Compositionality in rational analysis: Grammar-based induction for concept learning

Compositionality in Rational Analysis: Grammar-based Induction for Concept Learning

Noah D. Goodman[1], Joshua B. Tenenbaum[1], Thomas L. Griffiths[2], and Jacob Feldman[3]
[1]MIT; [2]University of California, Berkeley; [3]Rutgers University

Rational analysis attempts to explain aspects of human cognition as an adaptive response to the environment (Marr, 1982; Anderson, 1990; Chater, Tenenbaum, & Yuille, 2006). The dominant approach to rational analysis today takes an ecologically reasonable specification of a problem facing an organism, given in statistical terms, then seeks an optimal solution, usually using Bayesian methods. This approach has proven very successful in cognitive science; it has predicted perceptual phenomena (Geisler & Kersten, 2002; Feldman, 2001), illuminated puzzling effects in reasoning (Chater & Oaksford, 1999; Griffiths & Tenenbaum, 2006), and, especially, explained how human learning can succeed despite sparse input and endemic uncertainty (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001). However, there were earlier notions of the "rational" analysis of cognition that emphasized very different ideas. One of the central ideas behind logical and computational approaches, which previously dominated notions of rationality, is that meaning can be captured in the structure of representations, but that compositional semantics are needed for these representations to provide a coherent account of thought. In this chapter we attempt to reconcile the modern approach to rational analysis with some aspects of this older, logico-computational approach. We do this via a model—offered as an extended example—of human concept learning. In the current chapter we are primarily concerned with formal aspects of this approach; in other work (Goodman, Tenenbaum, Feldman, & Griffiths, in press) we more carefully study a variant of this model as a psychological model of human concept learning.

Explaining human cognition was one of the original motivations for the development of formal logic. George Boole, the father of digital logic, developed his symbolic language in order to explicate the rational laws underlying thought: his principal work, *An Investigation of the Laws of Thought* (Boole, 1854), was written to "investigate the fundamental laws of those operations of the mind by which reasoning is performed," and arrived at "some probable intimations concerning the nature and constitution of the human mind" (p. 1). Much of mathematical logic since Boole can be regarded as an attempt to capture the coherence of thought in a formal system. This is particularly apparent in the work, by Frege (1892), Tarski (1956) and others, on model-theoretic semantics for logic, which aimed to create formal systems both flexible and systematic enough to capture the complexities of mathematical thought. A central component in this program is *compositionality*. Consider Frege's Principle[1]: each syntactic operation of a formal language should have a corresponding semantic operation. This principle requires *syntactic* compositionality, that meaningful terms in a formal system are built up by combination operations, as well as *compatibility* between the syntax and semantics of the system.

When Turing, Church, and others suggested that formal systems could be manipulated by mechanical computers it was natural (at least in hindsight) to suggest that cognition operates in a similar way: meaning is manipulated in the mind by computation[2]. Viewing the mind as a formal computational system in this way suggests that compositionality should also be found in the mind; that is, that mental representations may be combined into new representations, and the meaning of mental representations may be decomposed in terms of the meaning of their components. Two important virtues for a theory of thought result (Fodor, 1975): productivity—the number of representations is unbounded because they may be boundlessly combined—and systematicity—the combination of two representations is meaningful to one who can understand each separately.

Despite its importance to the computational theory of mind, compositionality has seldom been captured by modern rational analyses. Yet there are a number of reasons to desire a compositional rational analysis. For instance, productivity of mental representations would provide an explanation of the otherwise puzzling ability of human thought to adapt to novel situations populated by new concepts—even those far beyond the ecological pressures of our evolutionary milieu (such as radiator repairs and the use of fiberglass bottom powerboats).

We will show in this chapter that Bayesian statistical methods can be fruitfully combined with compositional representational systems by developing such a model in the well-studied setting of concept learning. This addresses a long running tension in the literature on human concepts: similarity-based statistical learning models have provided a good understanding of how simple concepts can be learned (Medin & Schaffer, 1978; Anderson, 1991; Kruschke, 1992;

[1] Compositionality has had many incarnations, probably beginning with Frege, though this modern statement of the principle was only latent in Frege (1892). In cognitive science compositionality was best expounded by Fodor (1975). Rather than endorsing an existing view, the purpose of this chapter is to provide a notion of compositionality suited to the Bayesian modeling paradigm.

[2] If computation is understood as *effective* computation we needn't consider finer details: the Church-Turing thesis holds that all reasonable notions of effective computation are equivalent (partial recursive functions, Turing machines, Church's lambda calculus, etc.).

Tenenbaum & Griffiths, 2001; Love, Gureckis, & Medin, 2004), but these models did not seek to capture the rich structure surely needed for human cognition (Murphy & Medin, 1985; Osherson & Smith, 1981). In contrast, the representations we consider inherit the virtues of compositionality—systematicity and productivity—and are integrated into a Bayesian statistical learning framework. We hope this will signpost a road toward a deeper understanding of cognition in general: one in which mental representations are a systematically meaningful and infinitely flexible response to the environment.

In the next section we flesh out specific ideas of how compositionality may be interpreted in the context of Bayesian learning. In the remainder of the chapter we focus on concept learning, first deriving a model in the setting of feature-based concepts, which fits human data quite well, then extending to a relational setting for role-governed concepts.

Bayesian Learning and Grammar-based Induction

Learning is an important area of application for rational analysis, and much recent work has shown that inductive learning can often be described with Bayesian techniques. The ingredients of this approach are: a description of the data space from which input is drawn, a space of hypotheses, a prior probability function over this hypothesis space, 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 what data one would expect to observe if hypothesis h were correct. Inductive learning can then be described very simply: we wish to find the appropriate degree of belief in each hypothesis given some observed data, that is, the posterior probability $P(h|d)$. Bayes' theorem tells us how to compute this probability,

$$P(h|d) \propto P(h)P(d|h), \tag{1}$$

identifying the posterior probability as proportional to the product of the prior and the likelihood.

We introduce syntactic compositionality into this setting by building the hypothesis space from a few primitive elements using a set of combination operations. In particular, we will generate the hypothesis space from a (formal) grammar: the productions of the grammar are the syntactic combination rules, the terminal symbols the primitive elements, and the hypothesis space is all the well-formed sentences in the language of this grammar. For instance, if we used the simple grammar with terminal symbols a and b, a single non terminal symbol A, and two productions $A \rightarrow aA$ and $A \rightarrow b$, we would have the hypothesis space $\{b, ab, aab, aaab, ...\}$.

This provides syntactic structure to the hypothesis space, but is not by itself enough: compositionality also requires compatibility between the syntax and semantics. How can this be realized in the Bayesian setting? If "we understand a proposition when we know what happens if it is true" (Wittgenstein, 1921, Proposition 4.024), then the likelihood function captures the semantics of each hypothesis. Frege's principle then suggests that each syntactic operation should have a parallel semantic operation, such that the likelihood may be evaluated by applying the semantic operations appropriate to the syntactic structure of a hypothesis[3]. In particular, each production of the grammar should have a corresponding semantic operation, and the likelihood of a hypothesis is given by composition of the semantic operations corresponding to the productions in a grammatical derivation of that hypothesis.

Returning to the example above, let us say that our data space consists of two possible worlds—"heads" and "tails". Say that we wish the meaning of hypothesis *aab* to be "flip two fair coins and choose the 'heads' world if they both come up heads" (and similarly for other hypotheses). To capture this we first associate to the terminal symbol a the number $s(a) = 0.5$ (the probability that a fair coin comes up heads), and to b the number $s(b) = 1$ (if we flip no coins, we'll make a "heads" world by default). To combine these primitive elements, assign to the production $A \rightarrow aA$ the semantic operation which associates $s(a) \cdot s(A)$ to the left-hand side (where $s(a)$ and $s(A)$ are the semantic values associated to the symbols of the right-hand side). Now consider the hypothesis *aab*, which has derivation $A \rightarrow aA \rightarrow aaA \rightarrow aab$. By compatibility the likelihood for this hypothesis must be $P(\text{"heads"}|aab) = 0.5 \cdot 0.5 \cdot 1 = 0.25$. Each other hypothesis is similarly assigned its likelihood—a distribution on the two possible worlds "heads" and "tails". In general the semantic information needn't be a likelihood at each stage of a derivation, only at the end, and the semantic operations can be more subtle combinations than simple multiplication.

We call this approach *grammar-based induction*. Similar grammar-based models have long been used in computational linguistics (Chater & Manning, 2006), and have recently been used in computer vision (Yuille & Kersten, 2006). Grammars, of various kinds and used in various ways, have also provided structure to the hypothesis spaces in a few recent Bayesian models in high-level cognition (Tenenbaum, Griffiths, & Niyogi, 2007; Tenenbaum, Griffiths, & Kemp, 2006).

