# Explanatory Reasoning for Inductive Confidence

**David Landy (dlandy@richmond.edu)**
Department of Psychology, M03B Richmond Hall
University of Richmond, VA, 23173 USA

**John E. Hummel (jehummel@illinois.edu)**
Department of Psychology, 603 E. Daniel St.
Champaign, IL 61820 USA

**Abstract**

We present Explanatory Reasoning for Inductive Confidence (ERIC), a computational model of explanation generation and evaluation. ERIC combines analogical hypothesis generation and justification with normative probabilistic theory over statement confidences. It successfully captures a broad range of empirical phenomena, and represents a promising approach toward the application of explanatory knowledge in new situations.

**Keywords:** induction; analogy; probabilistic reasoning

## Introduction

We are constantly making guesses. When we come across something new, we know about it in part from its relations to other things and we attribute to the novel the properties of the familiar. For instance, when Apple announced the iPad, technology reporters alternately compared it to tablet PCs, which are similar in size and function, and the iPhone, which is similar in appearance and operating system. In each case, the game was to predict the features of the new object on the basis of the old ones.

Property inductions of this kind—extending known properties of one category to other categories—have been heavily studied in experimental psychology (see Heit, 2000, for a review). Such inductions seem to take advantage of taxonomic knowledge about category structures as well as specific knowledge about particular categories (Shafto, Kemp, Bonawitz, Coley, & Tenenbaum, 2008).

One intuition, pursued here, is that people make inductions by adapting explanations for known properties to novel categories. People are habitual generators of explanations: Scientists explain natural phenomena; engineers explain why structures will or will not support various loads; mathematicians explain why a formal property does or does not hold of a particular situation or object; and everyone routinely explains much more mundane things such as why the doorbell rang, why we smell gas in the kitchen and why a child has a fever. Explanations serve many cognitive functions, but perhaps none is more important than their ability to support inductive inferences: A person who can explain a novel observation can have much greater confidence in their inferences about the circumstances under which that observation is likely to be repeated than a person who cannot explain it—which is why, for example, your auto mechanic is better than you are at knowing whether that strange noise you car is making is likely to be dangerous.

In order to apply explanations of past experiences to novel situations, a cognitive architecture must solve several problems. First, it must be able to generate and retain explanations in the first place. Second, it must have a way to generate novel hypotheses about a current situation from its beliefs about past circumstances. Finally, it must be able to distinguish when a novel explanation is plausible in the current situation, and when it is not.

Bayesian models, and particularly hierarchical Bayesian models, are adept at the last of these goals. For example, the model of Kemp and Tennenbaum (2009) carves known situations into disjoint domains, and applies to novel situations the domain assumptions that appear most appropriate. However, human reasoners also adapt explanation patterns across multiple, dissimilar domains (Medin, Coley, Storms, & Hayes, 2003). Although such cross-domain reasoning is the sine qua non of analogical approaches to reasoning (Falkenhainer, Forbus, & Gentner, 1989; Hummel & Holyoak, 1997), models of analogy generally provide no basis for generating probabilistic estimates of confidence in their inferences.

In this paper, we present the model ERIC, Explanatory Reasoning for Inductive Confidence (see also Landy & Hummel, 2009). ERIC uses a combination of analogical and probabilistic reasoning to (a) generate explanations for newly learned facts, (b) evaluate the plausibility of those explanations in light of its existing knowledge, (c) use those explanations to update its confidence in its existing knowledge and (d) make judgments about the plausibility of new inferences. The resulting model accounts for a large body of empirical findings from the literature on inductive confidence (e.g., Heit, 2000; Shafto et al., 2008).

A central tenet of the model is that the mind uses analogy to adapt old explanations to new situations and then uses those new explanations both to determine its confidence in the new observation and to update its confidence in its existing knowledge—both existing basic facts and existing explanations. The knowledge updated includes both the source analogs (i.e., the old explanations used to generate the new ones) and the analogies themselves (i.e., the mappings from the old [source] explanations to the new [target] explanations). As a result, if an analogy results in a good explanation, then the model becomes more convinced both that the source was true and that the analogy was good.

