

Toward Robust Real-World Inference: A New Perspective on Explanation-Based Learning

Gerald DeJong

Computer Science Department,
University of Illinois at Urbana
mrebl@uiuc.edu

Abstract. Over the last twenty years AI has undergone a sea change. The once-dominant paradigm of logical inference over symbolic knowledge representations has largely been supplanted by statistical methods. The statistical paradigm affords a robustness in the real-world that has eluded symbolic logic. But statistics sacrifices much in expressiveness and inferential richness, which is achieved by first-order logic through the nonlinear interaction and combinatorial interplay among quantified component sentences. We present a new form of Explanation Based Learning in which inference results from two forms of evidence: analytic (support via sound derivation from first-order representations of an expert's conceptualization of a domain) and empirical (corroboration derived from consistency with real-world observations). A simple algorithm provides a first illustration of the approach. Some important properties are proven including tractability and robustness with respect to the real world.

1 Introduction

The classification problem has become so emblematic of supervised machine learning that the two terms are sometimes used synonymously. Yet one can certainly imagine object classification as an inference rather than a learning task.

For example, suppose we know that birds fly, parakeets are birds, grackles are birds, etc. and that a particular object, Tweety, possesses the properties of being yellow, talkative, a parakeet, feathered, and so on. We can then infer that Tweety is to be classified into the set of flying things.

More abstractly we can realize classification using automated reasoning as follows. Let x refer to a world object whose representation X specifies a conjunction of input features. Let $C(\cdot)$ be the predicate denoting membership in some class L of interest, Π be the axioms capturing our prior knowledge, and \vdash our inference relationship of choice. Then

$$X \cup \Pi \vdash C(x) \text{ implements } x \in L \quad (1)$$

Unfortunately, the automated reasoning of conventional logic has proven too brittle to be effective in the real world. Notoriously, the above would equally conclude that the neighbor's parakeet who was just killed by their cat can fly, as can Bugsy the Mafioso grackle whose feet are cast in a block of cement.

Consider an expression ϕ that is derivable through our inference relation \vdash from a set of statements Δ :

$$\Delta \vdash \phi \quad (2)$$

We would like our domain theories to exhibit real-world robustness; inferring a sentence should provide some guarantee that it holds in the real world. This is expressly not the case in conventional logic where formal guarantees about ϕ 's apply only to the microworld defined by Δ . If a conventional inferential system is robust in this sense, it is due exclusively to the human implementor's care and cleverness in crafting Δ and is beyond the scope of the formal inferential system.

Explanation-Based Learning (EBL) can be viewed as a path around this brittleness. We suggest a kind of paradigm shift to the interpretation of EBL. Instead of seeing EBL as bringing prior knowledge to the learning task, we explore EBL as bringing learning to the inference task; it formalizes robustness in deductive inference.

Suppose we could achieve real-world robustness within the scope of some formalism. What robustness properties would be desirable and what would be possible? We might at first hope that the ϕ 's of (2) be guaranteed to hold in the real world (i.e., that derivability from some appropriate set of axioms suffices to trust that a statement is true in the real world). We believe this to be impossible. Instead, we propose a slightly weaker guarantee in which the inferred ϕ 's hold in the real world only with high confidence in a sense that follows from statistical learning.

This new EBL is based on four tenets.

1. The human expert's domain representations reflect a deep appreciation of world's subtleties that the computer learner is unlikely to equal and should not question. In EBL, training examples guide the *interpretation* of the human-supplied domain theory with respect to a particular task, and not its refinement or improvement.
2. Robustness cannot be achieved generally but only with respect to a particular domain task. A formal connection to this real world task is a requisite for robust inference. Statistics supports robustness guarantees but conventional logic does not.
3. The absolute inferential confidence afforded by conventional logic, while seductive, is often unrealistic in the real world and can be a major source of brittleness. Human inference is generally weaker than this as is statistical inference.
4. Complex domains require an appropriately expressive language for the expert to articulate his/her understanding. Propositional models, relational models, description logics, etc. are sufficient for some domains, but first order representation with the combinatorial interaction that unification affords can provide a greater conceptual richness. This is realized in logic but is not easily incorporated into statistics.