Grammar-based Induction for Concept Learning

In this section we will develop a grammar-based induction model of concept learning for the "classical" case of concepts which identify kinds of objects based on their features. The primary use of such concepts is to discriminate objects within the kind from those without (which allows an organism to make such subtle, but useful, discriminations as "friend-or-foe"). This use naturally suggests that the representation of such a concept encodes its recognition function: a rule which associates to each object a truth value ("is/isn't"), relying on feature values. We adopt this view for now, and so we wish to establish a grammatically generated hypothesis space of

[3] It is reasonable that the prior also be required to satisfy some compatibility condition. We remain agnostic about what this condition should be: it is an important question that should be taken up with examples in hand.

rules, together with compatible prior probability and likelihood functions, the latter relating rules to observed objects through their features.

We will assume for simplicity that we are in a *fully observed* world $\mathbf{W}$ consisting of a set of objects $\mathbf{E}$ and the feature values $f_1(x), \ldots, f_N(x)$ of each object $x \in \mathbf{E}$. (In the models developed below we could use standard Bayesian techniques to relax this assumption, by marginalizing over unobserved features, or an unknown number of objects (Milch & Russell, 2006).) We consider a single labeled concept, with label $\ell(x) \in \{1, 0\}$ indicating whether x is a positive or negative example of the concept. The labels can be unobserved for some of the objects—we describe below how to predict the unobserved labels given the observed ones.

Let us say that we've specified a grammar $\mathcal{G}$—which gives rise to a hypothesis space of rules $\mathcal{H}_\mathcal{G}$—a prior probability $P(F)$ for $F \in \mathcal{H}_\mathcal{G}$, and a likelihood function $P(\mathbf{W}, \ell(\mathbf{E})|F)$. We may phrase the learning problem in Bayesian terms: what degree of belief should be assigned to each rule F given the observed world and labels? That is, what is the probability $P(F|\mathbf{W}, \ell(\mathbf{E}))$? As in Eq. 1, this quantity may be expressed:

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

We next provide details of one useful grammar, along with an informal interpretation of the rules generated by this grammar and the process by which they are generated. We then give a more formal semantics to this language by deriving a compatible likelihood, based on the standard truth-functional semantics of first-order logic together with a simple noise process. Finally we introduce a simple prior over this language that captures a complexity bias—syntactically simpler rules are *a priori* more likely.

Logical Representation for Rules

We represent rules in a concept language which is a fragment of first-order logic. This will allow us to leverage the standard, compositional, semantics of mathematical logic in defining a likelihood which is compatible with the grammar. The fragment we will use is intended to express definitions of concepts as sets of implicational regularities amongst their features (Feldman, 2006). For instance, imagine that we want to capture the concept "strawberry" which is "a fruit that is red if it is ripe." This set of regularities might be written $(T \Rightarrow \text{fruit}(x)) \wedge (\text{ripe}(x) \Rightarrow \text{red}(x))$, and the definition of the concept "strawberry" in terms of these regularities as $\forall x\ \text{strawberry}(x) \Leftrightarrow ((T \Rightarrow \text{fruit}(x)) \wedge (\text{ripe}(x) \Rightarrow \text{red}(x)))$.

The full set of formulae we consider, which forms the hypothesis space $\mathcal{H}_\mathcal{G}$, will be generated by the context-free "implication normal form" (INF) grammar, Fig. 1. This grammar encodes some structural prior knowledge about concepts: labels are very special features (Love, 2002), which apply to an object exactly when the definition is satisfied, and implications among feature values are central parts of the definition. The importance of implicational regularities in human concept learning has been proposed by Feldman (2006), and is suggested by theories which emphasize causal regularities in category formation (Ahn, Kim, Lassaline, & Dennis, 2000; Sloman, Love, & Ahn, 1998; Rehder, 1999). We have chosen to use the INF grammar because of this close relation to causality. Indeed, each implicational regularity can be directly interpreted as a causal regularity; for instance, the formula $\text{ripe}(x) \Rightarrow \text{red}(x)$ can be interpreted as "being ripe *causes* being red". We consider the causal interpretation, and its semantics, in Appendix A.

(1)	$S \rightarrow \forall x\ \ell(x) \Leftrightarrow I$	"Definition of ℓ"
(2)	$I \rightarrow (C \Rightarrow P) \wedge I$	"Implication term"
(3)	$I \rightarrow T$	
(4)	$C \rightarrow P \wedge C$	"Conjunction term"
(5)	$C \rightarrow T$	
(6)	$P \rightarrow F_1$	"Predicate term"
	$\vdots$	
	$P \rightarrow F_N$	
(7)	$F_1 \rightarrow f_1(V) = 1$	"Feature value"
(8)	$F_1 \rightarrow f_1(V) = 0$	
	$\vdots$	
	$F_N \rightarrow f_N(V) = 1$	
	$F_N \rightarrow f_N(V) = 0$	
(9)	$V \rightarrow x$	"Object variable"

Figure 1. Production rules of the INF Grammar. S is the start symbol, and I, C, P, F_i, V the other non-terminals. There are N productions each of the forms (6), (7), and (8). In the right column are informal translations of the meaning of each non-terminal symbol.

Let us illustrate with an example the process of generating a hypothesis formula from the INF grammar. Recall that productions of a context-free grammar provide re-write rules, licensing replacement of the left-hand-side non-terminal symbol with the string of symbols on the right-hand-side. We begin with the start symbol S, which becomes by production (1) the "definition" $\forall x\ \ell(x) \Leftrightarrow I$. The non-terminal symbol I is destined to become a set of implication terms: say that we expand I by applying production (2) twice (which introduces two implications), then production (3) (which "ties off" the sequence). This leads to a conjunction of implication terms; we now have the rule:

$$\forall x\ \ell(x) \Leftrightarrow ((C \Rightarrow P) \wedge (C \Rightarrow P) \wedge T)$$

We are not done: C is non-terminal, so each C-term will be expanded into a distinct substring (and similarly for the other non-terminals). Each non-terminal symbol C leads, by productions (4) and (5),[4] to a conjunction of predicate terms:

$$\forall x\ \ell(x) \Leftrightarrow ((P \wedge P \Rightarrow P) \wedge (P \Rightarrow P))$$

Using productions (6) and (7) each predicate term becomes a feature predicate F_i, for one of the N features, and using production (8) each feature predicate becomes an assertion

[4] The terminal symbol T stands for logical True—it is used to conveniently terminate a string of conjunctions, and can be ignored. We now drop them for clarity.

that the i^{th} feature has a particular value[5] (i.e. $f_i(V) = 1$, etc.):

$$\begin{aligned}&\forall x\ \ell(x)\Leftrightarrow\\&\quad((f_1(V){=}1)\wedge(f_3(V){=}0)\Rightarrow(f_2(V){=}1))\\&\quad\wedge((f_1(V){=}0)\Rightarrow(f_4(V){=}1))\end{aligned}$$

Finally, there is only one object variable (the object whose label is being considered) so the remaining non-terminal, V denoting a variable, becomes x:

$$\begin{aligned}&\forall x\ \ell(x)\Leftrightarrow\\&\quad((f_1(x){=}1)\wedge(f_3(x){=}0)\Rightarrow(f_2(x){=}1))\\&\quad\wedge((f_1(x){=}0)\Rightarrow(f_4(x){=}1))\end{aligned}$$

Informally, we have generated a definition for ℓ consisting of two implicational regularities relating the four features of the object—the label holds when: f_2 is one if f_1 is one and f_3 is zero, and, f_4 is one if f_1 is zero. To make this interpretation precise, and useful for inductive learning, we must specify a likelihood function relating these formulae to the observed world.

Before going on, let us mention a few alternatives to the INF grammar. The association of definitions with entries in a dictionary suggests a different format for the defining properties: dictionary definitions typically have several entries, each giving an alternative definition, and each entry lists necessary features. From this we might extract a disjunctive normal form, or disjunction of conjunctions, in which the conjunctive blocks are like the alternative meanings in a dictionary entry. In Table 2(a) we indicate what such a DNF grammar might look like (see also Goodman et al., in press). Another possibility, inspired by the representation learned by the RULEX model (Nosofsky, Palmeri, & McKinley, 1994), represents concepts by a conjunctive rule plus a set of exceptions, as in Table 2(b). Finally, it is possible that context-free grammars are not the best formalism in which to describe a concept language: graph-grammars and categorial grammars, for instance, have attractive properties.

(a)

S	$\rightarrow$	$\forall x\ \ell(x)\Leftrightarrow(D)$
D	$\rightarrow$	$(C)\vee D$
D	$\rightarrow$	T
C	$\rightarrow$	$P\wedge C$
C	$\rightarrow$	T
P	$\rightarrow$	F_i
F_i	$\rightarrow$	$f_i(V)=1$
F_i	$\rightarrow$	$f_i(V)=0$
V	$\rightarrow$	x

(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	$\rightarrow$	$f_i(V)=1$
F_i	$\rightarrow$	$f_i(V)=0$
V	$\rightarrow$	x

Figure 2. (a) A dictionary-like DNF Grammar. (b) A rule-plus-exceptions grammar inspired by Nosofsky et al. (1994).

Likelihood: Compositional Semantics and Outliers

Recall that we wish the likelihood function to be compatible with the grammar in the sense that each production rule has a corresponding "semantic operation". These semantic operations associate some information to the non-terminal symbol on the left-hand side of the production given information for each symbol of the right-hand side. For instance the semantic operation for $F_1\rightarrow f_1(V){=}1$ might associate to F_1 the Boolean value *True* if feature one of the object associated to V has value 1. The information associated to F_1 might then contribute to information assigned to P from the production $P\rightarrow F_1$. In this way the semantic operations allow information to "filter up" through a series of productions.