A second central tenet is that the mind generates these explanations permissively and habitually: Presented with any new "fact" or observation, the mind will generate as many potential explanations of that fact as possible and assign a likelihood or confidence value to each; in turn, these values are used to update its confidence in the very facts that participated in the explanations themselves.

Some models of induction (e.g., Kemp & Tennenbaum, 2009) explicitly carve knowledge into separate domains, and assume that categorically different processes apply to situations attributed to those domains (e.g., reasoning in one way about ontological knowledge and in a different way about geographical knowledge). A third tenet of the model is that knowledge, including knowledge about generating processes, is applied to relevant situations regardless of domain. That is, the processes underlying explanation and confidence estimation are the same across and within all areas of knowledge: Any differences between, say, ontological knowledge and other knowledge domains (e.g., geographical location, diet or behavioral traits) emerge as a natural consequence of the relationship between individual sets of facts, and not through an explicit and absolute categorization into domain.

Finally, in line with other integrative general knowledge models, the goal of ERIC is not to be entirely formally consistent (Wang, 2009). For instance, it will not necessarily be the case that $a \wedge \sim a$ is guaranteed to be false.

## Property Induction

We report a collection of simulations using ERIC to perform a property induction task (Osherson, Smith, Wilkie, López, & Shafir, 1990; Rips, 1975). In this task, a subject (or ERIC) is given a *premise*, which is assumed to be true (e.g., "robins get disease *d*"), based upon which they are asked to estimate the likelihood of a *conclusion* (e.g., "birds get *d*"). The dependent measure of interest is the estimated likelihood of the conclusion as a function of the relation between the major term in the premise (here, "robins") and that in the conclusion ("birds"), and of the relation between these categories and the property induced ("disease *d*").

# ERIC

## Overview

ERIC is based on the following assumptions about the nature of the property induction task:

1. A person enters the laboratory with knowledge (facts, explanations, theories) believed in with varying degrees of confidence.
2. Faced with the premise, the subject tries to explain it by building a fairly large set of potential explanations by analogy to known cases.
3. Each explanation is assigned an inductive confidence that combines confidence in the knowledge involved in the explanation and confidence in the generating analogies.
4. These explanations are added (provisionally) to knowledge, and the confidence of existing statements is updated using Bayesian inference.
5. Faced with a conclusion, the subject repeats process of explanation and confidence updating.
6. Confidence in the conclusion is high to the degree that the explanations are strong.

As input, ERIC takes an explanandum—either a premise or a conclusion. As output, it generates potential explanations, each with an assigned confidence, and an estimate of the confidence in the explanandum itself. Applied to property induction, the mechanism operates in two stages: First, ERIC explains the premise(s) and any knowledge gleaned from those explanations is added to the knowledge base. Next, it explains the conclusion using that augmented knowledge. The result of these processes is an estimate of the likelihood that the conclusion is true.

## Knowledge Representation

All of ERIC's knowledge is represented in standard propositional notation, augmented to capture the logical and causal relations that link propositions into explanations. Atoms are of the form *f(a), g(a, b, c)*, and so on. Connectives $\wedge$, $\vee$, and $\sim$ are used in their usual sense to mean *and*, *or*, and *not*.

Two less universal connectives provide a language for representing explanations and analogical mappings. The connective $\Rightarrow$ denotes an explanatory or causal relationship. For example, $q \Rightarrow r$ should be read as "*q* (if true) would tend to explain (cause) *r*." In contrast to some prior models (e.g., Falkenhainer et al., 1989; Hummel & Holyoak, 1997), causal connections are treated as special types, and not as generic two-place predicates (see also Hummel & Landy, 2009). Syntactically, they are equivalent in ERIC to a material conditional.

The second novel connective is the *mapping relation*, $q \leftrightarrows r$, which asserts that *q* and *r* map to each other in some analogy, and provides ERIC's initial estimate that *q* and *r* might map to each other in some future analogy. Mapping connections have learned confidences.

