: the supposed dichotomy between rules and statistics.

Preliminary Sketch

In this section we will sketch out a notion of concepts, and concept learning, using the three intuitions above; in the next section we formalize this discussion.

Concepts are mental representations, that is they are things "in the head" which have some structure that reflects the world. Since we are concerned with concepts that are used to discriminate between objects, a natural hypothesis is that concepts are simply rules for classifying objects based on their features. This hypothesis is bolstered by the common reports of participants in concept learning experiments that they "feel as if" they are using a rule (e.g. Armstrong, Gleitman, & Gleitman, 1983). How can we specify the structure of concepts, if we accept as a working hypothesis that they are rules? Since many concepts are formed by combining simpler concepts, many of the corresponding rules should be built by combining other rules.

The classic examples of such combinatory syntax come from generative grammar, and particularly context-free grammars (CFGs) (see e.g. Manning & Schütze, 1999; Russell & Norvig, 2002). A CFG is a collection of terminal symbols, non-terminal symbols, and production rules. For instance, we could have a CFG with single non-terminal A, two terminals a, b, and two productions $A \rightarrow Aa$ and $A \rightarrow b$. Beginning with A this grammar generates strings by applying the productions until no non-terminal remains: $A \rightarrow Aa \rightarrow Aaa \rightarrow baa$, etc. The strings of terminal symbols formed by applying production rules are the *language* specified by the CFG—for the example all strings of the form $ba...a$. Note that production rules specify possible expansions of single symbols, and can be seen as combination operations "in reverse" (for instance reversing the production $A \rightarrow Aa$ allows the combination of an A-constituent with the primitive a to get another A-constituent). We adopt CFGs to provide the combinatory syntax for our concepts, thus we have a *concept language* describing the representations of rules. However, a CFG describes only the structure of representations, if this structure is to influence the use of concepts it must be mirrored in the meaning we assign to a representation—each syntactic combination of rules must give us a new rule able to discriminate among objects in the world. Fortunately, if we think of rules as functions from objects to truth values (where $True$ is arbitrarily assigned to

"in the concept"), then there is a natural set of combination operations: the logical connectives. Indeed, classical logic is a paradigmatic example of a language with CFG syntax, a semantics of discriminative rules, and combination operations that work on both levels. If we supplement with a set of primitive concepts (which are features or, perhaps, other pre-existing concepts), mathematical logic appears to provide a simple and intuitive framework for concepts.

Before we embrace this framework for concepts, we should reflect on the reasons why rule-based representations have had such a troubled past in psychology. Since an underlying commitment of rule-based accounts (at least as classically interpreted) is that concepts have deterministic definitions, there seems to be no room for one object which satisfies the definition to be a "better" example than another. Rosch and colleagues showed that people are very willing to assign graded typicality to concepts, and that this gradation is reflected in many aspects of concept use (e.g. Mervis & Rosch, 1981). Combined with the difficulty of identifying definitions for common concepts (Wittgenstein, 1953), these results led many authors to suggest that the organizing principle of concepts is *similarity*—for example via "family resemblance" among exemplars—not rules. The conventional wisdom regarding the virtues of rule-based versus similarity-based models of cognition is that "rules provide precision, expressiveness, and generativity, and similarity provides flexibility and the means to deal with uncertainty" (Sloman, Love, & Ahn, 1998). Modern cognitive psychology has been especially concerned with the "fuzzy edges"—capturing the ways that people deal with uncertainty—so it is natural that similarity-based models of concept learning have come to dominate. Is it true, however, that proper treatment of uncertainty is antithetical to rule-based representation? In concept learning uncertainty arises primarily in two ways: examples are unreliable, and available evidence is sparse. We believe that these can be addressed in ways that are agnostic to representation: allow that some experiences are outliers, and employ a strong inductive bias, respectively. Both of these responses to uncertainty can be realized by viewing concept learning as an inductive inference problem.

A general approach to the analysis of inductive learning problems has emerged in recent years (Anderson, 1990; Tenenbaum, 1999b; Chater & Oaksford, 1999). Under this approach a set of hypotheses is posited, and degrees of belief are assigned using Bayesian statistics. This assignment relies on: a description of the data space from which input is drawn, a space of hypotheses, a prior probability function over these hypotheses, and a likelihood function relating each hypothesis to the data. The prior probability, $P(h)$, describes the belief in hypothesis h before any data is seen, and hence captures prior knowledge. The likelihood, $P(d|h)$, describes the data one would expect to observe if hypothesis h were correct. With these components, inductive learning can be described very simply: we wish to find the appropriate degree of belief in each hypothesis given some observed data—the posterior probability $P(h|d)$. Bayes' theorem tells us how to

compute this probability,

$$P(h|d) \propto P(h)P(d|h), \qquad (1)$$

identifying the posterior probability as proportional to the product of the prior and the likelihood. This Bayesian posterior provides a rational analysis of inductive learning: a coherent integration of evidence and a priori knowledge into posterior beliefs (optimal within the specified learning context).

By viewing concept acquisition as an inductive problem, we may employ this Bayesian framework to describe the learning of a concept from examples. As described earlier, our hypothesis space is the collection of all phrases in a grammatically-generated concept language. Since each concept of this language is a classification rule, a natural likelihood is given by simply evaluating this rule on the examples: the data have zero probability if they disagree with the classification rule, and constant probability otherwise. However, to account for the unreliability of examples, we will allow non-zero probability for a data set, even if some examples are misclassified. That is, we assume that there is a small probability that any example is an *outlier*, which should be ignored. We will see in the next section that this outlier assumption combines with rule-evaluation to give a simple likelihood function, which decreases exponentially with the number of "misclassifications" of the examples.

In Bayesian models the prior, $P(h)$, provides the inductive bias needed to solve under-constrained learning problems. Hence, equipping the concept language with an appropriate prior can address the uncertainty engendered by sparse evidence. A natural prior follows by extending the grammar into a probabilistic context free grammar. By viewing the generative process for phrases, which is specified by the grammar, as a probabilistic process, we get a probability distribution on phrases of the concept language. We will see that this prior has a *syntactic complexity bias*: the prior probability of a combined rule is less than the prior probability of either component. In fact, the prior probability of a rule decreases, roughly exponentially, in the number of symbols used to express it. There is some empirical evidence that the number of primitive feature symbols in a rule, called its *Boolean complexity*, is relevant to the inductive bias of human concept learning. Indeed, Feldman (2000) found that the Boolean complexity is a good predictor of the difficulty of remembering a wide variety of binary concepts. Feldman (2006) showed that other aspects of the algebraic complexity of concepts predict further facets of human learning and use. This suggests that the natural inductive bias provided by the grammar of our concept language may be sufficient to describe human learning, and particularly the ways that human learning copes with the uncertainty of sparse examples.

Using this formulation of concept learning as a Bayesian induction problem, we can address, within a rule-based framework, the uncertainty inherent in concept learning. How does this overcome the determinism of rule-based representations, which was such a stumbling block to early theories? It has been noted before (Shepard, 1987; Tenenbaum,

2000) that graded effects can arise out of mixtures of deterministic representations, and that such mixtures result from rational Bayesian use of deterministic representations under uncertainty. Though our ideal Bayesian learner is committed to there being *a* correct definition for each concept, there is rarely enough information to determine this correct definition completely. Instead, the posterior belief function will be spread over many different definitions. This spread can result in graded, similarity-like effects in the classification behavior of the ideal agent, or in more deterministic rule-like classification, depending on the pattern of examples the agent observes and the shape of the posterior distribution they give rise to.

An Analysis of Concept Learning

In light of the above discussion, we wish to formulate a concept language of rules, and analyze the behavior of a rational agent learning concepts expressed in this language. This analysis will describe an ideal concept learner that satisfies the three intuitions with which we began this paper. In later sections we will explore the relationship between human concept acquisition and this ideal learning model.

The learning problem can be phrased, using the Bayesian induction formalism, as that of determining the posterior probability $P(F|\mathbf{E}, \ell(\mathbf{E}))$, where F ranges over formulae in the concept language, $\mathbf{E}$ is the set of observed example objects (possibly with repeats) and $\ell(\mathbf{E})$ are the observed labels. Thus concept learning will roughly amount to "probabilistic parsing" of labeled examples into representations of the concept language. (We consider a single labeled concept, thus $\ell(x) \in \{1, 0\}$ indicates whether x is an example or a non-example of the concept. An example is given in Table 1, which we will use throughout this section.) The posterior may be expressed (through Bayes' formula) as:

$$P(F|\mathbf{E}, \ell(\mathbf{E})) \propto P(F)P(\mathbf{E}, \ell(\mathbf{E})|F) \qquad (2)$$

To use this relationship we will need our concept language (which describes the hypothesis space), the prior probability, $P(F)$, and a likelihood function, $P(\mathbf{E}, \ell(\mathbf{E})|F)$.

Concept Representation

The concept language in which we will represent rules is a fragment of first-order logic. This will later allow us to use the standard truth-evaluation procedure of mathematical logic in defining our likelihood. The terminal symbols of the language (those which can appear in finished rules) are logical connectives ($\wedge, \vee, \Leftrightarrow$), grouping symbols, a quantifier over objects ($\forall\ x$) with quantified variable x, and a set of feature predicates. The feature predicates can be thought of as simple preexisting concepts, or as perceptual primitives. (Thus each concept, once learned, potentially becomes a "feature" for future concepts; for a related view see Schyns, Goldstone, and Thibaut (1998).) We will focus on simple feature predicates formed from functions $f_i(x)$, which report the value of a physical feature, and operators $=c$, $<c$, and $>c$, which represents comparison with constant c. Initially each

Table 1

A set of labeled examples E={A1,...,A5,B1,...,B4} (from Medin & Schaffer (1978)). This example set admits, for example, the simple definition D_1: $f_1(x)=0$, by assuming A5 and B1 are outliers, and the more complex definition D_2: $(f_1(x)=0 \land f_3(x)=0) \lor (f_2(x)=0 \land f_4(x)=0)$, with no outliers.

Object	Label ℓ	Feature Values
A1	1	0001
A2	1	0101
A3	1	0100
A4	1	0010
A5	1	1000
B1	0	0011
B2	0	1001
B3	0	1110
B4	0	1111

feature predicate is of the form $f_i(x)=c$ (read "the ith feature of object x has value c"), with Boolean values ($c \in \{0, 1\}$). The extension to continuous-valued features by using the inequality comparison operators is straightforward, and will be used later in the paper.

(S1)	$S \rightarrow \forall x\, \ell(x) \Leftrightarrow (D)$	
(D1)	$D \rightarrow (C) \lor D$	
(D2)	$D \rightarrow False$	
(C1)	$C \rightarrow P \land C$	
(C2)	$C \rightarrow True$	
(P1)	$P \rightarrow F_1$	

$$\vdots$$

(PN)	$P \rightarrow F_N$	
($F_1 1$)	$F_1 \rightarrow f_1(x) = 1$	
($F_1 2$)	$F_1 \rightarrow f_1(x) = 0$	

$$\vdots$$

($F_N 1$)	$F_N \rightarrow f_N(x) = 1$	
($F_N 2$)	$F_N \rightarrow f_N(x) = 0$	

Figure 1. Production rules of the DNF grammar. S is the start symbol, and D, C, P, F_i the other non-terminals. Productions (F_i1) and (F_i2) can be naturally extended to "decision boundary" predicates, e.g., $F_1 \rightarrow f_1(x) < 2$.

The set of formulae in our language is generated by the context-free "disjunctive normal form", or DNF, grammar, Fig. 1. Informally, each formula provides a "definition" and asserts that that definition must hold anytime the label is true[1]: $\forall x\, \ell(x) \Leftrightarrow D$. Each definition has the form of a standard dictionary entry: a set of alternative "senses", each of which is a list of necessary and sufficient conditions on the features. For example, the labeled examples in Table 1 admit an imperfect definition with a single sense $(f_1(x)=0)$ and a complex definition with two senses $(f_1(x)=0 \land f_3(x)=0)$ and $(f_2(x)=0 \land f_4(x)=0)$. More formally, each D non-terminal becomes, by productions (D1) and (D2), a disjunction of C

non-terminals (the "senses"); each C-term becomes a conjunction of predicate P-terms, and each P-term becomes a specific feature predicate. Let us illustrate the generative process of the DNF grammar by considering the two examples from Table 1.
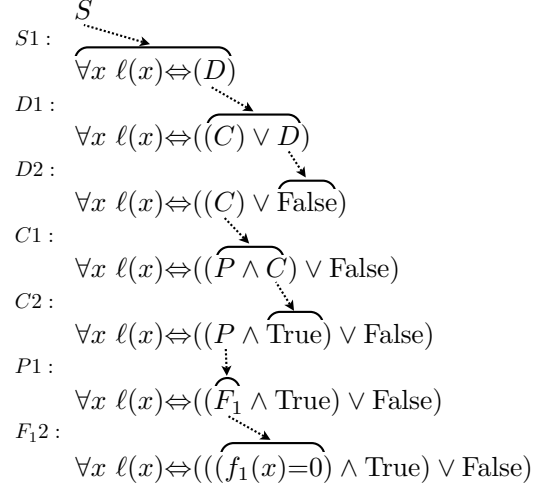

$$S$$
$$S1: \quad \overbrace{\forall x\, \ell(x) \Leftrightarrow (D)}$$
$$D1: \quad \forall x\, \ell(x) \Leftrightarrow (\overbrace{(C) \lor D})$$
$$D2: \quad \forall x\, \ell(x) \Leftrightarrow ((C) \lor \overbrace{False})$$
$$C1: \quad \forall x\, \ell(x) \Leftrightarrow ((\overbrace{P \land C}) \lor False)$$
$$C2: \quad \forall x\, \ell(x) \Leftrightarrow ((P \land \overbrace{True}) \lor False)$$
$$P1: \quad \forall x\, \ell(x) \Leftrightarrow ((\overbrace{F_1} \land True) \lor False)$$
$$F_1 2: \quad \forall x\, \ell(x) \Leftrightarrow (((\overbrace{f_1(x)=0}) \land True) \lor False)$$

Figure 2. Derivation of a formula from the DNF grammar.

First consider the derivation illustrated in Fig. 2. Beginning with the start symbol, S, the first step is to use production (S1) to derive $\forall x\, \ell(x) \Leftrightarrow D$. Next we expand the symbol D by applying production (D1), then production (D2). This leads to a single conjunct term:

$$\forall x\, \ell(x) \Leftrightarrow ((C) \lor False)$$

We continue by expanding the C-term using productions (C1) then (C2):

$$\forall x\, \ell(x) \Leftrightarrow ((P \land True) \lor False)$$

Using production (P1), the P-term is replaced with F_1, and using production ($F_1 2$) this becomes an assertion of the value for the first feature[2]:

$$\forall x\, \ell(x) \Leftrightarrow (f_1(x)=0)$$

We have thus derived a simple formula incorporating definition D_1 for the concept of Table 1.

More complex derivations are also possible from this grammar. For example, let us consider the derivation of the complex definition D_2 from Table 1. Beginning with $\forall x\, \ell(x) \Leftrightarrow D$, we expand the symbol D by applying production (D1) twice (instead of once), then production (D2). This

---

[1] In later sections we often write only the definition part of the formulae, for clarity.