Each hypothesis in the concept language has a grammatical derivation which describes its syntactic structure: a sequence of productions that generates this formula from the start symbol S. The semantic information assigned to most symbols can be of any sort, but we require the start symbol S to be associated with a probability value. Thus, if we use the semantic operations one-by-one beginning at the end of the derivation for a particular hypothesis, F, we will arrive at a probability—this defines the likelihood $P(\mathbf{W},\ell(\mathbf{E})|F)$. (Note that compositionality thus guarantees that we will have an efficient dynamic programming algorithm to evaluate the likelihood function.)

Since the INF grammar generates formulae of predicate logic, we may borrow most of the standard semantic operations from the model-theoretic semantics of mathematical logic (Enderton, 1972). Table 1 lists the semantic operation for each production of the INF grammar: each production which introduces a boolean operator has its conventional meaning, we diverge from standard practice only when evaluating the quantifier over labeled objects. Using these semantic rules we can evaluate the "definition" part of the formula to associate a function $D(x)$, from objects to truth values, to the set of implicational regularities. We are left (informally) with the formula $\forall x\ \ell(x)\Leftrightarrow D(x)$. To assign a probability to the S-term we could simply interpret the usual truth-value $\bigwedge_{x\in\mathbf{E}}\ell(x)\Leftrightarrow D(x)$ as a probability (that is, probability zero if the definition holds when the label doesn't). However, we wish to be more lenient by allowing exceptions in the universal quantifier—this provides flexibility to deal with the uncertainty of the actual world.

To allow concepts which explain only some of the observed labels, we assume that there is a probability e^{-b} that any given object is an outlier—that is, an unexplainable observation which should be excluded from induction. Any object which is not an outlier must satisfy the "definition" $\ell(x)\Leftrightarrow D(x)$. (Thus we give a probabilistic interpretation to the quantifier: its argument holds over a limited scope $S\subseteq\mathbf{E}$, with the subset chosen stochastically.) The likelihood be-

[5] For brevity we consider only two-valued features: $f_i(x)\in\{0,1\}$, though the extension to multiple-valued features is straightforward.

Table 1
The semantic type of each non-terminal symbol of the INF grammar (Fig. 1), and the semantic operation associated to each production.

Symbol	Semantic Type	Production	Semantic Operation
S	p	$S \rightarrow \forall x\ \ell(x) \Leftrightarrow I$	Universal quantifier with outliers (see text).
I	$e \rightarrow t$	$I \rightarrow (C \Rightarrow P) \wedge I$	For a given object, True if: the I-term is True, and, P-term is True if the C-term is True.
		$I \rightarrow T$	Always True.
C	$e \rightarrow t$	$C \rightarrow P \wedge C$	For a given object, True if both the P-term and C-term are True.
		$C \rightarrow T$	Always True.
P	$e \rightarrow t$	$P \rightarrow F_i$	True when the F_i term is True.
F_i	$e \rightarrow t$	$F_i \rightarrow f_i(V)$=val	True if the value of feature i for the object identified by the V-term is val.
V	e	$V \rightarrow x$	A variable which ranges over the objects $\mathbf{E}$.

Note: each semantic operation associates the indicated information with the symbol on the left-hand-side of the production, given information from each symbol on the right-hand-side. The semantic type indicates the type of information assigned to each symbol by these semantic rules: p a probability, t a truth value, e an object, and $e \rightarrow t$ a function from objects to truth values.

comes:

$$\begin{aligned} P(\mathbf{W}, \ell(\mathbf{E})|F) &\propto \sum_{S \subseteq \mathbf{E}} (1-e^{-b})^{|S|}(e^{-b})^{|\mathbf{E}|-|S|} \bigwedge_{x \in S} \ell(x) \Leftrightarrow D(x) \\ &= \sum_{S \subseteq \{x \in \mathbf{E} | \ell(x) \Leftrightarrow D(x)\}} (1-e^{-b})^{|S|}(e^{-b})^{|\mathbf{E}|-|S|} \\ &= e^{-b|\{x \in \mathbf{E}|\neg(\ell(x) \Leftrightarrow D(x))\}|}. \end{aligned} \tag{3}$$

The constant of proportionality is independent of F, so can be ignored for the moment, and the last step follows from the Binomial Theorem. If labels are observed for only a subset $\mathbf{Obs} \subseteq \mathbf{E}$ of the objects, we must adjust this likelihood by marginalizing out the unobserved labels. We make the *weak sampling* assumption (Tenenbaum & Griffiths, 2001), that objects to be labeled are chosen at random. This leads to a marginalized likelihood proportional to Eq. 3: $P(\mathbf{W}, \ell(\mathbf{Obs})|F) \propto P(\mathbf{W}, \ell(\mathbf{E})|F)$. In Appendix B we give the details of marginalization for both weak and strong sampling assumptions, and consider learning from positive examples.

A Syntactic Prior

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 grammatical derivation is assigned a probability, and the probability of the derivation is the product of the probabilities for these choices (the is the standard definition of a *probabilistic* context-free grammar used in computational linguistics (Chater & Manning, 2006)). The probability of a given derivation is:

$$P(T|\mathcal{G}, \tau) = \prod_{s \in T} \tau(s), \tag{4}$$

where $s \in T$ are the productions of the derivation T, and $\tau(s)$ their probability. The set of production probabilities, τ, must sum to one for each non-terminal symbol. Since the INF grammar is a unique production grammar—there is a single derivation, up to order, for each well-formed formula—the probability of a formula is given by Eq. 4. We will write F for both the formula and its derivation, hence Eq. 4 gives the prior probability for formulae. (In general, the probability of a formula is the sum of the probabilities of its derivations.) Note that this prior captures a syntactic simplicity bias: smaller formulae have shorter derivations, thus higher prior probability.

Since have no a priori reason to prefer one set of values for τ to another, we assume a uniform prior over the possible values of τ (i.e. we apply the principle of indifference (Jaynes, 2003)). The probability becomes:

$$\begin{aligned} P(T|\mathcal{G}) &= \int P(\tau) \prod_{s \in F} \tau(s) d\tau \\ &= \int \prod_{s \in F} \tau(s) d\tau \\ &= \prod_{Y \in N} \beta(|\{Y \in F\}| + 1), \end{aligned} \tag{5}$$

where $\beta(\vec{v})$ is the multinomial beta function (i.e. the normalizing constant of the Dirichlet distribution with vector of parameters $\vec{v}$, see Gelman, Carlin, Stern, and Rubin (1995)), and $|\{Y \in F\}|$ is the vector of counts of the productions for non-terminal symbol Y in the derivation of F.

The RR_{INF} Model

Collecting the above considerations, the posterior probability is:

$$\begin{aligned} &P(F|\mathbf{W}, \ell(\mathbf{Obs})) \\ &\qquad \propto \left(\prod_{Y \in N} \beta(|\{Y \in F\}| + 1) \right) e^{-b|\{x \in \mathbf{Obs}|\neg(\ell(x) \Leftrightarrow D(x))\}|}. \end{aligned} \tag{6}$$

This posterior distribution captures a trade-off between explanatory completeness and conceptual parsimony. On the

one hand, though some examples may be ignored as outliers, concepts which explain more of the observed labels are preferred by having a higher likelihood. On the other hand, simpler (ie. syntactically shorter) formulae are preferred by the prior.

Eq. 6 captures ideal learning. To predict empirical results we require an auxiliary hypothesis describing the judgments made by groups of learners when asked to label objects. We assume that the group average of the predicted label for an object e is the expectated value of $\ell(e)$ under the posterior distribution, that is:

$$P(\ell(e)|\mathbf{W}, \ell(\mathbf{Obs})) = \sum_{F \in \mathcal{H}_{\text{INF}}} P(\ell(e)|F) P(F|\mathbf{W}, \ell(\mathbf{Obs})). \quad (7)$$

Where $P(\ell(e)|F)$ will be 1 if $\ell(e)$ is the label of e required by F (this exists uniquely for hypotheses in our language, since they provide a "definition" of the label), and zero otherwise. This *probability matching* assumption is implicit in much of the literature on rational analysis. We will refer to this model, the posterior (Eq. 6) and the auxiliary assumption (Eq. 7), as the *Rational Rules* model of concept learning based on the INF grammar, or RR_{INF}.

We can also use Eq. 6 to predict the relative weights of formulae with various properties. For instance, the Boolean complexity of a formula (Feldman, 2000), $\text{cplx}(F)$, is the number of feature predicates in the formula. (E.g., $T \Rightarrow (f_1(x)=1)$ has complexity 1, while $(f_2(x)=0) \Rightarrow (f_1(x)=1)$ has complexity 2.) The weight of formulae with complexity C is the total probability under the posterior of such formulae:

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

Similarly, the weight of a feature in formula F is the number of times this feature is used divided by the complexity of F, and the total feature weight is the posterior expectation of this weight—roughly, the expected importance of this feature.