**Confidence** Each statement, *q*, is assigned a *confidence* value between 0 and 1, which is intended to work much like an intuitive probability that the statement is true. Indeed, we will refer to the confidence as "the probability of *q*," or p(*q*).

Statements in the initial knowledge set have a preset initial confidence. Regular property statements and cause relations (e.g., $q \Rightarrow r$) that do not appear in the initial knowledge have a confidence set to arbitrary low values (0.1 and 0.001).

**Explanations** An explanation is a recursive binary modal structure, with the pattern *E(explanation; explanandum)*, where the explanandum is a statement, and the explanation is a set of statements. They have the form of a modus ponens: Some set of (possibly recursively justified) causes

and an explanatory connective statement justify the effects. For instance, the explanation:

$$E_1(p,q,E_2(r,r \Rightarrow q;q),p \wedge q \Rightarrow s;s).$$

asserts that "*p*, *q* (where *q* is explained by *r*), and [*p* and *q* cause *s*] jointly cause *s*."

An explanation differs from a causal connective in several ways. First, a causal connective is purely dispositional, while an explanation asserts that *in fact*, the explanation explains the explanandum. An explanation thus encodes a derivation pattern, rather than a potential relationship. Further, an explanation carries its own internal semantics; it denotes a possible state of affairs.

**Knowledge base** ERIC's knowledge consists of three major classes of statements: simple property statements, such as *eats(Robin, Worm)*; simple explanations, such as generic taxonomic explanations of the form *isa(A, B) ^ x(B)* ⇒ *x(A)*; and taxonomic assertions, of the form *isa(Robin, Bird)*. It is worth noting here that taxonomic assertions are simply property statements, and not a special part of the model mechanism.

## Justification

ERIC revises its beliefs (e.g., explanations) using two kinds of justification: analogical and explanatory. For either, the effect of a justification, *j*, on an explanandum, *i*, is to update the probability of *i* according to a probabilistic-OR rule:

$$p(s) \leftarrow p(j) + (1 - p(j))p(s) \qquad (1)$$

Intuitively, (1) can be read as meaning that if the justification, *j*, is correct, then the assertion, *s*, it justifies must be correct, but if it is not, then *s* might still be correct with (base rate) probability *p*(*s*).

The initial confidence of an explanation is simply the probability that all the statements in the explanation are true:

$$p(j) = \prod_{e \in E} e \qquad (2)$$

**Analogical Justification** Intuitively, an analogy, $r \leftrightarrows q$, justifies *q* to the extent that the source analog (*r*) is true, and the mapping is reliable. Thus,

$$p(j) = p(r)p(r \leftrightarrows q) \qquad (3)$$

The target of an analogical justification is always a causal statement. These are updated by applying the justification to the cause statement via equation (1), just as with explanatory justification.

## Explanation Generation

When a new explanandum, *q*, is presented to ERIC, two steps are recursively applied to generate new explanations of *q*. First, each fact in the current knowledge base that shares any literals with *q* is postulated as a possible explanation for *q*. For example if *q* = *g(a)* and if *f(a)* is known, then one explanation postulated will be *f(a)⇒g(a)*. Confidence in this shallow explanation will initially be set to a very low value. Second, existing explanations (including those inside explanations) are expanded and justified by analogy to other explanations in knowledge.

Any potentially useful analogical mapping, e.g., $(a \Rightarrow b) \leftrightarrows (c \Rightarrow d)$, is computed by mapping the elements of *a,b* onto those of *c,d* using Holyoak and Thagard's (1989) ACME mapping algorithm. ACME's mapping strengths range between 0 and 1, and so translate conveniently into confidences. ACME combines structural isomorphism and semantic relationships. In ERIC, these semantic relationships are computed directly from the knowledge base (see Projectable Literals, below).

The best match produced by ACME is used as the basis for an analogy. This approach has two effects. First, the explanatory relation is justified by the analogical statement, using (3). Second, statements appearing in the analog but not in the current explanation are imported.