[2] The terminal symbols *True* and *False* stand for logical True and False—they are used to conveniently terminate a string of conjunctions or disjunctions, respectively, and can be ignored. We now drop them, and unnecessary parentheses, for clarity.

leads to a disjunction of two conjunct terms (the two "senses" of the definition). We now have the rule:

$$\forall x \, \ell(x) \Leftrightarrow (C \vee C)$$

Recall that C is a non-terminal, so each of these C-terms can ultimately result in a distinct substring (and similarly for the other non-terminals). Each non-terminal symbol C leads, by productions (C1) and (C2), to a conjunction of predicate terms:

$$\forall x \, \ell(x) \Leftrightarrow ((P \wedge P) \vee (P \wedge P))$$

Using production(s) (Pi), each predicate term becomes a feature predicate F_i, for one of the N features:

$$\forall x \, \ell(x) \Leftrightarrow ((F_1 \wedge F_3) \vee (F_2 \wedge F_4))$$

Finally, with productions (F_i1) and (F_i2), each feature predicate becomes an assertion that the ith feature has a particular value ($f_1(x)$=0, etc.):

$$\forall x \, \ell(x) \Leftrightarrow ((f_1(x)=0 \wedge f_3(x)=0) \vee (f_2(x)=0 \wedge f_4(x)=0))$$

Informally, this means that the label holds when: f_1 and f_3 are zero, or f_2 and f_4 are zero—this accurately classifies all the examples of Table 1.

Thus far we have used the generative nature of the DNF grammar only to specify which sequences of symbols are syntactically well formed (that is, those which represent valid concepts). However, generative processes can also be used to induce probability distributions: we can induce a probability distribution over the formulae of the concept language by providing a probability for each choice in the derivation process. Thus, the simple generative process that allows us to build syntactic formulae will also provide a prior over these formulae.

A Syntactic Prior

As illustrated above, each formula is generated from the start symbol S by a *derivation*: a sequence of productions, each replacing a single non-terminal, that ends when there are no non-terminals left to replace. At each step of a derivation we choose from among the productions which could be used to expand the next non-terminal symbol[3]—and if we assign a probability to each such choice we will have a probability for the complete derivation. Hence, by supplementing the context-free grammar with probabilities for the productions we get a prior over the formulae of the language: each production choice in a derivation is assigned a probability, and the probability of the complete derivation is the product of the probabilities for these choices. Of course, the set of production probabilities, τ, must sum to one for each non-terminal symbol. The probability of a derivation Deriv_F for formula F is:

$$P(\text{Deriv}_F | \mathcal{G}, \tau) = \prod_{s \in \text{Deriv}_F} \tau(s), \qquad (3)$$

where $s \in \text{Deriv}_F$ are the productions in the derivation, $\tau(s)$ the probability of each, and $\mathcal{G}$ denotes the grammar. The

DNF grammar is unambiguous—there is a single derivation for each well-formed formula[4]—so Eq. 3 gives the probability of the formula itself: $P(F|\mathcal{G}, \tau) = P(\text{Deriv}_F|\mathcal{G}, \tau)$. For a generic context-free grammar, $P(F|\mathcal{G}, \tau)$ would instead be the sum of the probabilities of its derivations. This would complicate our analysis, but not in any critical way.

Note that the prior in Eq. 3 captures a syntactic simplicity bias: smaller formulae have shorter derivations, thus higher prior probability. However, the precise values of the production probabilities may affect the inductive bias in important ways. For instance, if production (D2) (of the DNF grammar, Fig. 1) is much more likely that production (D1), but productions (C1) and (C2) are about equally likely, then complexity as measured by the number of disjunctions will be penalized more heavily than complexity as measured by the number of conjunctions. Given this sensitivity, how should a rational agent choose production probabilities? Any specific choice would be ad-hoc, and would preclude learning correct values from experience. Rather than committing to a specific choice, we can maintain uncertainty over τ. Since we have no a priori reason to prefer one set of values for τ to another we assume a uniform prior over the possible values—that is, we apply the principle of indifference (Jaynes, 2003). Now the prior probability is over both the formula and the production probabilities: $P(F, \tau|\mathcal{G})$. However, since we are primarily concerned with learning formulae F, we marginalize out τ (see Appendix A). After some simplification this leads to the expression:

$$P(F|\mathcal{G}) = \prod_{Y \in \text{ non-terminals of } \mathcal{G}} \frac{\beta(\mathbf{C}_Y(F) + \mathbf{1})}{\beta(\mathbf{1})}. \qquad (4)$$

Here $\beta(\mathbf{c})$ is the multinomial beta function (or normalizing constant of the Dirichlet distribution, see Appendix A). The vector $\mathbf{C}_Y(F)$ counts the uses of productions for non-terminal Y: it has an entry for each production s of Y, equal to the number of times s was used in Deriv_F to replace Y. For instance, in the above derivation of the formula $\forall x \, \ell(x) \Leftrightarrow (f_1(x)=0)$, the non-terminal D was expanded once with derivation (2) and once with derivation (3). Thus $\mathbf{C}_D(F)$=(1, 1), and D contributes $\frac{\beta(2,2)}{\beta(1,1)} = 0.1667$ to the product in Eq. 4. In contrast, the non-terminal F_2 is never used in this derivation, so $\mathbf{C}_{F_2}(F)$=(0, 0), and F_2 contributes $\frac{\beta(1,1)}{\beta(1,1)} = 1$ (i.e. nothing) to the product.

This prior probability has several important properties, which we now illustrate with examples. First, because the multinomial beta function is monotonically decreasing in each argument (over the domain which concerns us), this prior will again have a complexity bias penalizing longer

---

[3] There are multiple orders for any derivation, but this freedom can be eliminated by always expanding the right-most non-terminal first. We thus treat derivations as uniquely ordered, without loss of generality.

[4] To see this, note that it would change nothing to add additional parentheses in productions (D1) and (C1), making the formulae fully-parenthesized. The derivation of a formula could then be read off the structure of its parentheses.

derivations (hence longer formulae). For instance, we see in Fig. 2 that the formula $\forall x\, \ell(x) \Leftrightarrow (f_1(x){=}0)$ is derived by applying productions S1, D1, D2, C1, C2, P1, and $F_1$2. The prior probability of this formula is computed by applying Eq. 4, which leads to the product:

$$\overbrace{\frac{\beta(2)}{\beta(1)}}^{S} \times \overbrace{\frac{\beta(2,2)}{\beta(1,1)}}^{D} \times \overbrace{\frac{\beta(2,2)}{\beta(1,1)}}^{C} \times \overbrace{\frac{\beta(2,1,1)}{\beta(1,1,1,1)}}^{P} \times \overbrace{\frac{\beta(2,1)}{\beta(1,1)}}^{F_1}$$
$$= 1.7 \times 10^{-3}.$$

(Each term of this product comes from the non-terminal written over it; the remaining non-terminals are not used for this formula.) In contrast the more complex formula $\forall x\, \ell(x) \Leftrightarrow ((f_1(x){=}0 \wedge f_3(x){=}0) \vee (f_2(x){=}0 \wedge f_4(x){=}0))$ uses a number of the productions repeatedly (as illustrated above), and has prior probability:

$$\overbrace{\frac{\beta(2)}{\beta(1)}}^{S} \times \overbrace{\frac{\beta(3,2)}{\beta(1,1)}}^{D} \times \overbrace{\frac{\beta(5,3)}{\beta(1,1)}}^{C} \times \overbrace{\frac{\beta(2,2,2,2)}{\beta(1,1,1,1)}}^{P} \times$$
$$\overbrace{\frac{\beta(2,1)}{\beta(1,1)}}^{F_1} \times \overbrace{\frac{\beta(2,1)}{\beta(1,1)}}^{F_2} \times \overbrace{\frac{\beta(2,1)}{\beta(1,1)}}^{F_3} \times \overbrace{\frac{\beta(2,1)}{\beta(1,1)}}^{F_4}$$
$$= 5.9 \times 10^{-8}.$$

Again, this complexity bias follows from the monotonicity of the beta functions, and is the dominant contribution to the prior.

Another, subtler, contribution to the behavior of the prior comes from the property: $\beta(i, j, ...) < \beta(i + 1, j - 1, ...)$ if $i \geq j$. This means, in particular, that the prior favors *reuse* of features over use of multiple features. For instance, if we alter the above formula into $\forall x\, \ell(x) \Leftrightarrow ((f_1(x){=}0 \wedge f_3(x){=}0) \vee (f_1(x){=}0 \wedge f_4(x){=}0))$ (so that we use feature f_1 twice, and f_2 never), the formula receives a slightly higher prior probability:

$$\overbrace{\frac{\beta(2)}{\beta(1)}}^{S} \times \overbrace{\frac{\beta(3,2)}{\beta(1,1)}}^{D} \times \overbrace{\frac{\beta(5,3)}{\beta(1,1)}}^{C} \times \overbrace{\frac{\beta(3,1,2,2)}{\beta(1,1,1,1)}}^{P} \times$$
$$\overbrace{\frac{\beta(3,1)}{\beta(1,1)}}^{F_1} \times \overbrace{\frac{\beta(2,1)}{\beta(1,1)}}^{F_3} \times \overbrace{\frac{\beta(2,1)}{\beta(1,1)}}^{F_4}$$
$$= 1.6 \times 10^{-7}.$$

We will show below that this property of the prior allows us to naturally capture certain "selective attention" effects exhibited by human learners.

Likelihood: Evaluation and Outliers

We have given an informal description of the meaning of formulae in our concept language—they are definitions in disjunctive normal form. This meaning is captured formally by the likelihood function $P(\mathbf{E}, \ell(\mathbf{E})|F)$, which specifies the probability of a "world" of labeled examples, given a formula. Here we present the likelihood function, with intuitive justification; in Appendix B we provide a formal derivation of this likelihood.

Our informal description above suggests that examples should be labeled consistently with the classification rule, so we might constrain $P(\mathbf{E}, \ell(\mathbf{E})|F)$ to be uniform on worlds where the formula is true[5], and zero otherwise. If we knew that the observed labels were correct, and we required an explanation for each observation, this constraint would determine the likelihood. However, we wish to allow concepts that explain only some of the observations. For instance, it seems unreasonable to ignore the definition D_1 of Table 1, which is quite simple and correctly predicts seven of the nine examples (though it misses A5 and B1). Hence we assume that there is a small probability that any given example is an outlier (ie. an unexplainable observation which should be excluded from induction). For a given partition of the examples into outliers and in-liers the likelihood will still be a simple step function, but we must sum over outlier sets. This sum produces an exponential gradient[6] (see Appendix B), and the likelihood becomes:

$$P(\ell(\mathbf{E}), \mathbf{E}|F) \propto e^{-bQ_\ell(F)}, \qquad (5)$$

where $Q_\ell(F)$ is the number of example objects which do not satisfy the definition asserted by F. Note that this likelihood decreases exponentially with the number of mislabeled examples, since they must be treated as outliers, but is never zero. The probability that any given example is an outlier is e^{-b}—thus the parameter b is a measure of the *a prior* probability that an example is an outlier. (As we will see below, multiple exposures to an example will effectively decrease the probability that it is an outlier.)

Returning to the examples of Table 1, the simpler definition D_1 incorrectly classifies examples A5 and B1, hence $P(\ell(\mathbf{E}), \mathbf{E}|F_{D_1}) \propto e^{-2b}$. The complex definition D_2 correctly classifies all examples, hence $P(\ell(\mathbf{E}), \mathbf{E}|F_{D_2}) \propto 1$.

The Rational Rules Model

The above likelihood and prior, combined using Bayes' rule, constitute a model of concept learning, which we call the *Rational Rules* model (RR$_{\text{DNF}}$, to indicate the grammar). The posterior probability of formula F according to this model is:

$$P(F|\mathbf{E}, \ell(\mathbf{E})) \propto \left(\prod_{Y \in \text{ non-terminals of } \mathcal{G}} \beta(\mathbf{C}_Y(F) + \mathbf{1}) \right) e^{-bQ_\ell(F)}. \qquad (6)$$

In Table 2 we show the highest posterior formulae given the examples in Table 1, according to the RR$_{\text{DNF}}$ model.

---

[5] The concept language generates formulae of predicate logic. Hence we inherit from the standard truth-functional semantics of mathematical logic a procedure allowing us to decide whether the formula is true.

[6] This is similar to the derivation of Shepard's Law of Generalization (Shepard, 1987) by a sum over consequential regions.

The generalization probability that a test object t has label $\ell(t)=1$ is given by:

$$P(\ell(t)=1|\mathbf{E}, \ell(\mathbf{E})) = \sum_F P(\ell(t)=1|F)P(F|\mathbf{E}, \ell(\mathbf{E})) \quad (7)$$

where $P(\ell(t)=1|F)$ reflects the classification predicted by formula F (assuming t is not an outlier).

Let us summarize the ingredients of the Rational Rules model. At the most general level, we have assumed (i) that concepts are expressed in a grammatically defined concept language; (ii) that the generative process of this grammar leads to a prior over formulae in the language; and, (iii) that there is a likelihood which captures evaluation of concepts and allows that some examples may be outliers. These ingredients are combined by the standard techniques of Bayesian induction, and, all together, give a general framework for "grammar-based induction" (N. D. Goodman, Tenenbaum, Griffiths, & Feldman, In Press). The particular grammar we have used, the DNF grammar, describes concepts as classification rules: disjunctive normal form "definitions" for a concept label. This grammar incorporates some structural prior knowledge: labels are very special features (Love, 2002), which apply to an object exactly when the definition is satisfied, and conjunctions of feature values are useful "entries" in the definition. The final ingredient of the model is one free parameter, the outlier probability, which describes the prior probability that an example is an outlier which should be ignored.

Each of the three intuitions about concepts and concept learning that initially seeded our discussion is captured in the posterior belief function of the Rational Rules model, Eq. 6. First, each concept described in the concept language is a classification rule that determines when its label can be applied. Since the posterior is seldom degenerate, there will usually remain uncertainty about *which* rule corresponds to a given concept label. Second, this posterior deals gracefully with uncertainty by including a strong inductive bias, and the possibility of treating some examples as outliers; these two factors imply a trade-off between explanatory completeness and conceptual parsimony. Finally, concepts which can be expressed in the concept language are built by combining primitive concepts (the feature predicates); and these combinations are mirrored in the meaning of the concepts, through the likelihood.

Bridging to Empirical Studies

In this section we present a few additional assumptions and tools that we will need in order to bridge from the rational model to experimental results.

Individuals and choice rules. The posterior and generalization probabilities, Eqs. 6 and 7, capture the inferences of an ideal learner. However, to make experimental predictions, we will require an auxiliary hypothesis—a *choice rule*—describing the judgments made by individual learners on classification questions. One possibility, *probability matching*, is to assume that individuals maintain (in some

sense) the full posterior over formulae, and match the expected probability of labels when it comes time to make a choice. The expected portion of participants judging that a test object "t is an ℓ", would then be equal to the generalization probability, Eq. 7. The probability matching assumption is implicit in much of the literature on Bayesian learning (e.g. Tenenbaum & Griffiths, 2001), and also prevalent in the broader literature on decision making, via the Luce choice rule (Luce, 1959).