Explanation-Based Learning requires an expert-supplied domain theory, Δ . We take this to be a set of first order expressions. But the intent is to allow the expert's unencumbered conceptualization of the domain. As such, we expressly do not require the theory to possess the difficult-to-achieve global properties of consistency or robustness. A second input is a set of training observations, \mathbf{Z} . These provide a direct connection to the real world. The result of EBL is a specially tailored logic domain theory, Δ' which is likely to be robust in the task illustrated by the training examples. Thus,

$$\Delta \cup \mathbf{Z} \xrightarrow[\text{EBL}]{} \Delta' \quad (3)$$

In EBL the prior domain theory Δ is deemed to be correct but *not* believed. Unlike ILP or theory revision, EBL does not attempt to improve or augment the expert-supplied domain theory. Rather, Δ is viewed as the most comprehensive and maximally useful general domain description that the human expert can provide. Like all first-order theories, it is subject to the qualification problem:

Most universally quantified sentences will have to include an infinite number of qualifications if they are to be interpreted as accurate statements about the world. [9]

Flaws cannot be avoided, so encountering them cannot be taken as evidence against the worth of the human's statements. Revising Δ may simply compromise the human's expression of his/her expertise resulting in a worse theory. But by the same token, neither can Δ be believed. Due to the myriad unavoidable flaws, a particular statement cannot be accorded *any* degree of belief simply by virtue of its derivation via sound inference.

In EBL, inference over Δ is permitted only to tie together actual world observations (e.g., to *explain* labeled training examples). This constitutes *analytic* evidence, and Δ may support many incompatible explanations for the same training instance. All are causally well formed. Choosing which (if any) explanations to accept relies largely on statistics.

An explanation is syntactically identical to a theorem. The mechanism for building explanations is just theorem proving. Each explanation "derives" the teacher-assigned classification label from the object's observable features using the statements of Δ . In doing so, the explanation ascribes a causally well-formed set of hidden or latent features to the object. These additional (unobservable) properties are introduced by inference via the domain theory. They are compositions of distinctions that the expert has found useful in expressing his or her causal conceptualization of the domain.

Each explanation hypothesizes that a particular hidden causal structure is sufficient to determine an object's class label accurately in the context of the classification problem. Thus, while syntactically the process looks like theorem proving, semantically it amounts to conjecturing a statistical hypothesis about how to estimate the classification label from a (potentially complex) pattern of observable features. Hypotheses that are statistically confirmed by independent real-world examples (and therefore possess the desired robustness properties) are re-packaged into a conventional domain theory Δ' . Thus, the elements of Δ' are believed, but the elements of Δ are not the kinds of things that merit belief nor for which statistical evidence is even relevant.

It is instructive briefly to consider the classical view of EBL/EBG [14, 6, 15, 25]. Here one would logically deduce that a goal relation holds from the example's input features. Δ , assumed to be complete and correct, might include that x can be *safely-stacked* on y if x is lighter than y or y is not fragile. Other statements provide various methods for computing weight. Observing that a particular vase can be safely stacked atop a particular table, the learner constructs a general sufficient rule for concluding *safely-stacked* from the constituent volumes and densities and not depending on the identity of the objects, their colors, owners, etc. This classical EBL, sometimes referred to as speed-up learning, inherits the brittleness of conventional first-order logic upon which it is built.

2 Brittleness and Robustness

The brittleness of conventional logic can be largely traced to properties (2) and (3) above. In the conventional logical paradigm, an axiom set, Δ , represents a model of world interactions. The set of expressions that can be inferred about the [micro]world is precisely the set of expressions Φ entailed by the axioms:

$$\Delta \models \Phi \quad (4)$$

The brittleness follows from the afore-stated qualification problem and the unforgiving nature of logical semantics. Logical inference ascribes equal absolute belief in all logical consequences. Thus, the qualification problem assures us that there will be some anomalous consequences from the logical formalization of any reasonably interesting subset of the real world, while the semantics of logic assures us that these anomalies, if encountered in practice, may be devastating to our reasoner's robustness.