Comparison with Human Concept Learning

The RR_{INF} model provides a simple description of concept learning: from labeled examples one forms a posterior probability distribution over the hypotheses expressible in a concept language of implicational regularities. How well does this capture actual human concept learning? We compare the predicted generalization rates to human data from two influential experiments.

The second experiment of Medin and Schaffer (1978) is a common first test of the ability of a model to predict human generalizations on novel stimuli. This experiment used the category structure shown in Table 2 (we consider the human data from the Nosofsky et al. (1994) replication of this experiment, which counter-balanced physical feature assignments): participants were trained on labeled positive examples A1…A5, and labeled negative examples[6] B1…B4, the objects T1…T7 were unlabeled transfer stimuli.

As shown in Table 2 the best fit of the model[7] to human data is quite good: R^2=0.97. Other models of concept learning are also able to fit this data well: for instance R^2=0.98 for RULEX, a process model of rule learning (Nosofsky et al., 1994), and R^2=0.96 for the context model of Medin and Schaffer (1978). It is worth noting, however, that the RR_{INF} model has only a single parameter (the outlier parameter b), while each of these models has at least four parameters.

Table 2
The category structure of Medin & Schaffer (1978), with the human data of Nosofsky et al. (1994), and the predictions of the Rational Rules model at b=1.

Object	Feature Values	Human	RR_{INF}
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	0.22
B4	1111	0.15	0.08
T1	0110	0.56	0.57
T2	0111	0.41	0.44
T3	0000	0.82	0.94
T4	1101	0.40	0.44
T5	1010	0.32	0.29
T6	1100	0.53	0.57
T7	1011	0.20	0.14

We may gain some intuition for the RR_{INF} model by examining how it learns this concept. In Fig. 3(a) we have plotted the posterior complexity distribution after learning, and we see that the model relies mostly on single-feature rules. In Fig. 3(b) we have plotted the posterior feature weights, which show greater use of the first and third features than the others. Together these tell us that the RR_{INF} model focuses primarily on single feature rules using the first and third features (i.e. $\forall x\ \ell(x) \Leftrightarrow (T \Rightarrow (f_1(x)=0))$ and $\forall x\ \ell(x) \Leftrightarrow (T \Rightarrow (f_3(x)=0))$), with much smaller contributions from other formulae.

The object T3=0000, which never occurs in the training set, 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. This prototype is enhanced relative to the other transfer stimuli: T3 is, by far, the most likely transfer object to be classified as category A by human learners. The Rational Rules model predicts this prototype enhancement effect (Posner & Keele, 1968) because the dominant formulae $\forall x\ \ell(x) \Leftrightarrow (T \Rightarrow (f_1(x)=0))$ and $\forall x\ \ell(x) \Leftrightarrow (T \Rightarrow (f_3(x)=0))$

[6] Participants in this study and the next were actually trained on a pair of mutually exclusive concepts A and B. For simplicity, we account for this by averaging the results of the RR_{INF} model where A is the category and B the complement with vice versa. More subtle treatments are possible.

[7] We have optimized very roughly over the parameter b, taking the best fit from b=1, ..., 8. Model predictions were approximated by Monte Carlo simulation.

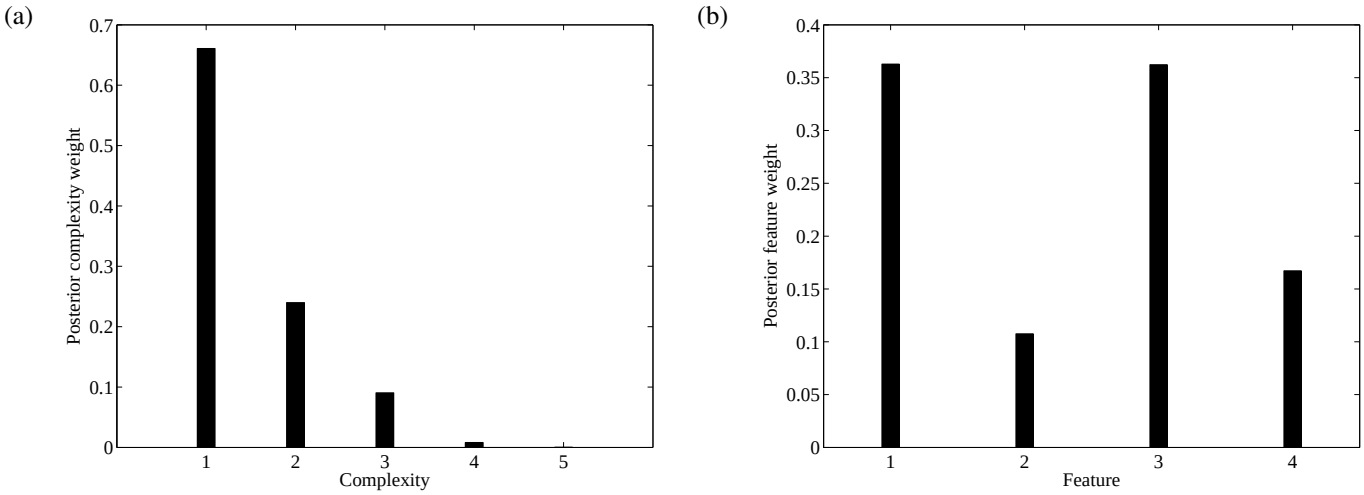

Figure 3. (a) Posterior complexity distribution (portion of posterior weight placed on formula with a given number of feature literals) for the category structure of Medin & Schaffer (1978), see Table 2. (b) Posterior feature weights.

agree on the categorization of T3 while they disagree on many other stimuli. Thus, together with many lower probability formulae, these hypotheses enhance the probability that T3 is in category A, relative to other training stimuli.

A similar effect can be seen for the prototype of category B, the object B4=1111, which *is* in the training set. Though presented equally often as the other training examples it is judged to be in category B far more often in the test phase. This enhancement, or greater degree of typicality, is often taken as a useful proxy for category centrality (Mervis & Rosch, 1981). The Rational Rules model predicts the typicality effect in a similar way.

Another important phenomenon in human concept learning is the tendency, called selective attention, to consider as few features as possible to achieve acceptable classification accuracy. We've seen a simple case of this already predicted by the RR_{INF} model: single feature concepts were preferred to more complex concepts (Fig. 3(a)). However selective attention is particularly interesting in light of the implied tradeoff between performance and number of features attended. Medin, Altom, Edelson, and Freko (1982) demonstrated this balance by studying the category structure shown in Table 3. This structure affords two strategies: each of the first two features are 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 more accurate, 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 subjects relied primarily on the individually diagnostic dimensions in the initial stage of learning, and confirmed reliance on the correlated features later in learning. (Similar results have been discussed by Smith and Minda (1998).) Our RR_{INF} model explains most of the variance in human judgments in the final stage of learning, R^2=0.99 when b=6, and a respectable amount early in learning: R^2=0.70 when b=3. These fits don't depend on precise value of the parameter; see Fig. 4 for fits at several values. We have plotted the posterior complexity weights of the model for several values of parameter b in Fig. 5(a), and the feature weights in Fig. 5(b). When b is small the model relies on simple formulae along features 1 and 2, much as human learners do early in learning. The model switches, as b becomes larger, to rely on more complex, but more accurate, formulae, such as the perfectly predictive rule $\forall x\, \ell(x) \Leftrightarrow ((f_3(x)=1)\Rightarrow(f_4(x)=1)) \wedge ((f_4(x)=1)\Rightarrow(f_3(x)=1))$.

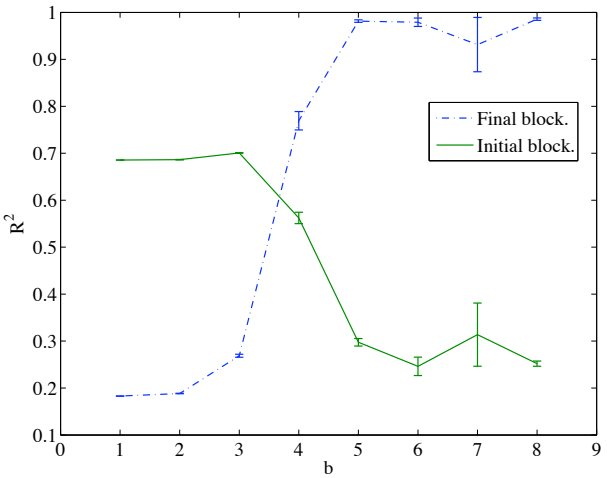

Figure 4. The fit (R^2) of RR_{INF} model predictions to human generalizations of McKinley & Nosofsky (1993) (see Table 3), both early and late in learning, for several different values of the parameter b. (Error bars represent standard error over five runs of the Metropolis algorithm used to approximate model predictions.)

These results suggest that grammar-based induction is a viable approach to the rational analysis of human concept learning. Elsewhere (Goodman et al., in press) we further

Table 3
The category structure of Medin et al. (1982), with initial and final block mean human classification responses of McKinley & Nosofsky (1993), and the predictions of the RR_{INF} model at parameter values b=3 and b=6.