A second possibility, *hypothesis sampling*, is that each individual has one (or a few) hypotheses drawn from the posterior over formulae. (That is, by the end of learning each individual has acquired such a hypothesis—we may remain agnostic about the process by which this is achieved.) Each individual then gives the most likely response to any query, given their hypothesis. The expected probability of generalization responses, averaged over a large enough population, is again given by Eq. 7. Thus the prediction for the population average of responses is identical between the probability matching and hypothesis sampling assumptions.

We favor the hypothesis sampling assumption for three reasons. First, it seems intuitively very plausible that individuals maintain only one, or a few, hypotheses rather than an entire distribution. This allows for the possibility that the process of learning resembles hypothesis testing, while sequentially sampling from the Bayesian posterior (as in Sanborn, Griffiths, & Navarro, 2006). Second, maintaining a small number of hypotheses is an efficient use of bounded computational resources. Indeed, memory constraints were a primary motivating consideration for the RULEX model. Finally, there is some experimental evidence that supports the idea that individuals in standard laboratory tasks learn a small number of rules. For instance, the classification behavior exhibited by individuals in the experiments of Nosofsky et al. (1994); Nosofsky and Palmeri (1998); Johansen and Palmeri (2002) are suggestive of hypothesis sampling (and the idea of idiosyncratic rule formation discussed by these authors is similar to the hypothesis sampling assumption). Lamberts (2000), has shown that over a large number of transfer blocks individual participants respond with far lower variance than expected based on the group average. This is consistent with the idea that individuals learn a nearly deterministic representation of concepts, such as a small set of alternative rules.

Where it is useful to be explicit, we phrase our discussion below in terms of the hypothesis sampling assumption. (The wider implications of this assumption will be considered in the General Discussion.) We further assume that each participant gives slightly noisy responses: there is a probability η of giving the subjectively wrong answer. This captures simple decision noise and a host of pragmatic effects (motor noise, inattention, boredom, etc.), and is a common assumption in models of concept learning (cf. Nosofsky et al., 1994; Smith & Minda, 1998). The effect of this response noise on the predicted (aggregate) response probability, Eq. 7, is a simple linear transformation—this parameter is thus absorbed into the correlation when R^2 values are used to compare with human data. Below we explicitly fit η only when discussing the

Table 2
The best formulae according to the RR_{DNF} model (with posterior probabilities) for the examples of Table 1, at two different values of the parameter b. (Only the "definition" part of the formulae are shown.) For small b the model favors single-feature rules, weighted by their predictiveness; for larger b the model begins to favor complex formulae.

$b=1$		$b=6$	
0.234	$f_3(x)=0$	0.059	$((f_1(x)=0) \wedge (f_3(x)=0)) \vee ((f_2(x)=0) \wedge (f_4(x)=0))$
0.234	$f_1(x)=0$	0.027	$f_1(x)=0$
0.086	$f_4(x)=0$	0.027	$f_3(x)=0$
0.031	$f_2(x)=0$	0.005	$((f_1(x)=0) \wedge (f_4(x)=0)) \vee (f_3(x)=0)$

performance of individual participants.

Parameter fitting. The Rational Rules model described above has two free parameters: the outlier parameter b, and the response noise parameter η. In the results reported below we have simulated the model for outlier parameter $b \in \{1, \dots, 8\}$ (these values were chosen for convenience only). When only a single fit is reported it is the best from among these eight parameter values. It is likely that this rough optimization does not provide the best possible fits of the RR_{DNF} model to human data, but it is sufficient to demonstrate the ability of the model predict human responses. Where η is explicitly fit, it is fit by grid search.

The model predictions were approximated by Monte Carlo simulation (30,000 samples for each run, with five runs averaged for most reported results). Details of the Monte Carlo algorithm and simulation procedure can be found in Appendix C.

Blocked-learning experiments. In many of the experiments considered below participants were trained on the category using a blocked-learning paradigm: each example in the training set was presented once per block, and blocks were presented until the training set could be classified accurately (relative to a predetermined threshold). It is often the case that different effects occur as training proceeds, and these effects can be tricky to capture in a rational model. One advantage of the Rational Rules model is that the effect of repeated examples on the posterior is related to the value of the outlier parameter b. Indeed, it is apparent from Eq. 6 that the Rational Rules model with outlier parameter b presented with N identical blocks of examples is equivalent to the model presented with only one block, but with parameter $b' = b \cdot N$. This makes intuitive sense: the more often an example is seen, the less likely it is to be an outlier. Thus we may roughly model the course of human learning by varying the b parameter—effectively assuming a constant outlier probability while increasing the number of trials.

Two-category experiments. In several of the experiments considered below participants were required to distinguish between two categories, A and B, which were mutually exclusive. (As opposed to distinguishing between a category A and it's complement "not A".) For simplicity in fitting the model we assume that the population is an even mixture of people who take A to be the main category, and B the con-

trast category, with vice versa. Since these experiments have similar numbers of A and B examples, this is probably a reasonable initial assumption. (A more realistic treatment would follow by treating these two categories as a system of interrelated concepts (Goldstone, 1996)—we leave this extension to future work.)

Descriptive measures of the posterior. We will shortly try to understand the behavior of the model in various concept learning experiments. Since the posterior (Eq. 6) describes what has been learned by the model, it will be useful to have a few descriptive measures of the posterior. In particular, we would like to know the relative importance of formulae with various properties.

The Boolean complexity of a formula (Feldman, 2000), written cplx(F), is the number of feature predicates in the formula: a good overall measure of syntactic complexity. For example, $(f_1(x)=1)$ has complexity 1, while $(f_2(x)=0) \wedge (f_1(x)=1)$ has complexity 2. The *posterior weight* of formulae with complexity C is the total probability under the posterior of such formulae:

$$\sum_{F \text{ st. cplx}(F)=C} P(F|\mathbf{E}, \ell(\mathbf{E})). \qquad (8)$$

Define the weight of feature i in formula F to be $\frac{\text{count}(f_i \in F)}{\text{cplx}(F)}$, that is, the number of times this feature is used divided by the complexity of the formula. The *posterior feature weight* is the posterior expectation of this weight:

$$\sum_{F} \frac{\text{count}(f_i \in F)}{\text{cplx}(F)} P(F|\mathbf{E}, \ell(\mathbf{E})). \qquad (9)$$

The posterior feature weights are a measure of the relative importance of the features, as estimated by the model. Indeed, it can be shown that Eq. 9 is related in a simple (monotonic) way to the posterior expectation of the production probability for production $P \to F_i$ (see Appendix D).

Comparison with Human Category Learning

In the preceding sections we have presented a Bayesian analysis of concept learning assuming that concepts are represented in a concept language of rules. In this section we begin to explore the extent to which this analysis captures human learning by comparing the RR_{DNF} model to human data

from several influential experiments. We consider four experiments from the Boolean concept learning literature that have often been used as tests of modeling efforts (e.g. Nosofsky et al., 1994), and one concept based on non-Boolean features which has been used in a similar way (e.g. Nosofsky & Palmeri, 1998). We close this section by considering the within-participants pattern of generalization judgments – a more refined test of the model.

We use data from human experiments in which physical features were counter-balanced against logical features. So, for instance, in an experiment with the two physical features Length and Angle, half of participants would see Angle playing the role of logical feature f_1 and for the other half Angle would be f_2. This counter-balancing allows us to focus on foundational questions about concept formation, without worrying over the relative saliency of the physical properties used to represent features. Except where otherwise noted physical features in these experiments were along psychologically separable dimensions.

Prototype Enhancement and Typicality Effects

The second experiment of Medin and Schaffer (1978), among the first studies of ill-defined categories, used the "5-4" category structure shown in Table 3 (we consider the human data from the Nosofsky et al. (1994) replication of this experiment, which counter-balanced physical feature assignments). This experiment is a common first test of the ability of a model to predict human generalizations on novel stimuli.

The overall fit of the Rational Rules model (Fig. 3) is good: R^2=0.98. Other models of concept learning are also able to fit this data quite well: for instance R^2=0.98 for RULEX, and R^2=0.96 for the context model (Medin & Schaffer, 1978). However, the Rational Rules model has only a single parameter (the outlier parameter), while each of these models has four or more free parameters; indeed, the full RULEX model has nine free parameters.

The best (highest posterior probability) formulae according to the Rational Rules model are shown in Table 2. For small values of b (those that are more permissive of outliers), the best formulae are single dimension rules. Note that the posterior probabilities of these formulae reflect their predictiveness on the training examples, and in particular that formula $f_4(x)$=0, though not the best, has significantly greater posterior probability than formula $f_2(x)$=0. In Fig. 4 we have plotted the posterior complexity weights and the posterior feature weights of the Rational Rules model for b=1. We see that this pattern is maintained when considering the entire posterior: most of the weight is on simple formulae along features 1 and 3, followed by feature 4, then feature 2.

The object T3=0000 is the prototype of category A, in the sense that most of the examples of category A are similar to this object (differ in only one feature) while most of the examples of category B are dissimilar. Though it never occurs in the training set, the importance of this prototype is reflected in the human transfer judgments (Table 3 and Fig. 3): T3 is, by far, the most likely transfer object to be classified as category A. The Rational Rules model predicts this "prototype enhancement effect" (Posner & Keele, 1968). This prediction results because the highest posterior probability formulae (Table 2) all agree on the categorization of T3, so combine (together with lower probability formulae) to enhance the probability that T3 is in category A.

The degree of typicality, or recognition rate for training examples, is often taken as a useful proxy for category centrality (Mervis & Rosch, 1981) because it correlates with many of the same experimental measures (such as reaction time). In Table 3 and Fig. 3, we see greater typicality for the prototype of category B, the object B4=1111, than for other training examples: though presented equally often it is classed into category B far more often. The Rational Rules model also predicts this typicality effect, in a manner similar to prototype enhancement: most high probability formulae agree on the classification of B4, while fewer agree on the classifications of the other training examples. However, note that the typicality gradient of training examples is not entirely determined by similarity to the prototypes. For instance, training example A2 and A1 are equally typical of category A and more typical than examples A4 and A5 (according to both humans and the Rational Rules model), however A1, A3, and A4 are each *more* similar to the prototype, T3, than is A2. This finding has often been taken as evidence of exemplar memory (Medin & Schaffer, 1978). However, for the Rational Rules model the enhancement of A2 reflects the greater reliance on rules involving features 1 and 3 (which support the classification of A2, but not of A4 or A5) and the relative unimportance of feature 2 (the only feature on which A1 and A2 differ).

In these model results graded typicality effects arise out of deterministic rules by maintaining uncertainty over the rule that defines the concept. Following the hypothesis sampling assumption outlined above, we might expect a single individual to learn a small set of rules, sampled from the posterior. Objects which satisfy all these rules would be considered more typical of the concept (after all, they are part of the concept under any of the competing definitions) than objects which satisfy only some. (This is similar to the proposal of Lakoff (1987), in which "idealized cognitive models" of a concept are composed of several "entries"; objects which satisfy many entries are considered better examples of the concept than those which satisfy few.) In this way similar graded typicality judgments would be expected from individuals, as from group averages.

Prototype and typicality effects led to great interest among the psychological community in prototype-based models of concept learning (e.g. Reed, 1972). Many such models represent prototypes as points in a similarity space. Because the curve equidistant to two points in a metric space is a line, these prototype models predict that linearly separable categories—those which admit a linear discriminant boundary—will be easier to learn than those that are not linearly separable. Medin and Schwanenflugel (1981) tested this prediction in four experiments, finding that linearly separable concepts could be harder for human participants to learn than closely matched concepts which were not linearly separable. As an example, consider Medin and Schwa-
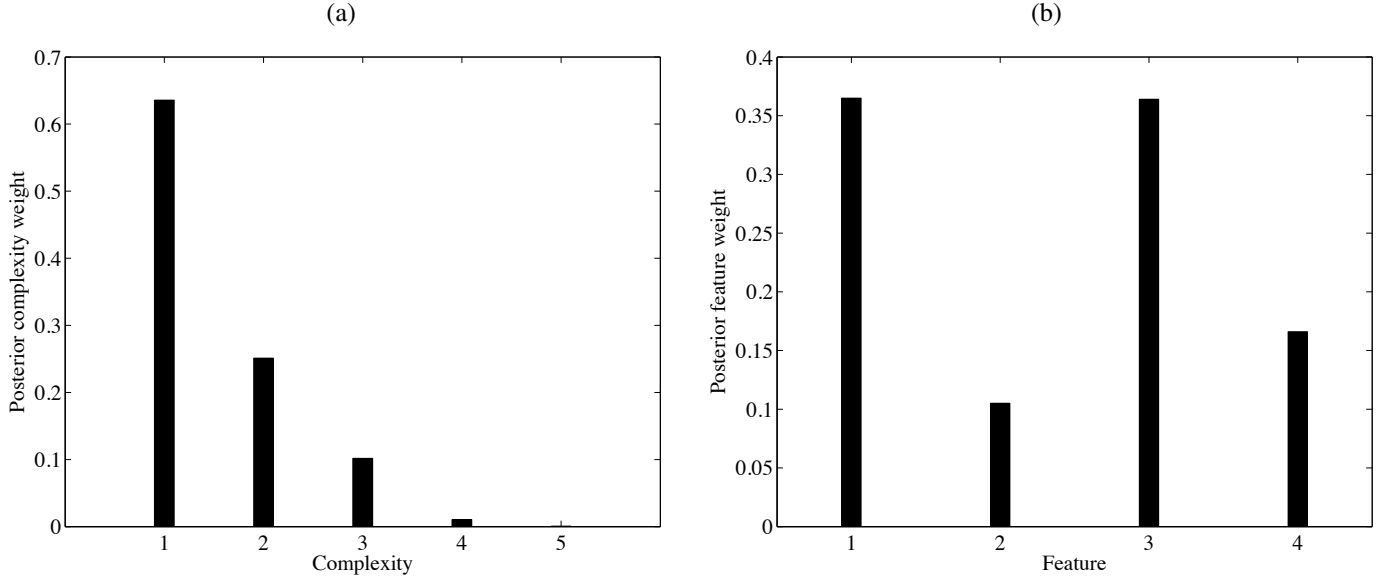
(a)

(b)

Figure 4. (a) Posterior complexity distribution of the RR$_{DNF}$ model (b=1) for the category structure of Medin & Schaffer (1978), see Table 3. (b) Posterior feature weights for this category structure. Together these weight distributions indicate that the RR$_{DNF}$ model focuses on simple rules along features 1 and 3.

Table 3
The category structure of Medin & Schaffer (1978), with the human data of Nosofsky et al. (1994), and the predictions of the Rational Rules model (b=1).