But statistical inference does not suffer from such brittleness. In the conventional statistical paradigm one adopts some parameterized family of (statistical) models, \mathcal{M} . Often in the statistical literature, the term *model* is used to refer to the family, but we reserve that term for a specific candidate or stand-in for the world. In the simplest version, a member of the family, $\mathcal{M} \in \mathcal{M}$, is chosen according to a set of world observations \mathbf{Z} . Then the set of expressions Φ judged to hold in the world are those that \mathcal{M} accepts as worth believing. As a specific illustration, \mathcal{M} might be the graphical structure of a Bayesian net. The observations \mathbf{Z} are used to estimate conditional probabilities which in turn individuate a specific explicit probability distribution. Perhaps we believe those things whose ascribed probability is greater than 0.5. Equally, we might choose a discriminative model. For example, \mathcal{M} might be a family of linear separators. Here \mathbf{Z} is a training set used to select a specific linear separator, \mathcal{M} . The collection of believed expressions Φ might be the ones that fall above the linear discriminant \mathcal{M} . To paraphrase (4) we might denote a weaker form of "entailment" of the set of new sanctioned beliefs Φ as:

$$\mathcal{M} \cup \mathbf{Z} \approx \Phi \quad (5)$$

Here the model family \mathcal{M} augmented with a set of world observations \mathbf{Z} provides sufficient justification to believe each $\phi \in \Phi$. The nonstandard symbol \approx is used (rather than \models) to denote that this inference provides less than the absolute confidence of logical entailment.

In the statistical paradigm exceptions are embraced within the formalism. Inferring ϕ but observing $\neg\phi$ in the real world results in a new augmented set of world observations $\mathbf{Z}' = \mathbf{Z} \cup \neg\phi$ which together with \mathcal{M} may eventually result in the selection of a different specific model \mathcal{M}' . Even after observing $\neg\phi$ it is quite possible that $\mathcal{M} \cup \mathbf{Z}' \approx \phi$, at least until *sufficient* contradictory evidence accrues. In this view, statistical inference naturally embodies a kind of nonmonotonicity; it requires no additional mechanisms.

EBL borrows the two statistical robustness characteristics above, letting Δ (with a sound inference procedure) play the role of the statistical prior commitments, \mathcal{M} . Thus, we examine inference systems that implement something like the following:

$$\Delta \cup \mathbf{Z} \approx \Phi \quad (6)$$

Δ , as before, is a set of first-order sentences. But this Δ is not required to be consistent or robust. EBL implements (6) via (3) so that Δ' is a compact approximate representation of Φ thus: $\Delta' \models \Phi'$ and Φ' approximates Φ . Inference is performed conventionally over Δ' .

3 A Simple Classification EBL Algorithm

To illustrate, we employ English sentences rather than first-order ones. The translations are straightforward. For example, sentence 2 is the notorious $\forall x Bird(x) \Rightarrow Flies(x)$.

As a domain, we are interested in which animals can fly. Each specific animal is defined by a conjunction of ground observable features (name= Tweety, species= parakeet, color= yellow, etc.). An expert supplies us with a (non-robust) domain theory, Δ :

- | | |
|--|---------------------------------------|
| 1. flying is kind of locomotion | 8. cooking causes animals to be dead |
| 2. birds fly | 9. sick animals are weak |
| 3. locomotion is a volitional act | 10. cement blocks are heavy |
| 4. dead things do not act volitionally | 11. birds are animals |
| 5. flying requires wings | 12. penguins, ostriches... do not fly |
| 6. wings need a particular geometry | 13. penguins are birds |
| 7. flying requires a favorable power to weight ratio | 14. robins are birds |
| | ... many other similar sentences |

This is not an acceptable conventional axiom set. There are contradictions (e.g., 2, 12, and 13). More importantly, there are many derivable expressions that incorrectly portray the world. For example, from 1-4 we deduce that birds cannot be dead.