These two processes are applied to each explanation in the current set a fixed number of times (three in the current simulations). Each explanation in the final set justifies the conclusion; the result is the confidence in the conclusion.

**Projectable Literals** Analogical similarity integrates structural overlap and semantic relationships (Taylor & Hummel, 2009). That is, structural relations being equal, ERIC prefers analogies about identical or similar terms to comparisons among distantly related items.

The semantic similarity—more accurately, *projectability* (Simmons & Estes, 2008; Sloutsky, Kaminski, & Heckler, 2005)—of *a* onto *b*, $p_{ab}$, can come from either of two sources. If two terms have been related by past explanatory analogies, then the projectability is stored in the form of a mapping statement. The projectability of two previously unrelated terms is calculated from ERIC's knowledge:

$$p_{ab} = e^{-d_{ab}} \qquad (4)$$

where

$$d_{ab} = \alpha s_a + \beta s_b - \gamma s_{ab} - \delta m_{ab}$$

α, β, γ and δ are free parameters (15/40, 2/40, 1/40, and 17/40, respectively). Here $s_a$ is the summed confidence in sentences in which *a* appears; $s_b$ and $s_{ab}$ are defined analogously. Intuitively, *a* is projectable onto *b* to the extent that they appear in similar relational roles in LTM ($\gamma s_{ab}$) or to the extent that *b* is a kind of *a* ($\delta m_{ab}$) and to the extent that *a* does not appear in roles in which *b* does not and vice-versa ($\alpha s_a + \beta s_b$). If a mapping connection exists between *a* and *b* then ERIC uses the mapping strength as $p_{ab}$: ERIC learns that facts about *a* generally apply to *b*.

The differential applicability of known explanations to novel situations constructs a kind of soft domain separation. Although any knowledge can be applied to a new situation in principle, close knowledge will be applied with far more

confidence. As a result, cross domain analogies have most effect in the absence of other good explanations. This differential applicability of old explanation replaces the construction of explicit domains of explanations (Kemp & Tenenbaum, 2009) used in other approaches, and in general may implement generic symbolic rules (Gentner & Medina, 1998; Sun, 2006).

### Knowledge Revision

In property induction, a certain number of premises (collectively $\pi$) are followed by a conclusion statement, *c*. In calculating confidence in a conclusion, ERIC first generates explanations of the premises. It uses these to update its knowledge base. If $\pi$ consists of multiple premises, then each individual premise is explained; the full set of explanations is the set of all possible combinations of explanations for individual premises.

Learning a new premise means adding it to the knowledge base with confidence=1. Learning a new fact should inform the learner to the degree that the fact was surprising; it should increase confidence in things that would explain that fact. Both intuitions can be captured by Bayes law, if we are careful about where our terms come from.

$$p(e|\pi) = \frac{p(\pi|e)p(e)}{p(\pi)} \qquad (5)$$

The prior probability, $p(\pi)$, is the confidence in $\pi$ resulting from the explanation process. Intuitively, $p(\pi|e)$ is the confidence we would have in $\pi$ if some particular fact $e$ were known with certainty. This value can be found by repeating the process of justifying $\pi$, setting the confidence of $e$ to 1 for each fact that appears in explanations for $\pi$, including analogy sources, and assertions of analogical validity. It should be clear that the use of this law is not normative here, since the values are not strictly probabilities. However, the law forms one good way to incorporate evidence into belief systems. ERIC postulates that people use something like this kind of inference.

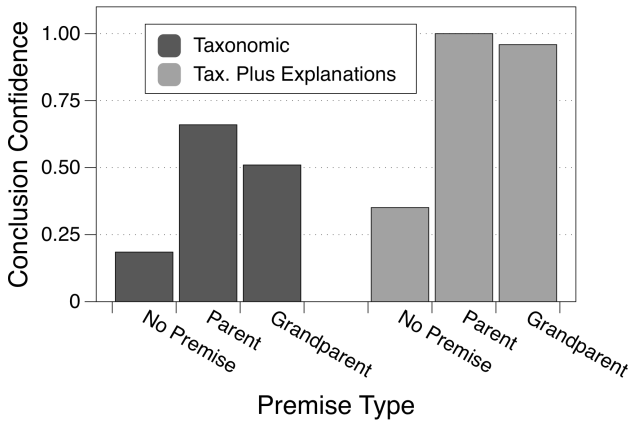


Figure 1: The strengths of induction of a property from one category to a related category. In general, ERIC makes stronger inductions from more closely related categories.