Object	Feature Values	Human	RR$_{DNF}$
A1	0001	0.77	0.82
A2	0101	0.78	0.81
A3	0100	0.83	0.92
A4	0010	0.64	0.61
A5	1000	0.61	0.61
B1	0011	0.39	0.47
B2	1001	0.41	0.47
B3	1110	0.21	0.21
B4	1111	0.15	0.07
T1	0110	0.56	0.57
T2	0111	0.41	0.44
T3	0000	0.82	0.95
T4	1101	0.40	0.44
T5	1010	0.32	0.28
T6	1100	0.53	0.57
T7	1011	0.20	0.13
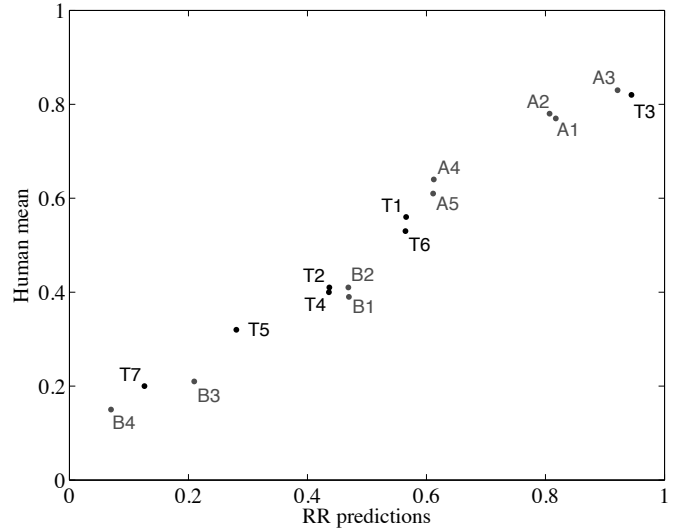
Figure 3. Comparison of human judgments with RR$_{DNF}$ model predictions: mean probability of category A judgments after training on the category structure of Medin & Schaffer (1978), see Table 3, for human and RR$_{DNF}$ model (b=1). The fit between model and human data is R^2=0.98.

nenflugel (1981), Experiment 3, in which participants were trained on the two concepts shown in Table 4, and tested on classification accuracy for the training set. Concept LS is linearly separable, Concept NLS is not, and the two concepts have matched single dimension strategies (that is, any single feature predicts category membership two thirds of the time, in each concept). Throughout the experiment learners

make fewer errors on Concept NLS (Fig. 5a). In Fig. 5b we see that the Rational Rules model provides good qualitative agreement with the human data, predicting more errors on the linearly separable concept (and note that no parameters were fit in these model results). The RULEX model also predicts the relative difficulty of Concept LS over Concept NLS (Nosofsky et al., 1994), which suggests that this is a rather

Table 4

The two concepts from Medin & Schwanenflugel (1981). Concept LS is linearly separable, Concept NLS is not.

Concept LS	
Category A	Category B
1000	0111
0001	1000
0110	1001

Concept NLS	
Category A	Category B
0011	1111
1100	1010
0000	0101

Figure 6. Posterior complexity distribution of the RR_{DNF} model (b=3) on the two category structures from Medin & Schwanenflugel (1981), see Table 4. The model shows greater dependence on simple rules for Concept LS than Concept NLS.

general prediction of rule-based models.

To understand this result, note that, though the two concepts support equally informative complexity 1 rules (that is single-feature strategies), Concept NLS supports more informative rules of complexity 2, 3, and 4 than does Concept LS. For example the complexity 4 formula ($f_1(x)$=0 ∧ $f_2(x)$=0) ∨ ($f_3(x)$=0 ∧ $f_4(x)$=0) discriminates perfectly for Concept NLS, while there is no complexity 4 formula which does so for Concept LS. The RR_{DNF} model relies more heavily on these rules of complexity 2, 3, and 4 for Concept NLS than for Concept LS, see the plots of posterior complexity in Fig. 6, which results in a difference in accuracy. The model does not, however, simply use the most informative rules (after all there are always perfectly predictive rules of very high complexity), but balances predictive accuracy against simplicity—it places weight on highly informative and moderately complex rules for Concept NLS, but, finding no such rules for Concept LS, places the majority of the weight on very simple rules.

Selective Attention Effects

It is important to be able to ignore uninformative features in a world, such as ours, in which every object has many features, any of which may be useful for classification. This motivates the long standing interest in *selective attention* in human concept learning (Kruschke, 1992): the tendency to consider as few features as possible to achieve acceptable classification accuracy. We have seen a simple case of this already predicted by the Rational Rules model: single feature concepts were preferred to more complex concepts for the 5-4 category structure (Fig. 4(a)). Indeed, each of the descriptive measures described above (the complexity and feature weights) is a measure of selective attention exhibited by the model: the posterior complexity weights describe the extent to which the model favors simpler formulae (which will have fewer features), while the posterior feature weights directly describe the informativeness of each feature, as estimated by the model. It has been noted before (Navarro, 2006) that selective atte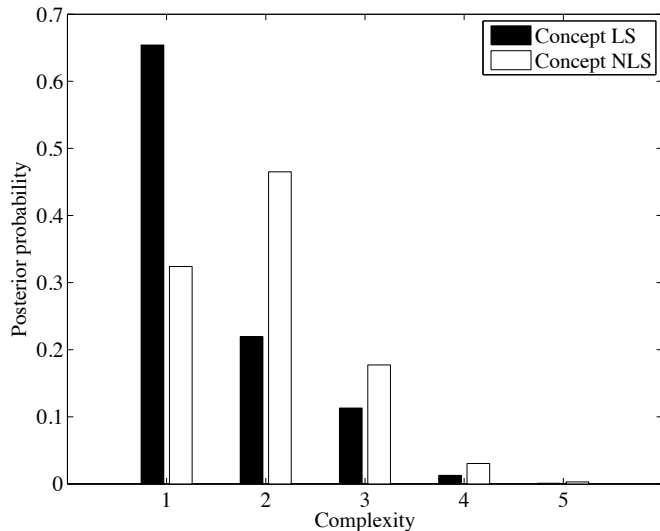ntion effects emerge naturally from the Bayesian framework. In our setting selective attention can be understood as the effect of updating the uncertainty over production probabilities as evidence accumulates. Indeed, as the prior over τ—initially uniform—is updated, it will often concentrate, becoming tightly peaked on a subset of productions. For instance, if only the first of three features is informative, the posterior distribution on production $P \rightarrow F_1$ will become larger, while the posteriors on $P \rightarrow F_2$ and $P \rightarrow F_3$ will be small (and these values will be reflected in the posterior feature weights, see Appendix D). The changing importance of the productions τ have been marginalized away in the summary prior Eq. 4, but the effects will still be felt in model predictions. As a result the inferences of the Rational Rules model will depend most sensitively on the informative features—this is the manner in which Bayesian models implement selective attention.

Shepard et al. (1961), in one of the first studies to demonstrate selective attention effects, compared difficulty in learning the six concepts in Table 5 (these are the six concepts with three Boolean features, four positive and four negative examples). These concepts differ in the number of dimensions which must be attended to, in the complexity of their simplest perfect rule, and in the number of imperfect, but useful, simple rules. To learn Concept I it is only necessary to consider the first feature, that is, the rule ($f_1(x)$=0) perfectly predicts category membership, and the remaining features are uninformative. For Concept II the first two features are informative; for example the complexity 4 formula

$$((f_1(x)=1) \land (f_2(x)=1)) \lor ((f_1(x)=0) \land (f_2(x)=0))$$

is the simplest perfect rule for this concept. In contrast, all three features are informative for Concepts III, IV, V, and VI.
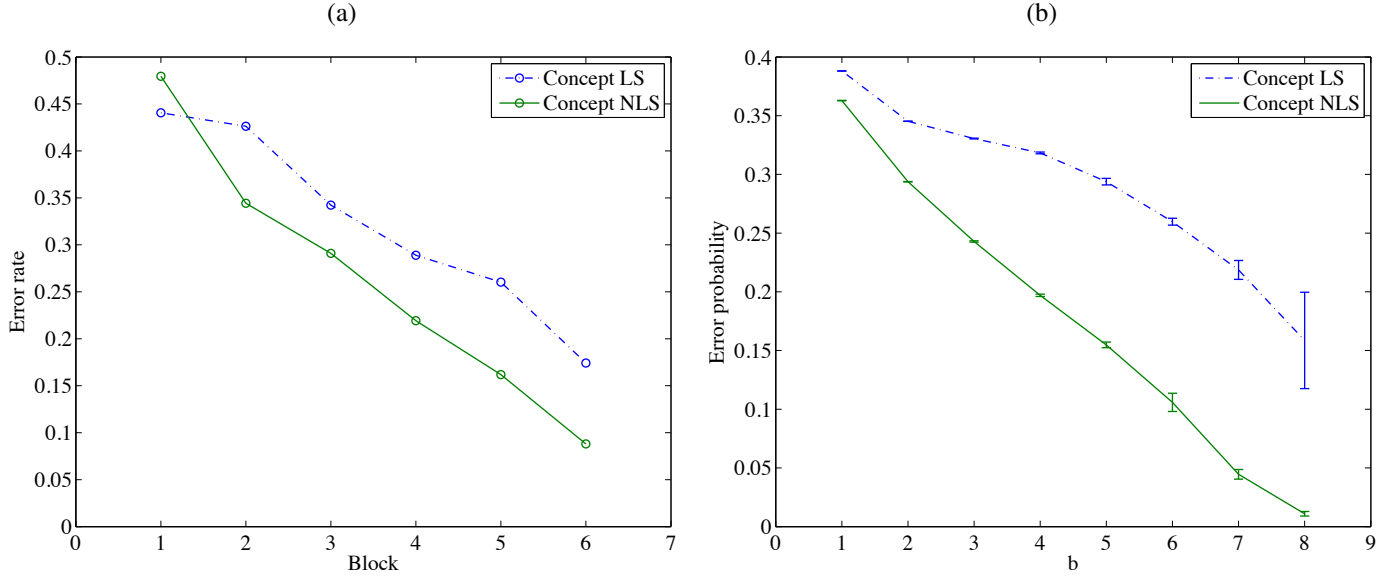
Figure 5. (a) The human data from Medin & Schwanenflugel (1981) for the category structures in Table 4, showing that linearly separable Concept LS was more difficult to learn than Concept NLS, which is not linearly separable. (b) Predictions of the RR_{DNF} model: the probability of an incorrect response vs. the outlier parameter b.

Concept III admits the relatively simple formula

$$((f_1(x){=}0) \wedge (f_3(x){=}0)) \vee ((f_2(x){=}0) \wedge (f_3(x){=}1)),$$

while Concepts IV, V, and VI don't admit any perfect rules of low complexity. However, IV, and V both admit imperfect, but useful, rules of low complexity, while VI has no useful simple rules at all.

A well replicated finding concerning human errors (Shepard et al., 1961) is that these concepts vary reliably in difficulty, reflecting the above complexity and informativeness considerations: I<II<III=IV=V<VI (ordered from least to most difficulty, where "=" indicates no reliable difference in difficulty). The RR_{DNF} model predicts these qualitative findings: error rates (via posterior probability, when $b{=}3$) of 0%, 17%, 24%, 24%, 25%, 48% for concepts I, II, III, IV, V, and VI, respectively. This ordering is predicted for a fairly wide range of parameter values, though an inversion is predicted at $b{=}1$: concept II is then more difficult than concepts III, IV, and V. It is intriguing that this inversion has been experimentally observed in humans when the stimulus dimensions are non-separable (Nosofsky & Palmeri, 1996), but further work will be needed to determine whether this is an illuminating or accidental prediction of the model. (This inversion, which is also seen in rhesus monkeys (Smith, Minda, & Washburn, 2004), is predicted by the ALCOVE model when attention learning is disabled.)

However, people are not bound to attend to the smallest set of informative features—indeed, selective attention is particularly interesting in light of the implied tradeoff between accuracy and number of features attended. Medin, Altom, Edelson, and Freko (1982) demonstrated this balance by studying the category structure shown in Table 6. This structure affords two strategies: each of the first two features are

Table 5
The six concepts with three features, four positive and four negative examples, studied first in Shepard et al. (1961).

I	II	III	IV	V	VI
+ 0 0 0	+ 0 0 0	+ 0 0 0	+ 0 0 0	+ 0 0 0	+ 0 0 0
+ 0 0 1	+ 0 0 1	+ 0 0 1	+ 0 0 1	+ 0 0 1	- 0 0 1
+ 0 1 0	- 0 1 0	+ 0 1 0	+ 0 1 0	+ 0 1 0	- 0 1 0
+ 0 1 1	- 0 1 1	- 0 1 1	- 0 1 1	- 0 1 1	+ 0 1 1
- 1 0 0	- 1 0 0	- 1 0 0	+ 1 0 0	- 1 0 0	- 1 0 0
- 1 0 1	- 1 0 1	+ 1 0 1	- 1 0 1	- 1 0 1	+ 1 0 1
- 1 1 0	+ 1 1 0	- 1 1 0	- 1 1 0	- 1 1 0	+ 1 1 0
- 1 1 1	+ 1 1 1	- 1 1 1	- 1 1 1	+ 1 1 1	- 1 1 1

individually diagnostic of category membership, but not perfectly so, while the correlation between the third and fourth features is perfectly diagnostic. It was found that human learners relied on the perfectly diagnostic, but more complicated, correlated features. McKinley and Nosofsky (1993) replicated this result, studying both early and late learning by eliciting transfer judgments after both initial and final training blocks. They found that human participants relied primarily on the individually diagnostic dimensions in the initial stage of learning, and confirmed human reliance on the correlated features later in learning. The RR_{DNF} model explains most of the variance in human judgments in the final stage of learning, $R^2{=}0.95$ when $b{=}7$; see Fig. 7. Correlation with human judgments after one training block is also respectable: $R^2{=}0.69$ when $b{=}1$. By comparison RULEX has $R^2{=}0.99$ for final, and $R^2{=}0.67$ for initial learning. We have

plotted the posterior complexity of the RR_{DNF} model against b in Fig. 8, and the posterior feature weights in Fig. 9. When b is small the Rational Rules model relies on simple rules, but gradually switches as b increases to rely on more complex, but more accurate, rules.
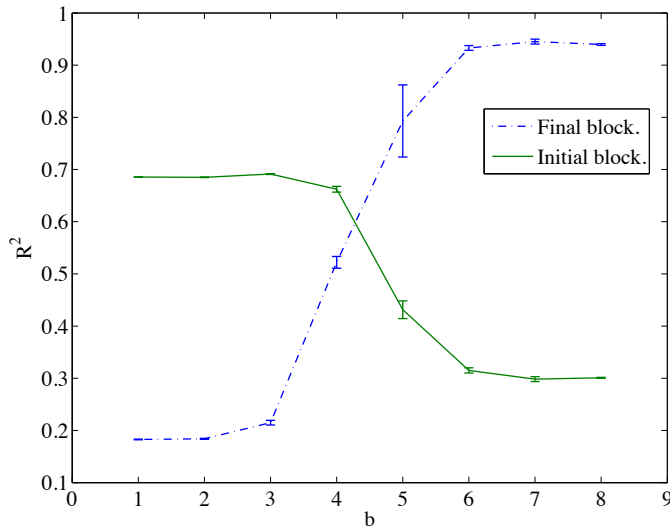

Figure 7. Fit of the RR_{DNF} model to human data (R^2), for data from initial and final learning blocks of McKinley & Nosofsky (1993), see Table 6. The fits are shown for eight values of the outlier parameter b. (Error bars represent standard error over five independent simulation runs.)
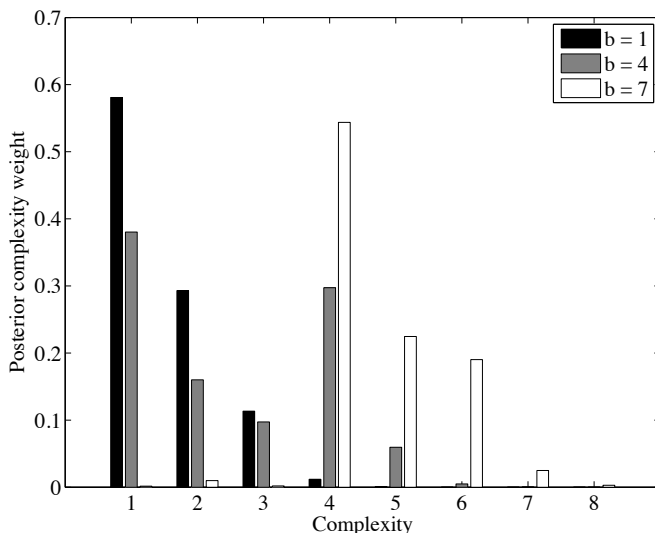

Figure 8. Posterior complexity distribution of the RR_{DNF} model on the category structure of Medin et al. (1982), see Table 6, for three values of the outlier parameter.