But an EBL system will never conclude that any particular bird both flies and does not fly because it will never observe a bird that requires an explanation of how it can simultaneously fly and not fly. Likewise, only after seeing an immortal bird would we entertain the explanation from 1-4 for why this animal might never die.

Robustness is defined with respect to a domain task. We conceive a **task** as an unlimited sequence of related questions posed to the reasoner. The questions are drawn randomly from a space of well-formed questions according to some fixed but unknown distribution. A theory is **robust** for a task if the answers produced by some sound inference procedure are usually the answers that the real world would give. An EBL system is robust if it usually produces robust theories. The user supplies an error tolerance parameter $0 < \epsilon \ll 1$ (bounding the probability of disagreement with the real world) and a confidence parameter $0 < \delta \ll 1$ (bounding the probability that the constructed theory is not robust).

Thus, the five inputs to an EBL system are Δ , \mathbf{Z} , a task, ϵ , and δ . The output is a new theory Δ' such that with probability of at least $1 - \delta$, the real-world accuracy of Δ' on the task is at least $1 - \epsilon$. Consider Δ :

$$\begin{aligned}
\Delta_1 &: \forall x \textit{ Flies}(x) \Rightarrow \textit{Locomotes}(x) \\
\Delta_2 &: \forall x \textit{ Bird}(x) \Rightarrow \textit{Flies}(x) \\
\Delta_3 &: \forall x \textit{ Locomotes}(x) \Rightarrow \textit{Volition}(x) \\
\Delta_4 &: \forall x \textit{ Dead}(x) \Rightarrow \neg \textit{Volition}(x) \\
\Delta_5 &: \dots
\end{aligned}$$

We posit the following procedures:

- EXPLAIN(Δ, S, L):** a sound theorem prover implementing the inference relation \vdash . This serves as an explanation generator. We require that explanations with shorter derivations are constructed before longer ones. Given a non-robust domain theory Δ , a set of world observations S , and a literal of interest L , EXPLAIN returns a succession of proof trees. Each derives the assigned literal truth value from one or more observations in S .
- RULEGEN(E):** a simple rule generation procedure such as that of Mooney and Bennett [15] which essentially lifts and flattens the proof tree. The result is a new first order statement that concludes L and tests only observable predicates.
- WORLD(L, N):** a protocol for monitoring the real world. We specify a literal of interest, L , and a positive integer N . It notifies us of new occurrences and succeeds after seeing N .

For each literal L of interest the following algorithm is invoked:

1. Set B to the singleton observation $\text{WORLD}(L, 1)$
2. Set E (an explanation) to $\text{EXPLAIN}(\Delta, B, L)$
3. Set R (a hypothesized rule) to $\text{RULEGEN}(E)$
4. Evaluate R on $\text{WORLD}(L, \ln(2/\delta)/2\epsilon^2)$
5. If R is correct on all of these, END returning R as the robust rule for L
6. Else reject R and add the new observations to B
7. Go to Step 2

We assume that the domain theory was created by a true expert and is *adequate* in a sense we will make formal in section 5. Basically, mixed in with all of the specious causal analyses there must be at least one that satisfies our robustness requirements, and the expert cannot intentionally make those ones more difficult to find.

Now suppose we see Rob, a flying robin. He is explained by an instantiation of Δ_2 . Lifting and flattening results in a statement identical to Δ_2 hypothesized to be included in Δ' :

$$\forall x \textit{ Bird}(x) \Rightarrow \textit{Flies}(x) \tag{7}$$

Next we see Tom, the non-flying turkey. Encountering Tom initiates two computations. First, he serves to refute 7. Second, a derivation is initiated to explain Tom's non-flight whose simplest explanation is:

$\neg Flies(Tom)$	Given
$\neg Flies(Tom) \Leftarrow \neg Locomotes(Tom)$	CP Δ_1
$\neg Volition(Tom) \Rightarrow \neg Locomotes(Tom)$	CP Δ_3
$\neg Volition(Tom) \Leftarrow Dead(Tom)$	Δ_4
$Dead(Tom)$	Given