Object	Feature Values	Human, initial block	Human, final block	RR_{INF}, b=3	RR_{INF}, b=6
A1	1111	0.64	0.96	0.96	1
A2	0111	0.64	0.93	0.59	0.99
A3	1100	0.66	1	0.96	1
A4	1000	0.55	0.96	0.60	0.99
B1	1010	0.57	0.02	0.41	0.01
B2	0010	0.43	0	0.04	0
B3	0101	0.46	0.05	0.41	0
B4	0001	0.34	0	0.04	0
T1	0000	0.46	0.66	0.14	0.64
T2	0011	0.41	0.64	0.14	0.63
T3	0100	0.52	0.64	0.51	0.64
T4	1011	0.5	0.66	0.51	0.64
T5	1110	0.73	0.36	0.86	0.36
T6	1101	0.59	0.36	0.86	0.36
T7	0110	0.39	0.27	0.49	0.36
T8	1001	0.46	0.3	0.5	0.36

investigate the ability of the Rational Rules model (based on the DNF grammar of Fig. 2(a)) to predict human generalization performance and consider in detail the relationship between the full posterior distribution and individual learners.

Role-governed Concepts

So far we have focussed on a concept language which can describe regularities among the features of an object. Is this feature-oriented model sufficient? Consider the following anecdote: A colleague's young daughter had been learning to eat with a fork. At about this time she was introduced to modeling clay, and discovered one of its fun properties: when you press clay to a piece of paper, the paper lifts with the clay. Upon seeing this she proclaimed "fork!" It is unlikely that in extending the concept "fork" to a lump of modeling clay she was finding common features with the spiky metal or plastic forks she had seen. However, it is clear that there is a commonality between the clay and those utensils: when pressed to an object, they cause the object to move with them. That is, they share a common *role* (in fact, a causal role—see Appendix A).

This anecdote reminds us that an object has important properties beyond its features—in particular, it has relationships with other objects. It also suggests that the defining property of some concepts may be that of filling a particular role in a relational regularity. Indeed, it is easy to think of such *role-governed* concepts: a key is something which opens a door, a predator is an animal which eats other animals, a mother is a female who has a child, a doctor is a person that heals illnesses, a poison is a substance that causes illness when ingested by an organism, and so forth. The critical commonality between these concepts is that describing them requires reference to a second object or entity; the contrast with simple feature-based concepts will become more clear in the formal representations below. The importance of relational roles in concept formation has been discussed recently by several authors. Markman and Stilwell (2001) introduced the term role-governed category and argued for the importance of this idea. Gentner and colleagues (Gentner & Kurtz, 2005; Asmuth & Gentner, 2005) have extensively considered relational information, and have found differences in the processing of feature-based and role-based categories. Goldstone, Medin, and Gentner (1991) and Jones and Love (2006) have shown that role information effects the perceived similarity of categories.

It is not difficult to imagine why role-governed concepts might be important. To begin, role-governed concepts are quite common. In an informal survey of high frequency words from the British National Corpus, Asmuth and Gentner (2005) found that half of the nouns had role-governed meaning. It seems that roles are also more salient than features, when they are available: children extend labels on the basis of functional role (Kemler-Nelson, 1995) or causal role (Gopnik & Sobel, 2000) in preference to perceptual features. For instance, in the study of Gopnik and Sobel (2000) children saw several blocks called "blickets" in the novel role of causing a box (the "blicket detector") to light when they were placed upon it. Children extended the term "blicket" to other blocks which lit the box, in preference to blocks with similar colors or shapes. However, despite this salience, children initially form feature-based meanings for many categories, such as "uncle" as a friendly man with a pipe, and only later learn the role-governed meaning (Keil & Batterman, 1984).

We have demonstrated above that grammar-based induction, using a concept language that expresses feature-based definitions, can predict effects found in concept learning that are often thought to be incompatible with definitions. It is interesting that many authors are more willing to consider

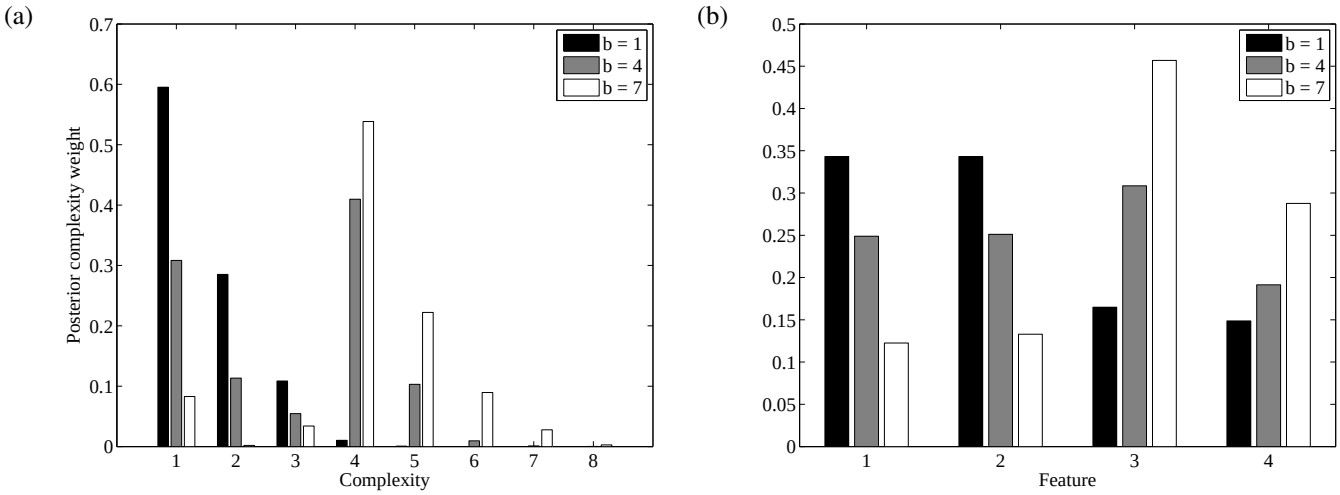

Figure 5. (a) Posterior complexity distribution on the category structure of Medin et al. (1982), see Table 3, for three values of the outlier parameter (b) Posterior feature weights.

role-governed concepts as definitional (Markman & Stilwell, 2001) or rule-like (Gentner & Kurtz, 2005), than they are for feature-based concepts. Perhaps then a concept language, like that developed above, may be especially useful for discussing role-governed concepts.

Representing Roles

Just as one of the prime virtues of compositionality in cognition is the ability to explain the productivity of thought, a virtue of grammar-based induction in cognitive modeling is a kind of "productivity of modeling": we can easily extend grammar-based models to incorporate new representational abilities. The hypothesis space is extended by adding additional symbols and production rules (with corresponding semantic operations). This extended hypothesis space is not a simple union of two sets of hypotheses, but a *systematic* mixture in which a wide variety of mixed representations exist. What's more, the inductive machinery is automatically adapted to this extended hypothesis space—providing a model of learning in the extended language. This extension incorporates the same principles of learning that were captured in the simpler model. Thus, if we have a model that predicts selective attention, for instance, in a very simple model of concepts, we will have a generalized form of selective attention in models extended to capture richer conceptual representation.

How can we extend the feature-based concept language, generated by the INF grammar, to capture relational roles? Consider the role-governed concept "key", which is an object that opens a lock. We clearly must introduce relation primitives, such as "opens", by a set of terminal symbols $r_1, \ldots, r_M$. With these symbols we intend to express "x opens y" by, for instance, $r_1(x, y)$; to do so we will need additional variables (such as y) to fill the other roles of the relation. With relation symbols and additional variables, and appropriate production rules, we could generate formulae like: $\forall x\ \ell(x) \Leftrightarrow (r_1(x, y)=1)$, but this isn't quite complete—*which* objects should y refer to? We need a quantifier to bind the additional variable. For instance, if there is *some* lock which the object must open, we might write $\forall x\ \ell(x) \Leftrightarrow (\exists y\ r_1(x, y)=1)$.

In Fig. 6 we have extended the INF grammar to simple role-governed concepts. The generative process is much as it was before. From the start symbol, S, we get $\forall x\ \ell(x) \Leftrightarrow (Qy\ I)$. The new quantifier symbol Q is replaced with either a universal or existential quantifier. The implication terms are generated as before, with two exceptions. First, each predicate term P can lead to a feature or a relation. Second, there are now two choices, x and y, for each variable term V. We choose new semantic operators, for the new productions, which give the conventional interpretations[8].

Let us consider the concepts which can be described in this extended language. The concept "key" might be expressed: $\forall x\ \text{Key}(x) \Leftrightarrow (\exists y\ (T \Rightarrow \text{Opens}(x, y))$. There is a closely related concept, "skeleton key", which opens *any lock*: $\forall x\ \text{Key}(x) \Leftrightarrow (\forall y\ (T \Rightarrow \text{Opens}(x, y))$[9]. Indeed, this formal language highlights the fact that any role-governed concept has a *quantification type*, $\forall$ or $\exists$, and each concept has a twin with the other type.