Concepts based on non-Boolean Features

We have focussed thus far on concepts with Boolean feature values, but the modular nature of the concept gram-
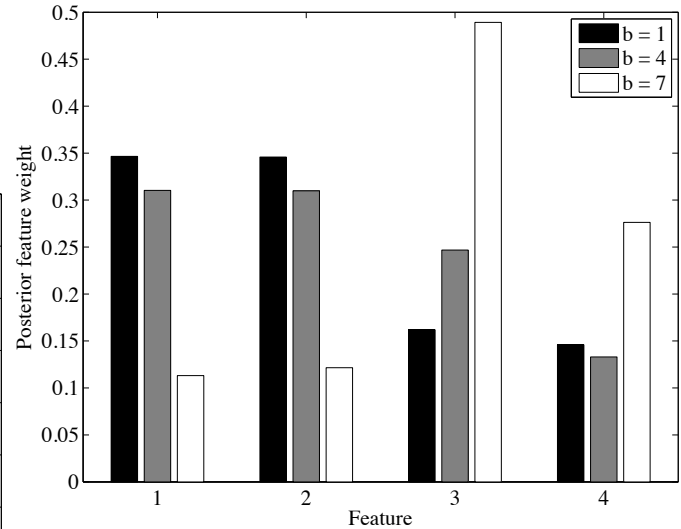
Figure 9. Posterior feature weights of the RR_{DNF} model on the category structure of Medin et al. (1982), see Table 6, for three values of the outlier parameter.

mar makes it quite easy to extend the model to other concept learning settings. Indeed, when the features take values on a continuous dimension we may, as described above, replace the simple boolean feature predicates with "decision boundary" predicates, e.g. $f_1(x) < 3$. This is quite similar to the strategy taken in Nosofsky and Palmeri (1998) to extend RULEX to continuous feature values. Indeed, the Nosofsky and Palmeri (1998) version of RULEX is in some ways most similar to the Rational Rules model. However the complications of the continuous RULEX model result in an awkward modeling process: a set of likely rules is chosen by hand, then the model is fit with several free parameters for each rule. In contrast, the Rational Rules model is easily extended, and acquires no additional free parameters or ad-hoc fitting steps.

As an initial test we have compared the RR_{DNF} model, using decision boundary predicates, with human data for the concept with two continuous features that first appeared in Nosofsky, Clark, and Shin (1989), using the human data from the Nosofsky and Palmeri (1998) replication. (The two experiments of this replication were identical except they used the two different assignments of logical dimensions to physical dimensions—we have averaged the results from these two experiments to counterbalance the data). The result suggests that the RR_{DNF} model captures a significant amount of human learning also for concepts based on continuous feature dimensions: $R^2 = 0.82$ (for $b=3$).

All of the results discussed so far depend on the choice of primitive feature predicates, but this is especially true of predictions for continuous-dimension stimuli. Indeed, there seems to be no *a priori* reason to choose two continuous features f_1 and f_2 rather than alternative "rotated" features $f_1' = f_1 + f_2$ and $f_2' = f_1 - f_2$. Yet the predictions of the Rational Rules model will differ significantly depending on the

Table 6
The category structure of Medin et al. (1982), with initial and final block mean human responses of McKinley & Nosofsky (1993), and the predictions of the Rational Rules model at b=1 and b=7.

Object	Feature Values	Human, initial block	Human, final block	RR$_{DNF}$, b=1	RR$_{DNF}$, b=7
A1	1111	0.64	0.96	0.84	1
A2	0111	0.64	0.93	0.54	1
A3	1100	0.66	1	0.84	1
A4	1000	0.55	0.96	0.54	0.99
B1	1010	0.57	0.02	0.46	0
B2	0010	0.43	0	0.16	0
B3	0101	0.46	0.05	0.46	0.01
B4	0001	0.34	0	0.16	0
T1	0000	0.46	0.66	0.2	0.56
T2	0011	0.41	0.64	0.2	0.55
T3	0100	0.52	0.64	0.5	0.57
T4	1011	0.5	0.66	0.5	0.56
T5	1110	0.73	0.36	0.8	0.45
T6	1101	0.59	0.36	0.8	0.44
T7	0110	0.39	0.27	0.5	0.44
T8	1001	0.46	0.3	0.5	0.43

set of features: simple concepts in terms of decision bounds along f_1 and f_2 may be quite complicated in terms of f_1' and f_2'. (This parallels the "grue/green" thought experiment often used to illustrate the "new riddle of induction" (N. Goodman, 1955).) The psychological reality and importance of basis features has been shown a number of ways. For instance, the "Filtration" concept of Kruschke (1993) may be simply describe as a decision bound along one dimensions, while the "Condensation" concept cannot—though it is identical to the Filtration concept but for a 45 degree rotation. Both human learners and the Rational Rules model find the Filtration concept considerably easier to learn than the Condensation concept.

It is likely that Rational Rules, viewed as a computational-level model, can guide the resolution of some of the ad-hoc assumptions in the existing process-level accounts of rule-based learning for continuous features. Conversely, the modeling assumptions used here to extend Rational Rules to continuous features can certainly be refined by incorporating insights from the (extensive) literature on continuous categories. In particular, empirical evidence (e.g. Maddox & Ashby, 1993) suggests that feature predicates capturing general linear or quadratic decision boundaries may be appropriate in some situations.

Individual Generalization Patterns

Nosofsky et al. (1994) investigated the pattern of generalizations made by individual participants, that is, they reported the proportion of participants giving each set of answers to the generalization questions. One may wonder whether it is necessary to consider these generalization patterns in addition to group averages for each question. As noted in Nosofsky and Palmeri (1998), even the best binomial model does very poorly at predicting individual gen-

eralization patterns (R^2=0.24 in the case of Nosofsky et al. (1994), Exp. 1), though, by construction, it perfectly predicts the group average for each generalization question. Therefore the pattern of generalizations provides an additional, more fine grained, probe for testing concept learning models.

To understand how the Rational Rules model can predict these generalization patterns, recall the hypothesis sampling assumption discussed above: each individual has a single hypothesis which is drawn from the posterior over formulae. The pattern of judgments made by each individual is then determined by this hypothesis, with additional response noise η. If we assume a single value of the parameters (b and η) for all participants, the best fit of the RR$_{DNF}$ model explains R^2=0.85 of the variance in human generalization for the 36 generalization patterns reported in (Nosofsky et al., 1994). The RULEX model also does well, R^2=0.86, but uses several more free parameters. As with RULEX, the qualitative match of the Rational Rules model to human judgments is good, see Fig. 10. Also as with RULEX, the generalization pattern ABABBAB is under-predicted by the model (cf. Johansen & Palmeri, 2002).

An Experiment

In the previous section we have discussed several important experiments from the concept learning literature, exploring human learning of concepts based on relatively few features. In each of these experiments participants were trained on many, or all, of the objects in the feature space, leaving only a few untrained objects available as transfer stimuli. (In fact, none of these experiments had fewer than half of possible objects as training examples.) In contrast, human learners must often cope with both large feature spaces, and relatively few labeled examples compared to the number of unseen objects. In a more natural setting, one with many

Figure 10. Individual generalization patterns: the probability of responding with the indicated categorizations for the seven transfer stimuli of Table 3. Human data from Nosofsky et al. (1994), Experiment 1. The model values are for parameters b=4, η=0.09. Agreement of model with human data is good: R^2=0.85, rmsd=0.016.

features and sparse training examples, one might expect different aspects of concept learning to come to the fore. For instance, when training examples are sparse, learning will be less constrained by available information and the inductive bias of the learning mechanism will play a relatively larger role. Further, when there are many features, the memory demands of remembering even a single exemplar become significant, so it is important to focus on informative rules based on a subset of the features. Given these considerations, it is important to test models of concept learning against human learning in settings with many features and sparse examples.

In addition, there is a danger of selecting the concepts to be tested in a way that biases the results. Historically, many concept learning experiments have used the same hand-picked concept structures, e.g. the Medin and Schaffer (1978) 5-4 concept, which has been used in dozens of studies. It is extremely plausible that some learning techniques work better on some types of concepts than others (see Briscoe & Feldman, 2006; Feldman, 2003, 2004), leaving doubt about whether performance on a small set of concepts is a reliable indicator of success more generally. This was one of the motivations for Shepard et al. (1961)'s famous concept set, which constitutes an exhaustive (and thus inherently unbiased) survey of concepts with three dimensions and four positive examples. When the number of features is large, it is impossible to be similarly exhaustive, but we can achieve a similar end by choosing our concepts *randomly*, so that we are at least guaranteed that our choices will be unbiased with respect to the performance of competing models—a level playing field. Thus in the experiment described below, the training set is a randomly selected subset of the complete set of objects.

The complexity of patterns formed by chance should vary with the number of examples: for example, with few examples there may be more "accidental" simple regularities. It

isn't feasible to vary the number of examples systematically over a wide range, but it is possible to do so for small numbers of examples. Hence, in the experiment that follows we use a large set of Boolean features (D=7), yielding 2^7=128 objects total, of which a small randomly drawn set of 3 to 6 are presented as positive examples, and two are presented as negative examples. (Some negative examples are necessary to give the participant a sense of the range of positive examples; for simplicity we always used two negative examples.) This leaves the vast majority of the space (at least 122 objects) as "transfer" objects. After brief training with the example objects, participants were asked to classify all 128 objects in random order. The goal is to apply the model to predict responses on the 128 generalization trials, as a function of the training set.

Method

Participants. Participants were 47 undergraduate students enrolled in a Psychology class, participating in the study in return for course credit. All were naive to the purposes of the study.

Materials and procedure. Objects were amoeba-like forms, each consisting of an outer boundary and one or more "nuclei" (smaller shapes in the interior). The amoebas varied along seven Boolean dimensions (body shape = rectangle or ellipse; boundary = solid or fuzzy; nucleus shape = triangle or circle; nucleus size = large or small; nucleus color = filled or unfilled; nucleus number = 1 or 2; fins present or absent). These features were chosen simply to be plainly perceptible and salient to participants.

Participants were told they were to play the role of a biologist studying a new species of "amoeba"-like organisms. They were instructed that they were to study a small number of known examples of each new species, and then attempt to

classify unknown examples as members or non-members of the species.

Each concept session began with a training screen, divided into top and bottom halves by a horizontal line. The top half of the screen contained the $P = 3, 4, 5$ or 6 positive example objects, drawn in a horizontal row in random order, and labeled "Examples"; the bottom half contained the two negative examples, again drawn in a horizontal row in random order, and labeled "NOT examples." The training screen remained up for a fixed period of time, equal to $5P$ seconds (i.e. $15 - 30$ seconds depending on the number of positives, yielding a constant average learning time per positive object).

After the training screen disappeared, participants were presented with a series of 128 individual classification trials in random order. On each classification trial, a single object was presented centrally, and participants were instructed to classify it as a member of the species or a nonmember. No feedback was given.

Each participant was run on several separate concepts, at different values of P, in random order. It was emphasized to participants that each new species (concept) was unrelated to the preceding ones. The entire sequence took about an hour per participant. In some cases we intentionally ran multiple participants on the same concepts, to facilitate comparisons between participants on identical inputs, but this was not pursued systematically. In total there were 140 participant/training set pairs, which we will refer to as individual *runs*.

Results and Discussion

We consider several ways of analyzing the experimental data, in order to address three main questions. First, to what extent can the probabilistic rule-based approach describe human concept learning and generalization behavior in this relatively unconstrained setting? Second, can we infer the specific rules that learners adopted on particular runs, and can we make any general conclusions about the kinds of rules they tend to use in this setting? Third, how well can we predict the graded generalization probabilities of individual learners for individual test objects? The RR_{DNF} model provides our basic tool for answering each of these questions.

To assess the first question we consider the overall ability of the RR_{DNF} model to predict participants' patterns of judgments for the 128 test stimuli within each run. We measure the fit of the model to human responses with the (natural) log-likelihood of the pattern of responses in an individual run according to the model:

$$\ln(P_{RR_{DNF}}(\text{response}_1, ..., \text{response}_{128}|\text{training examples})).$$

This is a measure of how likely the model considers the response pattern given the training set (or, roughly, the amount of information in the human data that cannot be accounted for by the model). It is always negative, and numbers with smaller magnitude indicate better fit. Without fitting any parameters, simply fixing $\eta{=}0.1$ and $b{=}1$, mean log-likelihood across runs is -54. That is, using the model the probability of correctly predicting responses to all 128 questions of a given

run is 34 orders of magnitude better than chance (chance is $\ln(\frac{1}{2^{128}}){=} -89$). Moreover this predictive success is quite broad: the response pattern of 85% of runs is predicted better than chance, and the mean predictive success for runs in each of the 43 unique training sets is above chance.
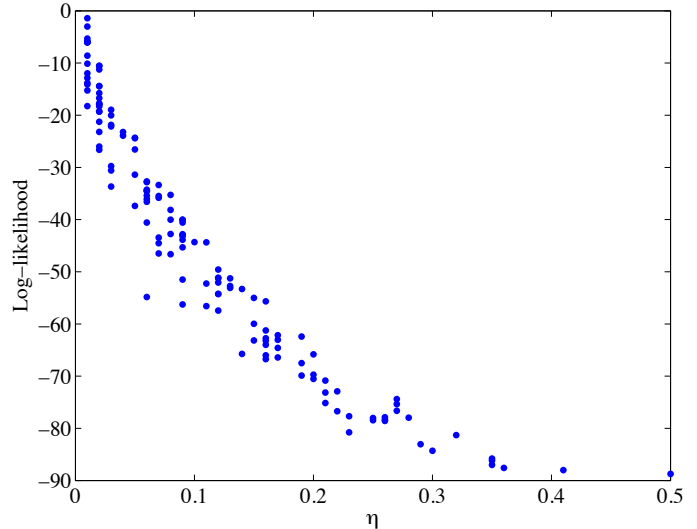

Figure 11. Distribution of (natural) log-likelihood scores vs. best-fit η parameter for each individual run. (Chance log-likelihood is -89, since there are 128 binary choices in each run.)