In property induction, ERIC uses the knowledge base that results from explaining the premise to explain the conclusion. Since each explanation justifies the conclusion, confidence in the conclusion results from the application of (1) once for each explanation.

In principle, the resulting confidence values could be matched directly to human probability estimates. In practice, current limitations of the model (especially its extremely impoverished "knowledge") make such point-by-point comparison uninformative, so our evaluation of the model will focus on the relative rankings of sets of explanations.

## Simulations and Results

ERIC predicts that inductions, and even patterns of inductions, will be strongly dependent on knowledge, and particularly on contextually relevant knowledge. For this reason, conclusions about the predictions of ERIC must be made relative to some particular set of knowledge.

### Taxonomic Simulations

Taxonomic relationships have received much attention in the literature on category inductions; we decided to explore two knowledge bases built largely around taxonomic knowledge. In the first, a taxonomic structure of "animals" was constructed with *isa* statements, including two mammals, six birds, and two reptiles. Animals were, in turn, defined by membership to the superordinate "living things." One general taxonomic explanation was included, over elements that did not appear in any other statements. The pattern of this explanation was: $isa(x,y) \wedge f(y) \Rightarrow f(x)$.

The second knowledge base included all of these taxonomic facts, but also included a fairly arbitrary set of about 200 facts, including property statements and casual explanations, both taxonomic and not taxonomic. This knowledge base tests the generality of the conclusions across a noisier knowledge base.[1]

Since inductions from a category to its subset are explanations, like all explanations, they are not certain. Furthermore, close ancestors generally provide more support than more distant ancestors. Figure 1 compares ERIC's inductions from immediate superordinates of a category ("parents"), and from the superordinates' superordinates ("grandparents" see Figure 1). Thus, a premise "birds have x" provides more support to the conclusion "robins have x" than does "animals have x". This pattern matches the empirically discovered category inclusion fallacy (Heit, 2000; Sloman, 1998). Figure 1 shows that this same pattern appears with the richer knowledge base, as well.

Within taxonomic categories at the same level (e.g., the species level), taxonomic proximity again can vary. Figure 2 shows the results of simulations varying the taxonomic proximity, and also the number of premises in the induction

[1] The full contents of all knowledge bases described here can be found online at http://www.richmond.edu/~dlandy/cogsci10/.

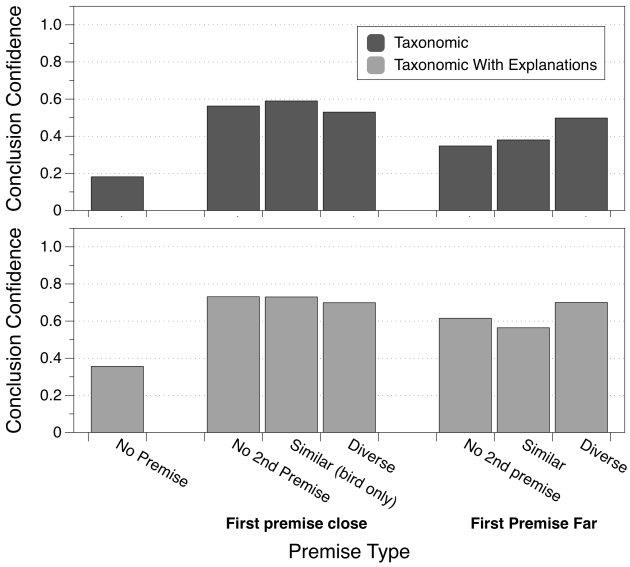


Figure 2: The strengths of induction of a property from zero, one or two categories to others at the same level.