(CP means contrapositive of while \parallel shows unifications) This explanation when lifted and flattened yields the conjectured rule:

$$\forall x Dead(x) \Rightarrow \neg Flies(x) \quad (8)$$

Next we see more world observations from our bird flying task. Some, the sparrows and blue jays eating from our backyard feeder, fly. Others, the roasted chicken we have for dinner later in the week, the cooked Cornish game hen the next day, the neighbor's parakeet killed by their cat, do not fly. Rule 7 is refuted. This does not change Δ (which still contains Δ_2). But 7 is dropped from consideration from Δ' . This reinvokes EXPLAIN on Rob, Tom, and the other evidence observations that participated in the evaluation of 7. Two additional inference rules are required by EBL. We will see these in the next section. The second one, (13), is used here to conjecture the rule:

$$\forall x Bird(x) \wedge \neg Dead(x) \Rightarrow Flies(x)$$

After a significant number of similar observations this and (8) are statistically confirmed and Δ' becomes:

$$\begin{aligned} \forall x Dead(x) &\Rightarrow \neg Flies(x) \\ \forall x Bird(x) \wedge \neg Dead(x) &\Rightarrow Flies(x) \end{aligned}$$

If we had seen a significant number of airplanes, penguins, emus, Mafioso birds, etc. these rules would be different. But in this task context, these rules encounter many confirming and no disconfirming examples.

4 Some Semantic Properties of Explanations

The expert provides us with a set of first-order sentences that capture his or her understanding of the domain:

$$\Delta = \{\theta_i \mid i = 1, r\} \quad (9)$$

Each first-order θ only approximates some underlying constraint of the real world. Thus, we interpret the meaning of each statement θ_i as $\alpha_i \Rightarrow \theta_i$ where the α 's denote the (possibly infinite) missing qualifiers from the qualification problem. The veracity of a deduction over Δ depends on the unmodeled α 's and the unifications performed.

Without loss of generality (by renaming variables as needed) assume that we have a single global unifier, Γ , constructed as a side effect of the explanation process. A necessary and sufficient condition for the deduction to hold in the real world is the conjunction of the implicit qualifications:

$$\left(\bigwedge_j \alpha_j \right) \circ \Gamma \quad j = 1, s \quad (10)$$

where \circ denotes the specialization of a first-order expression by the application of a unifier. The index j ranges over sentences in the deduction.

To achieve robustness, EBL insures that the corresponding expression (10) of each sentence in Δ' holds with high probability. But note that this needs to be met only *when the sentence is applied*. Thus, in EBL we are interested in

$$Pr \left(\left(\bigwedge_j \alpha_j \right) \circ \Gamma \quad j=1,s \right) \quad (11)$$

where the probability distribution is taken over just those situations in which the inference mechanism chooses to construct and employ the sentence. The EXPLAIN procedure draws inferences according to this very distribution when constructing explanations for real world observations. As a derivation grows, its (10) incurs an additional independent chances to fail so (11) tends to diminish. Let λ be the factor characterizing our expected erosion of confidence from a single additional inference step. Thus, $\lambda \leq 1$ with equality only in artificially clean domains where the qualification problem does not apply.

We assume the inference mechanism is paraconsistent or “inconsistency-tolerant” (e.g., relevance logic), and we disallow hyper-inference. We also require additional inference rules that define the inference relation \vdash (implemented here by EXPLAIN). The new inference rules are sound but unnecessary in conventional logic. For intuitive clarity we state them using implications. Of course, as with first-order inference rules, the identity match (of ϕ in each case) can be brought about by specializing first-order expressions through unification.

The first rule is a kind of AND introduction:

$$\frac{\begin{array}{l} \psi \Rightarrow \phi \\ \varphi \Rightarrow \phi \end{array}}{(\psi \wedge \varphi) \Rightarrow \phi} \quad (12)$$

While sound, this inference rule is logically unnecessary. Statistically it is quite useful. If alone, explanations from each of two pieces of evidence are insufficient statistically, then the desired accuracy might be achieved by insisting on both evidentiary sources together.