In light of the results on individual generalization patterns above, it makes sense to fit parameters η and b for each individual run, hoping to capture individual differences in learning in addition to the differences in what is learned. (Fitting parameters per run can quickly lead to over-fitting; however, since we have 128 responses per run in this experiment, it is not unreasonable to fit two parameters for each.) The mean best log-likelihood for each run is -44. This log-likelihood is significantly greater than that expected by a chance fit ($p{<}0.001$ by permutation test[7]). The distribution of log-likelihood scores against best-fit η values is shown in Fig. 11. Note first that the majority of η values are relatively small, indicating that behavior on most runs is best explained primarily in terms of the rational rule-based inferences from the examples participants saw rather than in terms of response noise. Second, the tight correlation between log-likelihood values and η values indicates that the amount of response noise (as measured by the inferred value of η) is a primary factor explaining the differences in fit between individual runs. If there was another factor that distinguished runs and explained differences in model fit, we would expect that factor to produce independent variance observable in Fig. 11.

The fit between model predictions and human data indicates that participants' performance may usefully be analyzed as probabilistic rule-based classification. Can we gain

---

[7] For this test we randomly permuted the data within individual runs. It is likely that the *p*-value is much smaller than 0.001: we estimated the permutation test by Monte Carlo simulation of 1500 random permutations of the data, and found no permutations with greater log-likelihood than the actual data.
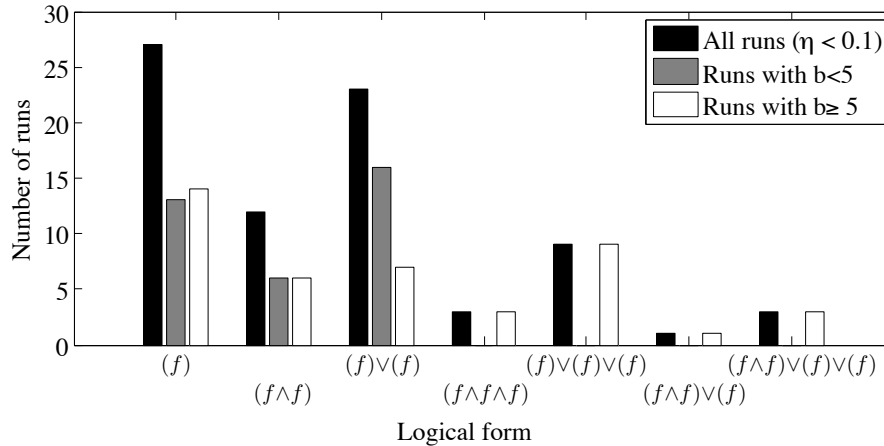
Figure 12. The distribution of logical forms of best formulae for individual runs with $\eta < 0.1$. The logical forms are schematic: f stands for any feature predicate (e.g., $f_1{=}1$). Runs with high b (less permissive outlier parameter) exhibit somewhat more complex classification behavior.

Table 7

Two sample individual runs, showing the training examples, best-fit parameters, and highest posterior formulae together with their agreement with the participant's responses (i.e. the portion of the 128 transfer questions on which participants' responses satisfy the formula). Formulae are listed in descending order of posterior probability given the training examples.

label	feature values	best-fit $b=1$, $\eta=0.03$	
		formula	agreement
-	0 1 1 0 1 0 1	($f_1{=}1$)	50%
-	0 1 1 0 1 1 0	($f_2{=}0$)	48%
+	1 0 0 0 1 0 1	($f_5{=}0$)	97%
+	1 0 1 1 0 1 0	($f_6{=}1$)	53%
+	1 1 1 0 0 1 1		

label	feature values	best-fit $b=8$, $\eta=0.07$	
		formula	agreement
+	0 0 0 0 0 0 1	($f_1{=}1$) $\vee$ ($f_5{=}0$)	76%
-	0 0 0 0 1 1 0	($f_1{=}1$) $\vee$ ($f_7{=}1$)	88%
+	0 0 0 1 1 1 0	($f_1{=}1$) $\vee$ ($f_3{=}0$)	84%
-	0 0 1 1 1 1 0	($f_1{=}1$) $\vee$ ($f_6{=}0$)	77%
+	1 0 1 1 1 1 0	($f_4{=}1$) $\vee$ ($f_7{=}1$)	74%
+	1 1 0 1 0 0 0		

further insight into *which* rules each participant used in each run? The rules that participants were using cannot be read off directly from their generalization behavior, but we can use the RR$_{\text{DNF}}$ model to infer which rules they were likely to be using. Consider two sample runs shown in Table 7, where we have given the training set, the highest posterior formulae according to the model under best-fitting parameters, and the agreement between participants' responses and the responses predicted by these high posterior formulae. As exemplified by the first run, some participants seem to pick a simple one-feature rule and use it consistently to classify the test stimuli.

The rule chosen may be somewhat arbitrary, not necessarily fitting the training examples perfectly or fitting them better than alternative rules of equivalent complexity. In the first run of Table 7, we see that a formula with high posterior probability captures the vast majority of the participant's responses but is not the formula with highest posterior probability; another simple rule exists ($f_1{=}1$) that can perfectly explain the training data. The inferred response noise (η) is low, because the participant's judgments are highly predictable.

Other participants, as illustrated by the second run shown in Table 7, appear to be using slightly more complex rules with multiple features, but these more complex rules may be used less consistently. Here the agreement between generalization behavior and any high-probability formula is less strong, which the model attributes to greater intrinsic noise (higher η). The agreement is also less tightly peaked, with multiple formulae having similarlly high agreement levels (largely because agreement between formulae forces correlation in agreements with participants' responses). Even in these cases, however, responess may be still usefully captured by a single high-agreement formula.

To generalize this analysis beyond single case studies, we define the "best" rule for each individual run to be the rule with maximum posterior probability *given the participant's responses*[8]. In Fig. 12 we have plotted the distribution of logical forms of the best formula for each of the individual runs. Overall, participants focused primarily on rules that use only one or two of the available seven features. This finding is similar in spirit to the classic finding of Medin, Wattenmaker, and Hampson (1987), that participants in unsupervised classification tasks tend to classify based on a sin-

---

[8] We may think of this as a Bayesian estimate of the rule used by an individual, in which the prior is given by the model's posterior over rules, and the likelihood is based on how well each rule predicts the noisy classification behavior of that individual. We obtain similar results if we define the "best" rule to be simply the formula that best agrees with a participant's responses, chosen from among those formulae with high posterior according to the model.

gle feature, even if the features of objects could also be explained by a family-resemblance category structure. Here, we see that participants in a "lightly supervised" concept-learning setting—in which there are relatively few examples, and brief familiarization—classify mostly based on simple one- or two-feature rules.

It is worth noting that runs which are best modeled with a permissive outlier parameter ($b<5$) rely mostly on one or two features, while runs with stricter outlier parameter values ($b \geq 5$) begin to exhibit more complex classification rules. Because the value of the outlier parameter is an estimate of how accurately participants have learned a concept consistent with the training set, this provides an indication of the course of learning (without manipulating it directly). The difference between small and large b runs is suggestive of a "representational shift" (Smith & Minda, 1998; Johansen & Palmeri, 2002; Briscoe & Feldman, 2006), from simple rules to more complex rules (or exemplar-type representations) as training is extended. In future work, it would be interesting to test whether this pattern of variation in representational complexity would result from more systematic variation of exposure time and number of examples relative to size of the training set.

Finally, we turn to the question of how well the RR_{DNF} model can predict the graded generalization probabilities of individual learners for individual test objects (as opposed to entire generalization patterns). This assessment is complicated by the fact that the model makes different predictions for each participant on each question, depending on the training set they saw—and because of the large number of training sets in this experiment only a few participants saw each. Yet if the model accurately predicts individual responses, across participants and questions, then of responses predicted with probability X% to be "yes", roughly X% should actually be answered "yes". Thus, instead of doing an object-by-object correlation as we did above for other data sets, we first sort the responses according to predicted probability of generalization—that is, we bin responses according to the prediction of the model, given the training set and the best-fit b and η. In Fig. 13(a) we have plotted the frequency of "yes" responses against the predicted generalization probability. We see that response frequencies are highly correlated with model generalization probability ($R^2=0.97$), indicating that the model is a good predictor of individual responses. Another way of interpreting Figs. 13(a) is that, across judgments, the model correctly predicts human responses in proportion to how confident it is in that prediction: when the model predicted a "yes" response with high probability most people did say yes, and when the model was unsure—assigning probability near 50%—people were evenly split between "yes" and "no" responses.

To find out whether this trend holds at the level of individual runs we did a similar analysis for each run: we binned the 128 responses in a run according to model posterior, and computed the frequency of "yes" responses in each bin. Fig. 13(b) shows the mean and standard error over runs, demonstrating that the model is a good predictor of individual responses on individual runs. (Since each run is an indi-

vidual participant this indicates good fit for the responses of individuals.)

One of the responsibilities of models of concept learning is to describe the inferences that people make across a broad range of natural situations. Thus it is important to verify that a model fits human data not only on simple, well-controlled concepts, but for more natural and generic circumstances. We have found good agreement between the RR_{DNF} model and human judgments when the number of features is large, the number of training examples small, and the specific training sets randomly generated. This is a necessary complement to the results discussed above which show that the Rational Rules model captures a number of well-known, specific learning effects.

General Discussion

We have suggested an approach for analyzing human concept learning: assume that concepts are represented in a concept language, propose a specific grammar and semantics for this language, then describe rational inference from examples to phrases of the language. Carrying out this scheme for concepts which identify kinds of things, by using a grammar for DNF formulae, we derived the Rational Rules (RR_{DNF}) model of concept learning. This model was shown to predict human judgments in several key category learning experiments, and to do so with only one, readily interpretable, parameter (and an additional, decision-noise, parameter for fits to individual subjects). Several phenomena characteristic of human concept learning—prototype enhancement, typicality gradients, and selective attention—were demonstrated. The model was used to predict categorization judgments based on continuous features, and the the pattern of generalization responses of individual learners. In a new experiment, we investigated the ability of the model to predict human behavior in generic natural environments: natural in the sense that there were many features and few training examples relative to the number of transfer stimuli, generic because these training sets were randomly chosen. Human generalization behavior was again well predicted by the model in this experiment.

Relation to Other Models of Concept Learning

For most fits of the Rational Rules model to human data we have provided a comparison with RULEX model fits. This comparison has shown that the RR_{DNF} model fits human data as well as one of the most successful existing models of concept learning, but it also shows how fits of the RR_{DNF} model generally parallel fits of RULEX, both better and worse. This reinforces the interpretation of the Rational Rules model as a computational-level analysis of the same species of rule-based inductive inferences that RULEX attempts to capture at the process level. Of course the RR_{DNF} model is not a rational analysis of RULEX per se, but rather of a class of models for which RULEX is one prominent example.

Our DNF representations are in some ways more similar to the cluster-based representations of the SUSTAIN
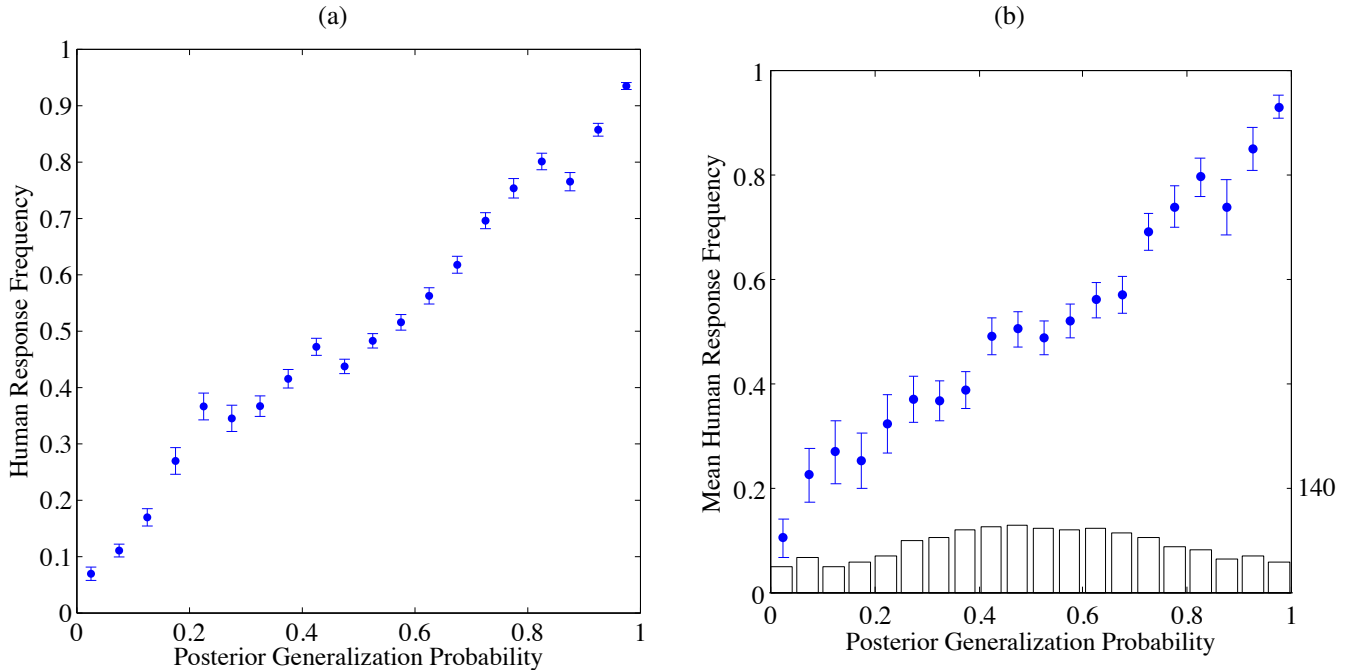
Figure 13. (a) Human categorization response frequency (proportion of "yes" judgments) against model posterior generalization probability, $R^2 = 0.97$; error bars represent standard error of frequency (assuming binomial distribution). (Frequencies are computed by first binning responses according to model prediction.) (b) The mean of response frequencies (binned according to model prediction) computed for each run separately; error bars represents standard error of the mean over runs; bars below each data point indicate number of runs contributing to that bin (scale on right).

model (Love et al., 2004) than they are to representations in RULEX; conjunctive blocks of RR_{DNF} formulae are analogous to the clusters that SUSTAIN learns, with features that are ommitted from a conjunctive clause analogous to features that receive low attentional weights in SUSTAIN. All three of these models—RULEX, SUSTAIN, and RR_{DNF}—navigate similar issues of representational flexibility, trade-offs between conceptual complexity and ease of learning, and generalization under uncertainty. The main advantages that Rational Rules offers over the other two models come from its focus on the computational-theory level of analysis and the modeling power that we gain at that level: the ability to work with a minimal number of free parameters and still achieve strong quantitative data fits, the ability to separate out the effects of representational commitments and inductive logic from the search and memory processes that implement inductive computations, and the ability to seamlessly extend the model to work with different kinds of predicate-based representations, such as those appropriate for learning concepts in continuous spaces, concepts defined by causal implications (see N. D. Goodman et al., In Press), or concepts defined by relational predicates (see below).