Though we have been speaking of role-governed and feature-based as though they were strictly different types of concept, most concepts which can be expressed in this language mix concepts and features. Take, for instance $\forall x\ \text{shallow}(x) \Leftrightarrow \forall y\ (\text{likes}(x, y) \Rightarrow \text{beautiful}(y))$, which may be translated "a shallow person is someone who only likes an-

[8] That is, '$R_j \rightarrow r_j(x, y)$=val' evaluates the j^{th} relation, '$Q \rightarrow \forall$' associates the standard universal quantifier to Q (and, *mutatis mutandis*, for '$Q \rightarrow \exists$'), and V is assigned independent variables over $\mathbf{E}$ for x and y. It would be more complicated, but perhaps useful, to allow outliers to the additional quantifier, as we did for the quantifier over labeled objects. This would, for instance, allow skeleton keys which only open *most* locks.

[9] We name relations and features in this discussion for clarity.

other if they are beautiful". It has been pointed out before that concepts may be best understood as lying along a feature–relation continuum (Gentner & Kurtz, 2005; Goldstone, Steyvers, & Rogosky, 2003). Nonetheless, there is a useful distinction between concepts which can be expressed without referring to an additional entity (formally, without an additional quantifier) and those which cannot. (Though note the concept "narcissist", a person who loves himself, which involves a relation but no additional entity.)

$$
\begin{array}{lcl}
S & \rightarrow & \forall x\, \ell(x) \Leftrightarrow (Qy\ I) \\
Q & \rightarrow & \forall \\
Q & \rightarrow & \exists \\
I & \rightarrow & (C \Rightarrow P) \wedge I \\
I & \rightarrow & T \\
C & \rightarrow & P \wedge C \\
C & \rightarrow & T \\
P & \rightarrow & F_i \\
P & \rightarrow & R_j \\
F_i & \rightarrow & f_i(V) = 1 \\
F_i & \rightarrow & f_i(V) = 0 \\
R_j & \rightarrow & r_j(V, V) = 1 \\
R_j & \rightarrow & r_j(V, V) = 0 \\
V & \rightarrow & x \\
V & \rightarrow & y
\end{array}
$$

Figure 6. The INF Grammar extended to role-governed concepts. (Indices $i \in \{1 \ldots N\}$ and $j \in \{1 \ldots M\}$, so there are M relation symbols R_i and etc.)

Learning Roles

The posterior for the feature-based $\mathrm{RR}_{\mathrm{INF}}$ model can be immediately extended to the new hypothesis space:

$$
P(F|\mathbf{W}, \ell(\mathbf{Obs})) \propto \left(\prod_{Y \in N} \beta(|\{Y \in F\}| + 1)\right) e^{-b|\{x \in \mathbf{Obs}|\neg(\ell(x) \Leftrightarrow (Qy\ D(x,y)))\}|}, \quad (9)
$$

where $D(x, y)$ is the set of implicational regularities, now amongst features and relations, and $Qy\ D(x, y)$ is evaluated with the appropriate quantifier. We now have a model of role-governed concept learning. Defining this model was made relatively easy by the properties of compositionality, but the value of such a model should not be underestimated: to the best of our knowledge this is the first model that has been suggested to describe human learning of role-governed concepts. (There have, however, been a number of Bayesian models that learn other interesting conceptual structure from relational information, for instance Kemp, Tenenbaum, Griffiths, Yamada, and Ueda (2006).)

The extended $\mathrm{RR}_{\mathrm{INF}}$ model is, unsurprisingly, able to learn the correct role-governed concept given a sufficient number observed labels (this limit-convergence is a standard property of Bayesian models). It is more interesting to examine the learning behavior in the case of an ill-defined role-governed concept. Just as a concept may have a number of characteristic features that rarely line up in the real world, there may be a collection of characteristic roles which contribute to the meaning of a role-governed concept. (This collection is much like Lakoff's idealized cognitive models (Lakoff, 1987); the 'entries' here are simpler yet more rigorously specified.) For instance, let us say that we see someone who is loved by all called a "good leader", and also someone who is respected by all called a "good leader". It is reasonable to think of these as two contributing roles, in which case we should expect that someone who is both loved and respected by all is an especially good "good leader". Let us see whether we get such a generalized prototype effect from the $\mathrm{RR}_{\mathrm{INF}}$ model.

Starting with our "good leader" example we construct a simple ill-defined role-governed concept, analogous to the concept of Medin and Schaffer (1978) considered above. In Table 4 we have given a category structure, for eight objects with one feature and two relations, that has no feature-based regularities and no simple role-based regularities. There are, however, several imperfect role-based regularities which apply to one or the other of the examples. Transfer object T4 is the prototype of category A in the sense that it fills all of these roles, though it is not a prototype by the obvious distance measure[10].

Table 5 shows formulae found by the extended $\mathrm{RR}_{\mathrm{INF}}$ model, together with their posterior weight. The highest weight contributors are the two imperfect role-based regularities ("someone who is loved by all" and "someone who is respected by all"), each correctly predicting 75% of labels. After these in weight comes a perfectly predictive, but more complex, role-governed formula ("someone who is respected by all those who don't love her"). Finally, there are a number of simple feature-based formulae, none of which predicts more than 50% of labels. The predicted generalization rates for each object (i.e. the posterior probability of labeling the object as an example of category A) are shown in Table 6. There is one particularly striking feature: transfer object T4 is enhanced, relative to both the other transfer objects and the examples of category A. Thus, the extended $\mathrm{RR}_{\mathrm{INF}}$ model exhibits a generalized prototype enhancement effect. This is a natural generalization of the well-known effect for feature-based concepts, but it is not a direct extension of similarity-based notions of prototype. The emergence of useful, and non-trivial, generalizations of known learning effects is a consequence of compositionality.

We can also explore the dynamics of learning for role-governed concepts. We would particularly like to know if the reliance on features relative to that on relations is expected to change over time. To investigate this we generated a world $\mathbf{W}$ at random[11], and assigned labels in accordance with the role-governed concept $\forall x\, \ell(x) \Leftrightarrow (\exists y\, r_1(x, y){=}1)$. As

[10] Prototypes are often treated as objects with smaller bit-distance (Hamming distance between feature vectors) to examples of the category than to its complement. If we extend this naively to bit-distance between both feature and relation vectors we find that the distance between A1 and T4 is larger than that between B1 and T4, so T4 is not a prototype of category A.

[11] Each random world had 15 objects, 5 features, and 2 relations. The binary features were generated at random with probability 0.5,

Table 4
An ill-defined role-governed category. The objects A1 and A2 are positive examples, B1 and B2 are negative examples, and T1-T4 are unlabeled transfer objects. It may be convenient to think of r_1 as "loved-by" and r_2 as "respected-by", and the concept label as "good leader".

Object	f_1	r_1:	A1	A2	B1	B2	T1	T2	T3	T4	r_2:	A1	A2	B1	B2	T1	T2	T3	T4
A1	0		1	1	1	1	1	1	1	1		0	1	0	0	0	0	1	0
A2	1		1	1	1	1	0	0	1	0		1	1	1	1	1	1	1	1
B1	0		0	1	1	1	0	1	0	1		1	0	1	1	0	1	1	1
B2	1		1	0	1	1	1	0	1	1		0	0	0	0	0	0	0	1
T1	0		0	0	0	0	1	0	0	0		0	0	1	0	1	0	1	0
T2	0		0	1	1	0	1	1	1	1		0	0	1	1	0	1	1	1
T3	1		1	1	1	1	1	1	1	1		1	1	0	0	1	0	1	1
T4	1		1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1

Table 5
The six highest posterior weight formulae from the extended RR_{INF} model applied to the world of Table 4.

weight	formula
0.178	$\forall x\, \ell(x) \Leftrightarrow (\forall y\, (T \Rightarrow (r_2(x,y)=1)))$
0.178	$\forall x\, \ell(x) \Leftrightarrow (\forall y\, (T \Rightarrow (r_1(x,y)=1)))$
0.049	$\forall x\, \ell(x) \Leftrightarrow (\forall y\, (((r_1(x,y)=0) \wedge T) \Rightarrow (r_2(x,y)=1)))$
0.016	$\forall x\, \ell(x) \Leftrightarrow (\forall y\, (T \Rightarrow (f_1(x)=1)))$
0.016	$\forall x\, \ell(x) \Leftrightarrow (\forall y\, (T \Rightarrow (f_1(y)=0)))$
0.016	$\forall x\, \ell(x) \Leftrightarrow (\exists y\, (T \Rightarrow (f_1(y)=1)))$

a measure of feature and relation weights we use the posterior expectation of the number of features or relations used in a formula; by averaging over many random worlds we get a qualitative prediction for typical learning. In Fig. 7 we have plotted these feature and relation weights against the number of observed labels. We see a clear feature-to-relation transition: early in learning features are of primary importance, as observations accumulate relations become more important, and eventually the correct role-governed concept is learned.

Recall children's shift for words like "uncle", from a feature-based interpretation to a role-based one. The qualitative feature-to-relation transition predicted by the extended RR_{INF} model suggests that this shift may in fact be the result of rational belief updating rather than limited resources or a domain general shift (e.g. from concrete to abstract understanding).

Discussion

The previous sections may be thought of as an extended example illustrating our view on what compositionality should mean in Bayesian rational analysis. The key features of our grammar-based induction approach to concept learning are the use of a concept language and a likelihood function compatible with the grammar of this language—the concept language lays the foundation for the virtues of compositionality, while compatibility gives the theory its semantic teeth. The RR_{INF} model is a case study of this approach.