(that is, the number of species of which the property was asserted). In the absence of knowledge, ERIC generally predicts that inductions tend to be stronger between categories that are closely related (see Figure 2). More premises tend to make inductions stronger; moreover, ERIC shows a general diversity effect: when multiple premises come from unrelated categories, that tends to increase inductions more than when they have a common superordinate. This is true in general because two close premises will tend to be best explained by explanations in terms of their common superordinate, while diverse premises are likely to be explained in terms of distant superordinates. This pattern is complicated, however, by an interaction between the diversity of the premises and their similarity to the conclusion. If one premise category is close to the conclusion category, a single premise category already generalizes fairly strongly, because most explanations for the premise are highly mappable into the conclusion; adding a second close premise improves the induction very slightly or not at all. However, if the second premise is from a very different category (making the premises more diverse), then ERIC's explanations are likely to be less finely tuned to the conclusion category, and confidence decreases slightly. This pattern again matches empirical literature (Osherson et al., 1990; Sloman, 1993).

## Typicality

To explore how ERIC uses typicality information, we augmented the taxonomic knowledge base with two kinds of information. Both involved four members of a common animal family ("birds"), with four features. The *typical* member had the same four features. The *typical plus* member had the same four features plus an additional two not shared by other members. The *typical minus* had only two of the features, and no additional features. The final *atypical* member had two shared features, and two unique features. A second knowledge base had the same exemplars and features, plus explanations for each feature.

ERIC computed confidence in the induction of a blank property from each premise bird to the conclusion bird. Figure 3 displays the results. Generally, as with people (Heit, 2000), increased typicality led to higher inductive confidence. One interesting exception to this pattern was that in the features only case, inductions were slightly stronger from the premise category with relatively few features than from the premise category with many typical categories. This is because this "unknown" category was exceptionally projectable, due to having very few features. When more explanations were available, the relatively high number of good potential explanations for the typical category dominated, leading to strong inductions.

## Causal Knowledge

Because ERIC extends its knowledge based on the overall analogical quality, the predicate attributed to a premise and conclusion category can also strongly impact induction, if facts involving that premise or a related one are part of prior knowledge. A predicate similar to those that appear as part of good, projectable explanations about similar categories sets will generally form strong inductions; projectable predicates known to apply to very different creatures, or those about which little is known, tend to project less well.

We illustrated this property by creating knowledge corresponding to the taxonomic and predatory structures explored by Shafto et al (2008). For a set of seven animals, predation and taxonomic facts were encoded in memory. Two generic explanations involved a "disease" spread by predation, and an "organ" shared by animals sharing a taxonomic category. Inductions were generated for each creature regarding a different "disease" and "bone."

As illustrated in Figure 4, inductions on the bone graded taxonomically. Premises involving species with the same parent (distance 0) generalized more strongly than more distantly related species. Diseases also showed a taxonomic structure, but less strongly than bones did. Furthermore, the disease was strongly affected by ecological relationships, generating an asymmetry such that predators were judged more likely to get diseases carried by their prey than were prey whose predator was known to catch the disease.

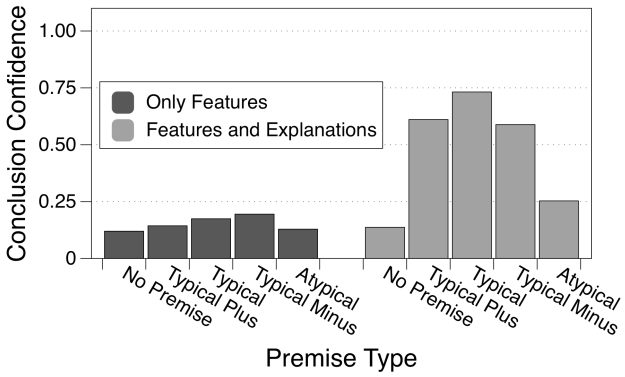


Figure 3: ERIC's predictions of induction strength, varying the typicality of the premise category.