A second important evidentiary rule that is logically unnecessary is:

$$\frac{\begin{array}{l} \psi \Rightarrow \phi \\ \varphi \Rightarrow \neg\phi \end{array}}{(\psi \wedge \neg\varphi) \Rightarrow \phi} \quad (13)$$

The first statement subsumes the inferential conclusion, making it logically redundant. But statistically it resembles conditioning. In the Rob / Tom illustration, given only one of the observations, we can conjecture that “birds fly” or that “dead things do not fly.” In the presence of both kinds of world observations, the first statement cannot be confirmed. This rule allows conjecturing the composite sentence that “birds that are not dead can fly.”

5 Analysis of Simple EBL

To prove that the simple EBL algorithm works, we will use λ , introduced in the previous section, β and γ , introduced below, and the parameters ϵ and δ , introduced in section 3.

Proposition 1: If simple EBL produces a rule whose actual real-world task accuracy is a , then

$$Pr(a < (1 - \epsilon)) \leq \delta/2 \quad (14)$$

That is, its true accuracy cannot be far from its measured accuracy (which is 100% or else it would not have been produced).

Proof: The additive Chernoff bound requires the true mean of a distribution μ and an estimated mean $\hat{\mu}$ based on m samples from the distribution respect $Pr(\mu < \hat{\mu} - \epsilon) \leq e^{-2m\epsilon^2}$

Let μ be the true real world accuracy of the rule. The rule is accepted only if it makes no errors so $\hat{\mu} = 0$. Letting m be $\ln(2/\delta)/2\epsilon^2$ (the algorithm’s number of world observations) yields (14).

Proposition 2: The set of explanations from Δ can grow no faster than exponentially as inference depth increases. This follows from the assumptions of a sound paraconsistent logic without hyper-inference. This is proved in [20].

Definition 1: The *Domain Adequacy Measure* γ of a theory is $\lambda \cdot \beta$ where λ is the expected inferential erosion of confidence from the previous section and β is the base of the exponential from Proposition 2.

Definition 2: An input domain theory Δ is *adequate* iff some robust rule can be derived for any real world questions of interest and $\gamma < 1$. The domain expert must include in Δ a sufficient set of conceptual distinctions and causally simple explanations cannot be hidden in a plethora of easy-to-derive Rube-Goldbergish ones.

Proposition 3: If given an adequate theory, a question of interest, and a set of world observations, then a robust explanation (one giving rise to a rule robustly answering the question of interest in the real world) can be found with probability at least $1-\delta/2$ in no more than $n + k$ inference steps where n is the number of inference steps needed to find the simplest explanation for the observations, and k grows no faster than logarithmically in δ and γ and is independent of the complexity parameter ϵ .

Proof: We are given an adequate domain theory, an untested explanation derivable with n inference steps, and the assurance that no explanation is possible with fewer than n steps (as this one is stipulated to be the simplest). The probability of finding the first robust explanation R at inference level j is geometrically distributed with base γ :

$Pr(R|j) = (1 - \gamma) \cdot \gamma^{j-1}$. We condition on the known information that there is zero probability of finding R before inference level n. Define k to be the expected number of additional inference steps so that the robust R is unlikely to be missed (i.e., so that the cumulative tail of $Pr(R|j)$ for $j > n + k$ is bounded by $\delta/2$). This is equivalent to bounding the cumulative tail of the original geometric distribution starting at k which, in this form, is just γ^k . Requiring $\gamma^k \leq \delta/2$ results in $k \geq \ln(\delta/2)/\ln\gamma$. By inspection this has the specified dependence on ϵ , δ , and γ .

Theorem: With probability at least δ Simple EBL produces rules with real world accuracy of at least $1 - \epsilon$ and requires no more than a number of relevant world observations polynomial in $1/\epsilon$ and $1/\delta$.