A central theme of our work is the complementary nature of rule-based representations and statistical inference, and the importance of integrating these two capacities in a model of human concept learning. Other authors have written about the need for both rule-based and statistical abilities—or of-

ten rules and similarity—in concept learning, and cognition more generally (Sloman, 1996; Pinker, 1997; Pothos, 2005). The standard approach to combining these notions employs a "separate-but-equal" hybrid approach: endowing a model with two modules or systems of representation, one specialized for rule-based representations and one for statistical or similarity-based representations, and then letting these two modules compete or cooperate to solve some learning task. The ATRIUM model of Erickson and Kruschke (1998) is a good example of this approach, where a rule module and a similarity module are trained in parallel, and a gating module arbitrates between their predictions at decision time.

We argue here for a different, more unified approach to integrating rules and statistics. Rules expressed in a flexible concept language provide a single unitary representation; statistics provides not a complementary form of representation, but the rational inductive mechanism that maps from observed data to the concept language. We thus build on the insights of Shepard (1987) and Tenenbaum (2000) that the effects of similarity and rules can both emerge from a single model: one with a single representational system of rule-like hypotheses, learned via a single rational inductive mechanism that operates according to the principles of Bayesian statistics.

Effect of the Specific Concept Grammar

Although we focused on one concept language, based on a DNF representation, other concept languages are possible and the choice of language should affect the performance of a grammar-based induction model. For instance, grammars which are incomplete (lacking conjunction, say, hence unable to capture all extensions) fail to predict the flexibility of human learning. We focused here on only the DNF grammar for simplicity of exposition, because our goal has been to show the viability of a probabilistic grammar-based approach to concept learning, but this should not be taken to imply an endorsement of the DNF grammar as *the* correct concept grammar. Other possibilities capture aspects of other proposals about concepts and concept learning, which may be appropriate for different settings.

For instance, there is increasing evidence that causal relationships play an important role in concept use and formation (see e.g. Rehder, 2003; Sloman et al., 1998). We could attempt to capture causal regularities amongst features by generating sets of "causal laws": conjunctions of implications amongst features (this representation for concepts was suggested in Feldman (2006)). In Fig. 14(a) we indicate what a grammar for this language might look like; see N. D. Goodman et al. (In Press) for studies with this concept grammar. Another possible grammar (Fig. 14(b)), inspired by the representation learned by the RULEX model (Nosofsky et al., 1994), represents concepts by a conjunctive rule plus a set of exceptions. Finally, it is possible that context-free grammars are not the best formalism in which to describe a concept language: graph-grammars and categorial grammar, for instance, both have attractive properties.

(a)

$S \rightarrow \forall x\, \ell(x) \Leftrightarrow I$
$I \rightarrow (C \Rightarrow P) \wedge I$
$I \rightarrow T$
$C \rightarrow P \wedge C$
$C \rightarrow T$
$P \rightarrow F_i$
$F_i \rightarrow f_i(x) = 1$
$F_i \rightarrow f_i(x) = 0$

(b)

$S \rightarrow \forall x\, \ell(x) \Leftrightarrow ((C) \wedge E)$
$E \rightarrow \neg(C) \wedge E$
$E \rightarrow T$
$C \rightarrow P \wedge C$
$C \rightarrow T$
$P \rightarrow F_i$
$F_i \rightarrow f_i(x) = 1$
$F_i \rightarrow f_i(x) = 0$

Figure 14. (a) An Implication Normal Form grammar. (b) A rule-plus-exceptions grammar inspired by Nosofsky et al. (1994).

Rationality of Hypothesis Sampling

In the first part of this paper we gave a rational analysis of rule-based concept learning, assuming that the output of this learning was a posterior distribution on possible concept meanings. In subsequent sections we showed that this rational analysis predicted human behavior at the group and individual level, at least when supplemented by the additional assumption of hypothesis sampling (that each learner arrives at one or a small set of rules distributed according to the posterior). What is the status of the supplemented analysis? We may ask this along two dimensions: the level of analysis (in

the sense of Marr (1982)), and the claim of rationality. Thus we have two questions: is hypothesis sampling a process-level correction to the computational-level analysis, or an extension at the computational level of analysis? And, in either case, is it a "rational" extension to the analysis?

There are at least two ways in which it is reasonable to think of hypothesis sampling as a rational correction to the Bayesian analysis. First, as the scale of the learning problem increases, exact evaluation of Bayesian analyses quickly becomes intractable. This suggests that a rational analysis should account for limited computational resources available to the organism by describing a rational process for approximating the Bayesian solution. Bounded-optimal algorithmic approximation of this sort was an important component of the rational model of Anderson (1991). In recent years, Bayesian models in AI and machine learning have relied heavily on methods for efficiently computing approximate probabilities from complex models (Russell & Norvig, 2002). The most general and straightforward approximation methods are based on Monte Carlo sampling—a randomized procedure for constructing a set of hypotheses that are drawn from the Bayesian posterior. The ubiquity of sampling-based approximation methods suggests that they could provide the basis for bounded-optimal analysis of human learning (Sanborn et al., 2006)—an approach which could be consistent with our findings that individual subjects' behavior can be well explained as samples from the posterior. One appealing version of this approach, explored by Sanborn et al. (2006) for similarity-based representations, is based on sequential Monte Carlo (or "particle filter") algorithms (Doucet, De Freitas, & Gordon, 2001). These algorithms update a set of hypotheses in response to new evidence in a way that looks very much like simple hypothesis testing, but the details of the "hypothesis update" procedure guarantee that the resulting hypotheses are samples from the full Bayesian posterior. (The results for several simulations reported in this paper were verified by using such a sequential Monte Carlo algorithm, based on Chopin (2002), but we have not yet investigated the behavior of this algorithm in detail.) Thus, it is possible that the hypothesis sampling assumption is rational in the sense of being a rational process-level correction to the computational-level rational analysis.

There may also be a deeper sense in which the hypothesis sampling assumption corrects the Bayesian analysis, at the computational level, by altering the competence that is being described. If the output of the concept-learning competence is a mental representation, not merely classification behavior, then we must take care that the end result of a model of concept learning is a reasonable candidate for mental representation. In particular, because the full distribution on formulae in the Rational Rules model is infinite, the output representation can only be a summary of this distribution. The hypothesis sampling assumption then embodies the claim that this summary takes the form of a small set of rules—and that this approximates the format of concepts in the mind.

Additional work will be required to establish either the optimality of hypothesis sampling as a process-level approximation, or the plausibility of sets of rules as a mental rep-

resentation formed by concept learning. In the mean time, viewing hypothesis sampling as a useful and intriguing auxiliary hypothesis to the rational analysis may be the best course.

Limitations, Extensions, and Future Directions

Learning tasks. In this paper we have only modeled supervised learning of a single concept. As pointed out in a number of places (e.g. Love et al., 2004) it is important for models of concept learning to account for human behavior over a range of learning situations and a variety of inference tasks.

It should be possible to extend the Rational Rules model to unsupervised and semi-supervised learning tasks, by employing a "strong-sampling" likelihood (which assumes that examples are sampled from those with a given label), as in Tenenbaum and Griffiths (2001). Similar effects should emerge as those found in other Bayesian models which use strong-sampling (Tenenbaum, 1999a)—foremost a size principle, favoring more restrictive hypotheses over more general. This size principle also enables learning concepts from only positive examples (Tenenbaum & Xu, 2000; Xu & Tenenbaum, 2007).

We indicated earlier that the feature predicates used as primitives throughout this paper should be thought of as just one simple case of preexisting concepts that can be used as atoms for new concepts. Indeed, compositionality suggests that once a concept has been learned it is available as a building block for future concepts—and this effect has been demonstrated for human learning (Schyns & Rodet, 1997). We may extend the Rational Rules model to *systems* of concepts in exactly this way: by adding each learned concept as a primitive predicate to the concept grammar used to learn the next concept. This will predict certain synergies between concepts which should be empirically testable. (For instance, having learned that "daxes" are red squares, it might be easier to learn that a "blicket" is a fuzzy dax than it would have been to learn the meaning of blicket alone—a fuzzy red square.) In such synergies we begin to see the real power of compositional representations for concepts.

Process-level description. While we have not attempted here to predict processing measures of concept learning, there are natural approaches to bridge between our analysis and such measures. For instance, it is likely that response times for individual categorization decisions can be modeled by making processing assumptions similar to those in Lamberts (2000). Specifically, we may assume that noisy perceptual observations of features are accumulated at a constant rate, that the log-likelihood of a particular response given the (set of) concept formulae and observations is accumulated over time, and that a decision is made when this accumulating log-likelihood reaches a confidence threshold. This would result in a diffusion process (Luce, 1986; Ratcliff, Zandt, & McKoon, 1999) in which the diffusion rate depends on properties of the stimulus with respect to the concept language. It is important to note that the speed-limiting step

in this process is perceptual information gathering, not evaluation of rules. Thus, contra Fodor (1998), we would not expect that the representational complexity of concepts need be reflected in response times.

Sequential effects in learning have been taken to be a particularly puzzling set of phenomena from the Bayesian point of view (Kruschke, 2006). A parallel puzzle for our analysis, is how individuals are able to arrive at a set of samples from the posterior. Both of these puzzles may be addressed by considering the process of learning over the course of an experiment. Indeed, it is likely that rational process models, such as the sequential Monte Carlo models described above, will be able to give a more detailed treatment of the course of human learning. This includes refined learning curves and explanations of sequential effects—such as the greater weight of early examples in concept learning (Anderson, 1990)—which might otherwise be puzzling from the Bayesian point of view. A proper investigation of these questions could become a whole research program on its own.

Representation. One of the primary advantages of the Rational Rules model is the ease with which it can be extended to incorporate new, more powerful, representational abilities. We have already seen a simple case of this flexibility, when we extended the Rational Rules model to continuous feature dimensions by simply adding decision-boundary predicates (e.g. $f_1(x) < c$). In N. D. Goodman et al. (In Press) we have explored concepts defined by their role in a relational system (the importance of such concepts has been pointed out recently (Markman & Stilwell, 2001; Gentner & Kurtz, 2005)). For instance, the concept "poison" can be represented

$$\forall x \, \text{poison}(x) \Leftrightarrow (\forall y \, \text{in}(x, y) \wedge \text{organism}(y) \Rightarrow \text{injured}(y)),$$

or, "a poison is something that causes injury when introduced into an organism". To capture this rich set of concepts only a simple extension of the concept grammar is needed (including relational feature predicates and additional quantifiers). One can imagine making similar alterations to the concept language to include representations required, for instance, in social cognition or naive physics.

Conclusion

Our work here can be seen as part of a broader theme emerging in cognitive science, AI, and machine learning, where logical representations and statistical inference are seen as complementary rather than competing pararadigms. In the context of concept learning, the integration of statistical learning and rule-based representations may help us to understand how people can induce richly structured concepts from sparse experience. This is crucial if concepts are to play a foundational role in our understanding of the mind: concepts must serve many purposes—classification, theory building, planning, communication and cultural transmission—which require the inferential flexibility of statistics and the representational power of mathematical logic.

The proposal that concepts are represented by phrases in a concept language is not new in cognitive science—indeed this is a principal component of the language of thought hypothesis (Fodor, 1975). Nor is the idea of analyzing cognition by considering a rational Bayesian agent new: ideal observers have been prominent in vision research (Geisler, 2003) and cognitive psychology (Shepard, 1987; Anderson, 1990; Chater & Oaksford, 1999). However, the combination of these ideas leads to an exciting project not previously explored: Bayesian analysis of the language of thought. Although this paper represents only the first steps in such a program, we have shown that rigorous results are possible and that they can provide accurate models of basic cognitive processes.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*, 263–308.

Briscoe, E., & Feldman, J. (2006). Conceptual complexity and the bias-variance tradeoff. In *Proceedings of the conference of the cognitive science society* (pp. 1038–1043).

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Cassirer, E. (1946). *Language and myth*. Harper & Bros. (Translated by Suzanne K. Langer)

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, *3*(2), 57–65.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, *89*(3), 539–552.

Doucet, A., De Freitas, N., & Gordon, N. (Eds.). (2001). *Sequential monte carlo methods in practice*.

Enderton, H. B. (1972). *A mathematical introduction to logic*. New York: Academic Press.

Erickson, M. A., & Kruschke, J. K. (1998, Jun). Rules and exemplars in category learning. *J Exp Psychol Gen*, *127*(2), 107–140.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630-633.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227–232.

Feldman, J. (2004). How surprising is a simple pattern? Quantifying "Eureka!". *Cognition*, *93*, 199–224.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology*, *50*, 339–368.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press: Cambridge, MA.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. New York: Oxford University Press.

Geisler, W. S. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 825–837). MIT press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.

Gentner, D., & Kurtz, K. (2005). Categorization inside and outside the lab. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), (pp. 151–175). APA.

Goldstone, R. (1996). Isolated and interrelated concepts. *Mem Cognit*, *24*(5), 608–28.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge: Harvard University Press.

Goodman, N. D., Tenenbaum, J. B., Griffiths, T. L., & Feldman, J. (In Press). Compositionality in rational analysis: Grammar-based induction for concept learning. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for bayesian cognitive science*. Oxford: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 285–386.

Halpern, J. Y., & Pearl, J. (2001). Causes and explanations: a structural-model approach. Part II: Explanations. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*.

Humboldt, W. von. (1863). *On language*.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Johansen, M., & Palmeri, T. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*(4), 482–553.

Kruschke, J. K. (1992, Jan). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: implications for backpropagation models. *Connection Science*, *5*(1), 3–36.

Kruschke, J. K. (2006). Locally Bayesian Learning with Applications to Retrospective Revaluation and Highlighting. *Psychological Review*, *113*, 677–699.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychol Rev*, *107*(2), 227–260.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835.

Love, B. C., Gureckis, T. M., & Medin, D. L. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Luce, R. D. (1959). *Individual choice behavior*. New York: John Wiley.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford, England: Oxford University Press.

Maddox, W. T., & Ashby, F. G. (1993, Jan). Comparing decision bound and exemplar models of categorization. *Percept Psychophys*, *53*(1), 49–70.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, *13*(4), 329–358.

Marr, D. (1982). *Vision*. Freeman Publishers.

McKinley, S. C., & Nosofsky, R. M. (1993). *Attention learning in models of classification*. ((Cited in Nosofsky, Palmeri, and McKinley, 1994))

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 37–50.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability

in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 355–368.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242–279.