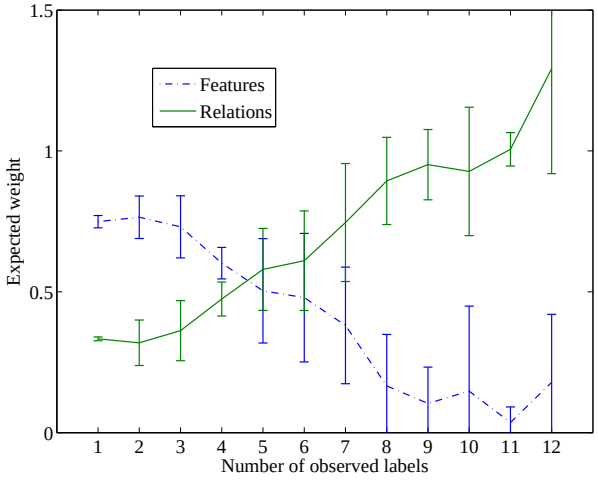

Figure 7. A feature-to-relation transition: the posterior weight of features and relations versus number of labels observed. Labels are consistent with the role-governed concept $\forall x\, \ell(x) \Leftrightarrow (\exists y\, r_1(x,y)=1)$. Error bars are standard deviation over 10 randomly generated worlds.

We compared the feature-based version of the RR_{INF} model to human data from two concept learning experiments, and found extremely good fits, comparable to the best existing models. This is particularly encouraging because the RR_{INF} model has only one free parameter—far fewer than other models. The RR_{INF} model has similarities with several well established models of concept learning. Like the RULEX model of Nosofsky et al. (1994), the RR_{INF} model learns a mixture of rule-like representations. However, while RR_{INF} is a computational-level rational model, the RULEX model is a process-level model. Thus the RR_{INF} model complements other efforts by providing a missing level of explanation to the rule-based approach to concept learning. RR_{INF}

the binary relations had probability 0.05 (providing sparse matrices).

Table 6
Generalization rates predicted by the extended RR_{INF} model for the category of Table 4.

Object:	A1	A2	B1	B2	T1	T2	T3	T4
Rate:	56%	59%	25%	23%	25%	26%	56%	81%

shares with other rational analyses of concept learning, such as Anderson's rational model (Anderson, 1990), its underlying Bayesian inductive principles—but existing rational models learn similarity-based representations that lack compositionality. One of the primary objections to similarity-based theories of concepts has been the lack of compositionality, which makes it difficult to to express rich relationships between concepts (Murphy & Medin, 1985) or to usefully combine concepts (Osherson & Smith, 1981). The grammar-based induction approach is well suited to address these concerns while still providing precise, and testable, computational models.

We leveraged compositionality to extend the RR_{INF} model to role-governed concepts by adding additional primitives to the concept language (relations and an additional quantified variable), expanding the set of productions to make use of these primitives, and specifying semantic operations for the new productions. Because the inductive semantics of RR_{INF} is compositional, the new pieces integrate naturally with the old, and the learning model continues to "makes sense" in the extended setting. This extension provides usefully expressive representations, and several interesting learning effects were noted, but future work is needed to empirically test and refine this form of the model.

The representations used in the RR_{INF} model were derived from mathematical logic. We drew not only on the syntax of logic, but also, in order to build a compatible likelihood, on the model-theoretic semantics of logic. These model-theoretic methods have already been used in several branches of cognitive science. In the psychology of reasoning, for instance, mental-model theory (Johnson-Laird, 1983) was inspired by model-theoretic notions of deductive truth. A bit farther afield, formal semantics, beginning with Montague (1973), has elaborated a detailed mathematical logic of natural language semantics. This *intensional* logic is also founded on mathematical model theory, though on the more modern notion of possible worlds. Formal semantics has developed a unique array of rich, and mathematically rigorous, representations relevant to cognition. These representations go far beyond the first-order logic used here, and are sure to be of interest as grammar-based induction is extended.

We have emphasized the notion of a concept language in which hypotheses are represented. This concept language, of course, is meant to be an internal mental language, which may not be verbalizable or even consciously accessible. Philosophical theories of mind based on an internal mental language ("mentalese") have a considerable history, via the language of thought hypothesis (Fodor, 1975). A key argument for these theories is that a language of thought might be sufficiently expressive for cognition because it inherits the virtues of compositionality; but an unresolved puzzle is what form the compositional semantics might take, in order to usefully connect mentalese to observations of the world. Our efforts in this chapter offer one solution: the language of thought can find its semantics in compositional prescriptions for induction—in particular, in likelihood and prior functions compatible with the syntax of mentalese. A moral can also be drawn in the other direction: the hypothesis spaces of Bayesian cognitive modeling, to the extent that they describe actual mental representations, can be seen as portions of the language of thought.

References

Ahn, W.-K., Kim, N., Lassaline, M. E., & Dennis, M. J. (2000). Casual status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 361–416.

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.

Asmuth, J., & Gentner, D. (2005). Context sensitivity of relational nouns. In *Proceedings of the twenty-seventh annual meeting of the cognitive science society* (pp. 163–168).

Boole, G. (1854). *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*. Walton and Maberly.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *TRENDS in Cognitive Sciences*, *10*, 335–344.

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

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006, July). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

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

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

Feldman, J. (2001). Bayesian contour integration. *Perception & Psychophysics*, *63*(7), 1171–1182.

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.

Frege, G. (1892). Uber Sinn und Bedeutung. *Zeitschrift fur Philosophie und philosophische Kritik*, *100*, 25–50.

Geisler, W. W., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, *5*(6), 508–510.

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. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependance of features in similarity judgments. *Cognitive Psychology*, *23*, 222–262.

Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory and Cognition*, *31*, 169–180.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (in press). A rational analysis of rule-based concept learning. *Cognitive Science*.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004, Jan). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *17*(5), 1205–1222.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Halpern, J. Y., & Pearl, J. (2001). Causes and explanations: A structural-model approach. part i: Causes. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence.*

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

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Jones, M., & Love, B. C. (2006). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology*.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior*, *23*, 221–236.

Kemler-Nelson, D. G. (1995). Principle-based inferences in young children's categorization: Revisiting the impact of function on the naming of artifacts. *Cognitive Development*, *10*, 347–380.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the twenty-first national conference on artificial intelligence (aaai-06).*

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

Lagnado, D. A., & Sloman, S. (2002). Learning causal structure. In *Proceedings of the Twenty-Fourth Annual Meeting of the Cognitive Science Society.* Erlbaum.

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

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.

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.

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

Milch, B., & Russell, S. (2006). General-purpose mcmc inference over relational structures. In *Proc. 22nd conference on uncertainty in artificial intelligence (uai)* (pp. 349–358).

Montague, R. (1973). The proper treatment of quantification in ordinary English. In J. Hintikka, J. M. E. Moravcisk, & P. Suppes (Eds.), *Approaches to natural language* (pp. 221–242). Dordrecht: D. Reidel.

Muggleton, S. (1997). Learning from positive data. In *Selected papers from the 6th international workshop on inductive logic programming* (p. 358-376). Springer-Verlag.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychol Rev*, *92*(3), 289–316.

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

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

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.

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

Rehder, B. (1999). A causal-model theory of categorization. In M. Hahn & S. C. Stones (Eds.), *21st annual conference of the cognitive science society* (pp. 595–600). Vancouver.

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.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453-489.

Tarski, A. (1956). The Concept of Truth in Formalized Languages. *Logic, Semantics, Metamathematics*, 152–278. (Originally "Der Wahrheitsbegriff in den formalisierten Sprachen", 1935.)

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

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

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

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Wittgenstein, L. (1921). *Tractatus logico philosophicus (routledge classics)*. Routledge.

Woodward, J. (2003). *Making things happen : a theory of causal explanation*. New York: Oxford University Press.

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

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301–308.

Appendix A
Causal Regularities

The INF grammar, which we have used to generate the concept language of the RR_{INF} model, represents the defining properties of a concept as a set of implicational regularities. As we indicated in the main text, it is intuitive to interpret these implications as causal relations. For instance the regularity $\text{Ripe}(x) \Rightarrow \text{Red}(x)$ for the concept "strawberry", might be interpreted as meaning that, if x is a strawberry, then x being ripe *causes* x to be red. This interpretation becomes especially interesting for the extension of RR_{INF} to role-governed concepts. Take the example of a "poison"; a poison is a substance that, when inside an organism, causes injury. We might write this as:

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

To see why the causal interpretation is important, consider the case of Socrates, who drank hemlock and died. If Socrates had been cured in the nick of time, should we conclude that hemlock is not a poison? This is the conclusion we must draw if we interpret the '$\Rightarrow$' as material implication. From the causal interpretation, however, hemlock may still be a poison: by intervening on the injury of socrates we supersede the causal regularity that would otherwise hold (Pearl, 2000). Indeed, interpreted causally this definition of "poison" is useful for crafting many interventions—e.g. if we'd like to injure someone, we may introduce a poison into them—and answering counterfactuals—e.g. if Socrates had been a rock, the poison hemlock would not have injured him.