Proof: We split the δ resource into two halves. Proposition 1 limits the probability to $\delta/2$ that a rule passes the robustness test but fails to actually achieve a real world accuracy of $1 - \epsilon$. Each robustness test consumes a polynomial number of relevant real-world observations (Proposition 1). With k additional inference levels only a polynomially growing number of additional explanations can be encountered: The set of explanations may grow exponentially in k (Proposition 2), but k grows only logarithmically in the complexity parameters (Proposition 3). The first adequate explanation might not be found in k additional inference steps but this is unlikely. By Proposition 3 this occurs only with probability $\delta/2$. Thus, the algorithm fails at most δ of the time (half the time by poorly testing a rule, and half the time by failing to search far enough to find the first good rule).

6 Domain Adequacy, Scaling, and Algorithmic Complexity

Will our new form of EBL scale to non-toy problems? Our first rather tentative step does not answer this important question. There is reason for concern: using full first-order explanations, EBL's worst case time (NB: not example) complexity is at least exponential for derivation membership and undecidable for non-membership. Of course, a less expressive explanation engine would force more favorable algorithmic complexity guarantees.

But we believe that the key to scaling may lie in the notion of *domain adequacy*. Domain adequacy provides a new quality measure on domain theories upon which algorithmic complexity can depend. Here we employed a simple adequacy measure $\gamma = \lambda \cdot \beta$ sufficient for example complexity. But this is just a first cut at a deep and important contrast to logical adequacy. The rules of chess and Peano's axioms may appear at first to be perfectly adequate for their respective domains. But an expert chess tutor prefers to describe games using less-precise invented domain terms such as "center control," "weak pawn structure," and "underdeveloped queen side." From the viewpoint of EBL, these expert-introduced terms are both more informative and more flexible. As latent variables or sub-concepts, they must be learned by the student. Their "correct" definition can (and will) depend on the student's own emerging idiosyncratic style of play. We believe this new EBL approach opens the door to a richer notion of "domain theory" and "training example." In this new EBL, the role of prior knowledge is to linearize the learning problem by conjecturing alternative sufficient sets of manageable sub-problems. Conventional notions of accuracy, satisfiability, or sets of possible

worlds cannot express these important characteristics of a domain theory. Perhaps domain theory adequacy can asymptotically bound the number of wasted inference steps by employing appropriate sub-concepts.

7 Related Work and Conclusions

The approach owes much to earlier work on learning with domain theories and declarative bias [24, 3, 4, 2, 18]. Inductive Logic Programming is also relevant [13, 17], as is the work on theory revision [26, 19, 10, 1]. These also acquire expressive representations but theirs is a much more ambitious goal of improving the expert-supplied domain theory rather than constructing a new specialized theory for a particular narrow task. The work combining first-order knowledge with statistics is also relevant (e.g., [16, 5]), as is learning in probabilistic logics (see [7]). However, in our Explanation-Based Logic no statistical or numerical manipulations take place during inference; there are no probabilities attached to the sentences in the robust output domain theory Δ' . This avoids a computational pitfall [23] without constraining the expressiveness of the result to a (propositional) Bayesian net as [11, 12]. The Knowledge-Based Neural Networks approach [27] is similar, utilizing a propositional neural net rather than Bayesian net. The burgeoning area of relational learning ([22, 21]) is also relevant, although link learning, relational learning, learning with description logics, etc. all employ knowledge representations that fall short of first-order expressiveness.

For some application domains, this new form of EBL may allow complex concepts to be learned from small more human-proportioned training sets. Possible applications include intelligent interfaces in which the system can learn to fit its user rather than forcing the human user to learn about it, and context adaptive computing in which a computer system specializes itself to its perceived deployment context. A preliminary less declarative illustration of this direction can be found in [8].

The main contribution is tolerating semantic approximation in the expert-supplied logic-like statements, and the use of world observations as evidence to interpret this domain knowledge.

Acknowledgements The Illinois EBL group, particularly Arkady Epshteyn, Shiao Hong Lim, Geoff Levine, and Li-Lun Wang, contributed immensely to the work. It also benefited greatly from comments from Ron Brachman that helped to clarify a number of issues. The research was supported by the National Science Foundation under Award NSF IIS 04-13161 and by the Information Processing Technology Office of the Defense Advanced Research Projects Agency under award HR0011-05-1-0040. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of NSF or DARPA.