Mervis, C. B., & Rosch, E. H. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.

Metropolis, A. W., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

Navarro, D. J. (2006). From natural kinds to complex categories. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 621–626).

Nosofsky, R. M. (1986, Mar). Attention, similarity, and the identification-categorization relationship. *J Exp Psychol Gen*, *115*(1), 39–61.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989, Mar). Rules and exemplars in categorization, identification, and recognition. *J Exp Psychol Learn Mem Cogn*, *15*(2), 282–304.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review*, *3*(2), 222–226.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*, 345–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition.* Oxford: Oxford University Press.

Osherson, D. N., & Smith, E. E. (1981, Feb). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35–58.

Pinker, S. (1997). *How the mind works.* W. W. Norton & Company.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Pothos, E. M. (2005, Feb). The rules versus similarity distinction. *Behav Brain Sci*, *28*(1), 1–14; discussion 14-49.

Ratcliff, R., Zandt, T. V., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychol Rev*, *106*(2), 261–300.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393-407.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1141-1159.

Russell, S. J., & Norvig, P. (2002). *Artificial intelligence: A modern approach* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization. In *Proceedings of the twenty-eighth annual conference of the cognitive science society.*

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts (with commentary). *Behavioral and Brain Sciences*, *21*, 1–54.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *23*, 681–696.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shepard, R. N., & Chang, J. J. (1963). Stimulus generalization in the learning of classifications. *J Exp Psychol*, *65*, 94–102.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Sloman, S. A., Love, B. C., & Ahn, W. kyoung. (1998). Feature Centrality and Conceptual Coherence. *Cognitive Science*, *22*, 189–228.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004, Sep). Category learning in rhesus monkeys: a study of the shepard, hovland, and jenkins (1961) tasks. *J Exp Psychol Gen*, *133*(3), 398–414.

Tenenbaum, J. B. (1999a). *A Bayesian framework for concept learning.* Unpublished doctoral dissertation, Massachussets Institute of Technology, Cambridge, MA.

Tenenbaum, J. B. (1999b). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems 11* (p. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems 12.* Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.

Tenenbaum, J. B., & Xu, F. (2000). Word learning as Bayesian inference. In *Proceedings of the 22nd annual conference of the cognitive science society.* Hillsdale, NJ: Erlbaum.

Wittgenstein, L. (1953). *Philosophical investigations.* New York, MacMillan.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*.

Appendix A
Derivation of the Prior

We begin with the expression for the prior probability of a formula given production probabilities τ (Eq. 3):

$$P(F|\mathcal{G}, \tau) = \prod_{s \in \text{Deriv}_F} \tau(s), \qquad (10)$$

However, we wish to allow uncertainty over the values of the production probabilities, thus:

$$
\begin{aligned}
P(F|\mathcal{G}) &= \int P(F, \tau|\mathcal{G}) d\tau \\
&= \int P(\tau) P(F|\tau, \mathcal{G}) d\tau \\
&= \int P(\tau) \left(\prod_{s \in \text{Deriv}_F} \tau(s) \right) d\tau,
\end{aligned}
\qquad (11)
$$

where $P(\tau)$ is the prior probability for a given set of production probabilities. (Note that the integral of Eq. 11 is over all production probabilities—a product of simplices.) Since

we have no a priori reason to prefer one set of values for τ to another, we apply the principle of indifference (Jaynes, 2003) to select the least informative prior: $P(\tau) = 1$. The probability of a formula becomes:

$$P(F|\mathcal{G}) = \int \left(\prod_{s \in \text{Deriv}_F} \tau(s) \right) d\tau. \qquad (12)$$

We may simplify this equation by recognizing the integral as a Multinomial-Dirichlet form (see Gelman, Carlin, Stern, & Rubin, 1995):

$$P(F|\mathcal{G}) = \prod_{Y \in \text{ non-terminals of } \mathcal{G}} \frac{\beta(\mathbf{C}_Y(F) + \mathbf{1})}{\beta(\mathbf{1})}. \qquad (13)$$

Here the expression $\mathbf{C}_Y(F)$ is the vector of counts of the productions for non-terminal symbol Y in Deriv_F (and $\mathbf{1}$ is the vector of ones of the same length). The function $\beta(c_1, c_2, ..., c_n)$ is the multinomial beta function (i.e. the normalizing constant of the Dirichlet distribution for pseudo-counts $c_1, c_2, ..., c_n$, see Gelman et al. (1995)), which may be written in terms of the Gamma function:

$$\beta(c_1, c_2, ..., c_n) = \frac{\prod_{i=1}^{n} \Gamma(c_i)}{\Gamma(\sum_{i=1}^{n} c_i)} \qquad (14)$$

Appendix B
Derivation of the Likelihood

In this appendix we will derive the likelihood function, Eq. 5. The DNF grammar generates a concept language of formulae, and it will be useful in what follows to write the formulae in two parts, the "quantified" part $\forall x \ (\ell(x) \Leftrightarrow \text{Def}(x))$, and the "definition" part $\text{Def}(x)$, where the definition part is a disjunction of conjunctions of predicate terms.

We wish to define the truth-value of $\text{Def}(x)$ for a given object x. Following the usual approach in mathematical logic (Enderton, 1972) the evaluation of $\text{Def}(x)$ is given recursively:

1. $\text{Def}(x)$ is a *term*.

2. If a term is a feature predicate, such as $f_1(x)=0$, then it can be evaluated directly (presuming that we know the feature values for the object x).

3. If a term is a conjunction of other terms, $A(x) \wedge B(x)$, then it is True if and only if each of the other terms is True.

4. If a term is a disjunction of other terms, $A(x) \vee B(x)$, then it is True if and only if any of the other terms is True.

At first this definition may appear vacuous, but it provides a concrete procedure for reducing evaluation of $\text{Def}(x)$ to evaluation of (primitive) feature predicates. The steps of this reduction exactly parallel the syntactic derivation from which the formula was built, hence providing a *compositional semantics*. (For a more careful treatment of the general issue of compositionality in Bayesian models, see N. D. Goodman et al. (In Press).) We now have an evaluation procedure that assigns a truth value to the definition part of each formula for each object; we write this truth value also as $\text{Def}(x)$.

The natural reading of the "quantified" part of the formula is "the formula is true if, for every example, the label is true if and only if $\text{Def}(x)$ is true". Using this as the only constraint on the likelihood, $P(\ell(\mathbf{E}), \mathbf{E}|F)$ is nonzero only when F is true, and otherwise uniform. (This is again an application of the principle of indifference: we add no additional assumptions, and are thus indifferent among the worlds compatible with the formula.) Thus, with logical True/False interpreted as probability 1/0:

$$P(\ell(\mathbf{E}), \mathbf{E}|F) \propto \bigwedge_{x \in \mathbf{E}} \ell(x) \Leftrightarrow \text{Def}(x). \qquad (15)$$

We needn't worry about the constant of proportionality in Eq. 15, because it is independent of the formula F (this holds because there is exactly one labeling which makes the formula true for each set of examples).

If we knew that the observed labels were correct, and we required an explanation for each observation, we could stop with Eq. 15. However, we assume that there is a probability e^{-b} that any given example is an outlier (ie. an unexplainable observation which should be excluded from induction)[9]. Writing I for the set of examples which are *not* outliers (which ranges over subsets of $\mathbf{E}$), the likelihood becomes:

$$
\begin{aligned}
P(\ell(\mathbf{E}), \mathbf{E}|F) &= \sum_{I \subseteq \mathbf{E}} P(I) P(\ell(I), I|F) \\
&\propto \sum_{I \subseteq \mathbf{E}} (1 - e^{-b})^{|I|} (e^{-b})^{|\mathbf{E}| - |I|} \bigwedge_{x \in I} \ell(x) \Leftrightarrow \text{Def}(x) \\
&= \sum_{I \subseteq \{x \in \mathbf{E} | \ell(x) \Leftrightarrow \text{Def}(x)\}} (1 - e^{-b})^{|I|} (e^{-b})^{|\mathbf{E}| - |I|} \\
&= e^{-b Q_\ell(F)},
\end{aligned} \qquad (16)
$$

where the last step follows from the Binomial Theorem. We have used the abbreviation $Q_\ell(F) = |\{x \in \mathbf{E} | \neg(\ell(x) \Leftrightarrow \text{Def}(x))\}|$ (this is the number of example objects which do not satisfy the definition asserted by F).

Appendix C
A Grammar-based Monte Carlo Algorithm

The expected generalization probability of Eq. 7 cannot be directly evaluated, since the set of formulae is infinite. However, this expectation may be approximated by importance sampling from the posterior distribution (Eq. 6). We now sketch a Markov chain Monte Carlo algorithm for sampling from the posterior distribution[10]. This algorithm applies generally for inference over a grammatically structured hypothesis space.

---

[9] We have written the outlier probability as an exponential merely for later convenience.

[10] Software implementing this algorithm is available from the authors' web site.

We wish to define a Markov chain on the space of parse trees (the grammatical derivations, up to order), in which each step is compatible (in some intuitive sense) with the structure of the grammar, $\mathcal{G}$. We do so by the Metropolis-Hastings procedure (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), using *subtree-regeneration* proposals which formalize the intuitive idea "generate a proposal by changing a (syntactically coherent) part of the given formula."

First fix an auxiliary probabilistic CFG, to be used in proposal generation, by choosing a convenient set of production probabilities, σ, for $\mathcal{G}$. Now the subtree-regeneration proposal is generated from parse tree T as follows: select $n \in T$ uniformly at random from among the non-terminal nodes of T, remove all nodes of T below n, now regenerate the tree below n according to the stochastic rules of $\mathcal{G}_\sigma$, to get proposal T'. (If $T' = T$, repeat the process.) Each proposal is accepted with probability equal to the minimum of 1 and:

$$\frac{P(\mathbf{E}, \ell(\mathbf{E})|F_{T'})}{P(\mathbf{E}, \ell(\mathbf{E})|F_T)} \cdot \frac{P(T'|\mathcal{G})}{P(T|\mathcal{G})} \cdot \frac{|T|}{|T'|} \cdot \frac{P(T|\mathcal{G}, \sigma)}{P(T'|\mathcal{G}, \sigma)}. \quad (17)$$

Where F_T is the formula associated with parse T, and $|T|$ is the number of non-terminal symbols in T. For the Rational Rules posterior each of the terms in Eq. 17 may be easily evaluated: the likelihood by Eq. 5, the prior by Eq. 4, and $P(T'|\mathcal{G}, \sigma)$ by Eq. 3.

Detailed balance follows as usual in the Metropolis-Hastings prescription, supplemented with some graph and counting arguments. Since we could amputate at the root, generating a completely new parse tree from the start symbol, this proposal scheme is ergodic. From ergodicity and detailed balance we may conclude that this Markov chain converges to $P(T|\mathbf{E}, \ell(\mathbf{E}), \mathcal{G})$ in distribution. By interpreting each parse T as its formula F_T we generate samples from the Rational Rules posterior.

In the reported results the RR$_{\text{DNF}}$ model was approximated by using this Monte Carlo algorithm. Except where otherwise noted 30,000 iterations were used to approximate each reported value, and convergence of most approximations was verified by doing five independent simulations (reported results are the mean of these five runs—hence an aggregate of 150,000 samples—and error bars, where given, are standard error over the five runs).

Appendix D
Derivation of Feature Weights

In the main text we have defined the posterior feature weights by:

$$\sum_F \frac{\text{count}(f_i \in F)}{\text{cplx}(F)} P(F|\mathbf{E}, \ell(\mathbf{E})), \quad (18)$$

and used them as an intuitive measure of the importance of each feature, as estimated by the model. We will now show that these weights are related in a simple way to the posterior expectations of the production probabilities for productions

$P \rightarrow F_i$. Because these production probabilities determine the relative importance of the features in generating a concept, their posterior expectations capture the relative informativeness of the features.

The posterior probability of the production probabilities τ_P for non-terminal P, given a formula, is:

$$\begin{aligned}
P(\tau_P|F) &= \frac{P(\tau_P)P(F|\tau_P)}{\int P(\tau_P)P(F|\tau_P)d\tau_P} \\
&= \frac{P(F|\tau_P)}{\int P(F|\tau_P)d\tau_P} \\
&= \frac{\prod_{i=1}^{N}(\tau_{P,i})^{\text{count}(f_i \in F)}}{\int \prod_{i=1}^{N}(\tau_{P,i})^{\text{count}(f_i \in F)}d\tau_P} \\
&= \frac{\prod_{i=1}^{N}(\tau_{P,i})^{\text{count}(f_i \in F)}}{\beta(\text{count}(f_i \in F) + \mathbf{1})}.
\end{aligned} \quad (19)$$

Where $\mathbf{1}$ indicates the vector of all ones. The expected value of production probability τ_{Pk} (for production $P \rightarrow F_k$), given formula F, is then:

$$\begin{aligned}
E_{P(\tau_P|F)}(\tau_{Pk}) &= \int \tau_{Pk} P(\tau_P|F)d\tau_P \\
&= \frac{\int \tau_{Pk} \prod_{i=1}^{N}(\tau_{P,i})^{\text{count}(f_i \in F)}d\tau_P}{\beta(\text{count}(f_i \in F) + \mathbf{1})} \\
&= \frac{\beta(\text{count}(f_i \in F) + \mathbf{1} + \delta_{\mathbf{k}})}{\beta(\text{count}(f_i \in F) + \mathbf{1})}.
\end{aligned} \quad (20)$$

Where δ_k indicates the vector with a 1 in the k-place, and zeros elsewhere. If we expand the beta functions in terms of gamma functions, most terms cancel giving us:

$$\begin{aligned}
E_{P(\tau_P|F)}(\tau_{Pk}) &= \frac{\Gamma(\text{count}(f_k \in F) + 2) \cdot \Gamma(N + \sum_i \text{count}(f_i \in F))}{\Gamma(\text{count}(f_k \in F) + 1) \cdot \Gamma(N + 1 + \sum_i \text{count}(f_i \in F))} \\
&= \frac{\Gamma(\text{count}(f_k \in F) + 2) \cdot \Gamma(N + \text{cplx}(F))}{\Gamma(\text{count}(f_k \in F) + 1) \cdot \Gamma(N + 1 + \text{cplx}(F))} \\
&= \frac{1 + \text{count}(f_k \in F)}{N + \text{cplx}(F)}
\end{aligned}$$
$$(21)$$

where the last simplification follows from the recursion $\Gamma(z + 1) = z\Gamma(z)$. Finally, the posterior expectation for production probability τ_{Pk} is given by:

$$E_{P(\tau_P|\mathbf{E}, \ell(\mathbf{E}))}(\tau_{Pk}) = \sum_F \frac{1 + \text{count}(f_k \in F)}{N + \text{cplx}(F)} P(F|\mathbf{E}, \ell(\mathbf{E})) \quad (22)$$

Thus the feature weights, Eq. 18, are monotonically related to the posterior expectations of the production probabilities, Eq. 22. The primary difference between the two, which is unimportant for our purposes, is that features which are never used will have non-zero posterior expectation, but zero posterior weight.