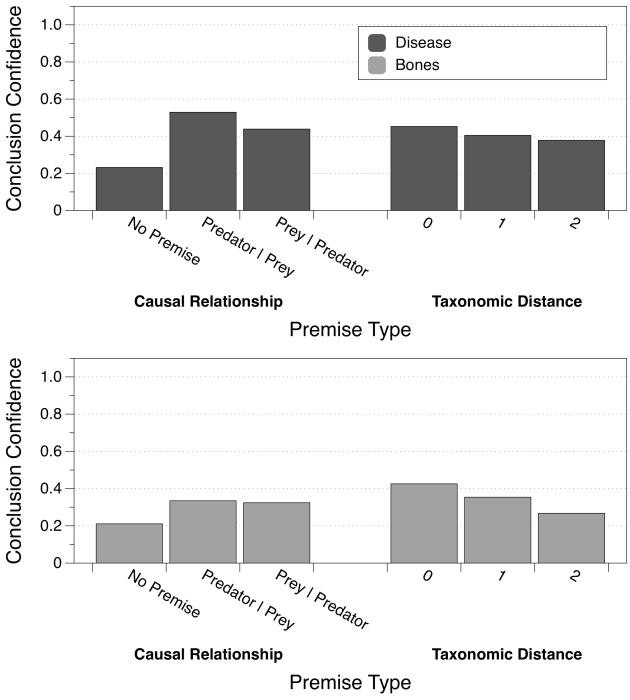


Figure 4: Dependency of inductive strength on both property and category relationships.

The latter still formed a strong induction in the disease case, because a prey carrying a disease made a good explanation for why a predator would have it; this explanation was thus well-supported during the premise explanation phase of ERIC's reasoning process. These patterns are quite similar to human judgments (Shafto et al., 2008), and demonstrate ERIC's ability to adjust the application of "rules" to different areas of knowledge.

Both properties showed taxonomic degradation. This is because both kinds of knowledge are in the system, and so both affect, to some degree, the same judgments. The model predicts that people will also blend different theories and domains of knowledge when making inductions.

## Conclusions

ERIC combines deductive probabilistic inference with inductive analogical inference to generate and evaluate the likelihood of explanations, the propositions they comprise and the observations they explain. The resulting model, still in an early stage of development, successfully predicts and explains a wide range of phenomena in the property induction literature. Much work remains to be done (e.g., representing probabilities more realistically, allowing explanations to decrease as well as increase confidence, and making the generation of analogical explanations psychologically plausible rather than computationally exhaustive, among many others), but at this point ERIC seems a promising way to overcome the limitations of purely analogical, and purely Bayesian approaches to explanation generation and evaluation.

## Acknowledgments

This research was funded by AFOSR grant # FA9550-07-1-0147. Thanks to Eric Taylor, Brian Ross, and Derek Devnich for thoughts and comments during the development of ERIC.

## References

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: algorithm and examples, *Artificial Intelligence, 41*, 1-63.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*(2-3), 263–297.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, *7*(4), 569–592.

Holyoak, K. J., & Thagard, P. (1989). Analogical Mapping by Constraint Satisfaction. *Cognitive Science*, *13*, 295-355.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.

Hummel, J. E., & Landy, D. (2009). From analogy to explanation: Relaxing the 1:1 mapping constraint...Very carefully. In *New Frontiers in Analogy Research: Proceedings of the Second International Conference on Analogy*. Sofia, Bulgaria.

Kemp, C., & Tenenbaum, J. B. (2009). Structured Statistical Models of Inductive Reasoning. *Psychological Review*, *116*(1), 20-58.

Landy, D., & Hummel, J. E. (2009). Explanatory reasoning for inductive confidence. In *New Frontiers in Analogy Research: Proceedings of the Second International Conference on Analogy*. Sofia, Bulgaria.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185-200.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, *14*, 665-681.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*(2), 175–192.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231-280.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33.

Sun, R. (2006). Accounting for a variety of reasoning data within a cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, *18*(2), 169-191.

Wang, P. (2009). Formalization of Evidence: A Comparative Study. *Journal of Artificial General Intelligence*, *1*, 25–53.