Causal knowledge tells you how to make things happen—this is the thesis that has recently been argued in philosophy (Woodward, 2003) and cognitive science (Pearl, 2000; Gopnik et al., 2004; Lagnado & Sloman, 2002; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Thus, if we have a causal feature-based definition of a concept A, we know how to make things happen to the properties of an A, by manipulating other properties of that A (e.g. we know to make a strawberry red by ripening it). If we have a causal role-governed definition of concept A, we know how to make things happen to an A using another object, or vice versa (e.g. we can open a lock by using its key). (Notice that it is a very small step from here to the notion of using an object as a tool.) Formally, a regularity is causal if we know how to evaluate it under all possible sets of interventions.

We can formally extend the compositional semantics of the RR_{INF} model to this causal interpretation by specifying the likelihood function under intervention. We follow the *structural equation* approach of Halpern and Pearl (2001), with some embellishments to maintain compatibility of the likelihood with the grammar. Say that the intervention condition $\mathcal{I}$ is the set of interventions performed: each intervention is on one feature-object pair or relation-object-object triple. Let $\mathbf{W}_{\mathcal{I}}$ and $\ell(\mathbf{E})_{\mathcal{I}}$ be the observed features and labels under this intervention condition. We first extend the semantic information associated with the evaluation of a feature or relation: $f_i(x)$=val is now associated to a function from objects to a pair of truth values. The second truth value, the *intervention value*, indicates whether an intervention has been performed at this feature-object pair (i.e. whether the pair f_i, x is in $\mathcal{I}$). The same extension holds, *mutatis mutandis*, for relations. Next we alter the semantic operation associated to the (production which introduces an) implication $C \Rightarrow P$: when the intervention value of the P-term is True, the implication is evaluated as True (i.e. it is ignored), otherwise it is evaluated as usual. The remaining semantic operations are unchanged. Putting these together we may evaluate the extended likelihood $P(\mathbf{W}_{\mathcal{I}}, \ell(\mathbf{E})_{\mathcal{I}} | F, \mathcal{I})$ under a given intervention condition.

This extension is conceptually simple—ignore any implicational regularities on whose implicand an intervention is performed—the slight subtlety comes from the need to maintain compatibility between the likelihood and grammar. We achieved this by extending the semantic information given a *specific* intervention condition, and adjusting the semantic operations. Another, more general, option would be to extend the semantic types to include interventions as an argument (e.g. $e \rightarrow t$ would becomes $(e, i) \rightarrow t$). This is similar to the possible worlds technique used in intensional logic (Montague, 1973).

Appendix B
Sampling Assumptions

In the main text we derived an expression for the likelihood of a world with observed labels for all objects:

$$P(\mathbf{W}, \ell(\mathbf{E}) | F) \propto e^{-bO_\ell(\mathbf{E})}. \tag{10}$$

Where $O_\ell(\mathbf{E}) = |\{x \in \mathbf{E} | \neg(\ell(x) \Leftrightarrow D(x))\}|$ is the number of labeled examples which don't satisfy the "definition" part of F. If labels are observed for only a subset $\mathbf{Obs} \subseteq \mathbf{E}$ of the objects, then we must marginalize over the unobserved labels (we write $\overline{\mathbf{Obs}} = \mathbf{E} \setminus \mathbf{Obs}$ for the set of objects with unobserved labels):

$$\begin{aligned} P(\mathbf{W}, &\ell(\mathbf{Obs}), \mathbf{Obs} | F) \\ &= \sum_{\ell(\overline{\mathbf{Obs}})} P(\mathbf{W}, \ell(\mathbf{E}), \mathbf{Obs} | F) \\ &= \sum_{\ell(\overline{\mathbf{Obs}})} P(\mathbf{Obs} | F, \mathbf{W}, \ell(\mathbf{E})) P(\mathbf{W}, \ell(\mathbf{E}) | F) \\ &\propto \sum_{\ell(\overline{\mathbf{Obs}})} P(\mathbf{Obs} | F, \mathbf{W}, \ell(\mathbf{E})) e^{-bO_\ell(\mathbf{E})} \\ &= e^{-bO_\ell(\mathbf{Obs})} \sum_{\ell(\overline{\mathbf{Obs}})} P(\mathbf{Obs} | F, \mathbf{W}, \ell(\mathbf{E})) e^{-bO_\ell(\overline{\mathbf{Obs}})} \end{aligned} \tag{11}$$

Now we will need to make a *sampling assumption* (Tenenbaum & Griffiths, 2001) about how the objects with observed labels were chosen from among all the objects. Two standard choices are *weak sampling*, that **Obs** are chosen at random,

and *strong sampling*, that **Obs** are chosen explicitly to be positive (or negative) examples of the concept.

Weak Sampling

If **Obs** is chosen independent of $F, \mathbf{W}, \ell(\mathbf{E})$, for instance if it is chosen at random from among all objects, then:

$$\begin{aligned} P(\mathbf{W}, \ell(\mathbf{Obs}), \mathbf{Obs}|F) &\propto e^{-bO_\ell(\mathbf{Obs})} \sum_{\ell(\overline{\mathbf{Obs}})} e^{-bO_\ell(\overline{\mathbf{Obs}})} \\ &= e^{-bO_\ell(\mathbf{Obs})} \sum_{i=0}^{|\overline{\mathbf{Obs}}|} \binom{|\overline{\mathbf{Obs}}|}{i} e^{-bi} \\ &\propto e^{-bO_\ell(\mathbf{Obs})} \end{aligned} \tag{12}$$

Where we have used the fact that $\ell(x) \Leftrightarrow D(x)$ is true for exactly one of the two values of $\ell(x)$, and the sum is then independent of F.

Weak sampling is a reasonable assumption when there are both positive and negative examples, and no external reason to assume that examples are chosen to be exemplary. We have used this weak sampling likelihood for the examples in the main text, since it is both reasonable and simple in those settings.

Strong Sampling

If, on the other hand, only positive examples are observed (Tenenbaum, 1999), or in certain pedagogical situations (Xu & Tenenbaum, 2007), it is more reasonable to make a strong sampling assumption: positive (and negative) examples are chosen from among those objects which satisfy (respectively, don't satisfy) the concept at hand.

For simplicity let us assume that we have only positive examples, that is $\ell(x)=1$ for all $x \in \mathbf{Obs}$. We will assume that **Obs** is chosen at random from among objects which satisfy the concept, thus:

$$P(x \in \mathbf{Obs}|F, \mathbf{W}, \ell(\mathbf{E})) \propto \frac{1}{|\{y \in \mathbf{E}|\ell(y)=1\}|}. \tag{13}$$

(Note that we have allowed repeated samples, for simplicity.) From this we can derive an expression for the likelihood:

$$\begin{aligned} &P(\mathbf{W}, \ell(\mathbf{Obs}), \mathbf{Obs}|F) \\ &\quad \propto e^{-bO_\ell(\mathbf{Obs})} \sum_{\ell(\overline{\mathbf{Obs}})} \left(\prod_{x \in \mathbf{Obs}} P(x \in \mathbf{Obs}|F, \mathbf{W}, \ell(\mathbf{E})) \right) e^{-bO_\ell(\overline{\mathbf{Obs}})} \\ &\quad = e^{-bO_\ell(\mathbf{Obs})} \sum_{\ell(\overline{\mathbf{Obs}})} \left(\prod_{x \in \mathbf{Obs}} \frac{1}{|\{y \in \mathbf{E}|\ell(y)=1\}|} \right) e^{-bO_\ell(\overline{\mathbf{Obs}})} \\ &\quad = e^{-bO_\ell(\mathbf{Obs})} \sum_{\ell(\overline{\mathbf{Obs}})} \frac{e^{-bO_\ell(\overline{\mathbf{Obs}})}}{|\{y \in \mathbf{E}|\ell(y)=1\}|^{|\mathbf{Obs}|}} \end{aligned} \tag{14}$$

Learning from Positive Examples

Using the strong-sampling-based likelihood, Eq. 14, we may describe learning from only positive examples. Importantly, the denominator in Eq. 14 causes the smallest concept consistent with the labeled examples to be the most likely (this is the *size principle* of Tenenbaum (1999)); this leads the learner to select the most restrictive (hence informative) concept even when the evidence is equally consistent with broader concepts. This learning situation has been studied extensively in the context of feature-based categories (Tenenbaum, 1999; Xu & Tenenbaum, 2007; Muggleton, 1997), but it is particularly interesting in the setting of role-governed concepts. Indeed, note that any set of positive examples that is compatible with a universally quantified regularity is also compatible with an existential regularity (but not vice versa: the existential form is less restrictive than the universal form). For instance, if we see several "poisons" injure several different people, should we infer that "poison" is governed by a universal or existential quantifier (i.e. that there is *someone* who every poison injures, or that a poison injures *anyone*)? Under the weak sampling likelihood we must wait until these are distinguished by negative evidence (e.g. a non-poison which injures a few people), but with the strong sampling likelihood the more restrictive hypothesis—the universal regularity—is favored, all other things being equal.