References

1. S. D. Bay, D. G. Shapiro, and P. Langley. Revising engineering models: Combining computational discovery with knowledge. In *Thirteenth European Conference on Machine Learning*, pages 10–22, 2002.

2. Clifford Brunk and Michael J. Pazzani. A lexical based semantic bias for theory revision. In *ICML*, pages 81–89, 1995.
3. Peter Clark and Stan Matwin. Using qualitative models to guide inductive learning. In *ICML*, pages 49–56, 1993.
4. William W. Cohen. Incremental abductive ebl. *Machine Learning*, 15(1):5–24, 1994.
5. James Cussens. Parameter estimation in stochastic logic programs. *Machine Learning*, 44(3):245–271, 2001.
6. Gerald DeJong and Raymond Mooney. Explanation-based learning: An alternative view. *Machine Learning*, 1(2):145–176, 1986.
7. Luc DeRaedt and Kristian Kersting. Probabilistic logic learning. *SIGKDD Explorations*, 5(1):31–48, 2003.
8. A. Epshteyn, M. Garzaran, G. DeJong, D. Padua, G Ren, X. Li, K. Yotov, , and K. Pingali. Analytic models and empirical search. In *The Eighteenth International Workshop on Languages and Compilers for Parallel Computing*, 2005.
9. Michael Genesereth and Nils Nilsson. *Logical Foundations of Artificial Intelligence*. Kaufmann, Los Altos, CA, 1987.
10. Russell Greinter. The complexity of theory revision. *Artificial Intelligence*, 107(2):175–217, 1999.
11. Peter Haddawy. Generating bayesian networks from probability logic knowledge bases. In *UAI*, pages 262–269, 1994.
12. Niels Landwehr, Kristian Kersting, and Luc De Raedt. nfoil: Integrating naïve bayes and foil. In *AAAI*, pages 795–800, 2005.
13. Nada Lavrac and Saso Dzeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, , New York, 1994.
14. Tom Mitchell, Richard Keller, and Smadar Kedar-Cabelli. Explanation-based learning: A unifying view. *Machine Learning*, 1(1):47–80, 1986.
15. Raymond Mooney and Scott Bennett. A domain independent explanation-based generalizer. In *AAAI*, pages 551–555, 1986.
16. Katharina Morik, Peter Brockhausen, and Thorsten Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *ICML*, pages 268–277, 1999.
17. Stephen Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *J. Log. Program.*, 19/20:629–679, 1994.
18. C. Nédellec, C. Rouveirol, H. Ade, F. Bergadano, and B. Tausend. Declarative bias in ILP. In L. De Raedt, editor, *Advances in Inductive Logic Programming*, pages 82–103. IOS Press, Amsterdam, 1996.
19. Dirk Ourston and Raymond Mooney. Theory refinement combining analytical and empirical methods. *Artif. Intell.*, 66(2):273–309, 1994.
20. David Plaisted and Yunshan Zhu. *The Efficiency of Theorem Proving Strategies*. Vieweg & Sohn, Wiesbaden, 1999.
21. Luc De Raedt, Thomas Dietterich, Lise Getoor, and Stephen H. Muggleton, editors. *Probabilistic, Logical and Relational Learning*, Dagstuhl Seminar Proceedings, 2005.
22. Luc De Raedt and Stefan Kramer, editors. *Workshop on Attribute-Value and Relational Learning: Crossing the Boundaries at ICML 2000*, 2000.
23. Dan Roth. On the hardness of approximate reasoning. In *IJCAI*, pages 613–619, 1993.
24. S. J. Russell and B. N. Grosz. A declarative approach to bias in concept learning. In *Proc. of AAAI-87*, pages 505–510, Seattle, WA, 1987.
25. Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition edition, 2003.
26. Lorenza Saitta, Marco Botta, and Filippo Neri. Multistrategy learning and theory revision. *Machine Learning*, 11(2-3):153–172, 1993.

27. Geoffrey G. Towell and Jude W. Shavlik. Knowledge-based artificial neural networks. *Artif. Intell.*, 70(1-2):119-165, 1994.