An Alternative To Deduction

Daniel Oblinger
Beckman Institute, University of Illinois, 405 N. Mathews, Urbana, IL 61801
Email: oblinger@cs.uiuc.edu Phone: (217) 244–1503

Gerald DeJong
Beckman Institute, University of Illinois, 405 N. Mathews, Urbana, IL 61801
Email: dejong@cs.uiuc.edu Phone: (217) 333–0491

Abstract

Deductive representations are well defined, easily inspected, and precise. However, they are also brittle, inflexible and difficult to debug. We propose a *plausible* representation whose inference mechanism is weaker than its deductive counterpart. This will allow it to reason with knowledge which is less precise, and replaces the notion of global consistency, with the weaker constraint of local consistency in its explanations[1].

1. PI–EBL approach

There are many advantages to a reasoning model based on a declarative representation of world knowledge. Such representations may be leveraged by many knowledge intensive mechanisms: it can be communicated to other agents, multiple models can be combined, parts of a model can be isolated and independently verified, etc. Because deduction is well understood it is natural to ground these models in a deductive framework. Requiring all knowledge to be expressed in a deductive form is restrictive, and causes systems to be very brittle. Connectionist approaches, on the other hand, are not based on such a restricted language, which is a source of their flexibility [Rumelhart86]. The symbolic/connectionist distinction may appear unimportant since the sources of knowledge potentially available are the same. However, research in connectionism is impeded by its representation—its very difficult to understand what abstract principles are revealed by a connectionist solution to a problem. The advantages of both approaches can be achieved by a reasoning paradigm based on an explicit world model (like symbolic approaches) using a non–deductive representation (like connectionist approaches). Research is facilitated by such a paradigm, high–level knowledge intensive mechanisms can be more easily constructed, and the principles underlying their success are much more available for inspection. In this paper we introduce a logic of plausible reasoning with these desired properties. As in distributed connectionist systems, our approach crucially involves learning through observations of the world..

1. This research was supported by the National Science Foundation under grant NSF–IRI–87–19766

The Explanation Based Learning paradigm is extended to support learning over our logic of plausible reasoning. EBL is a learning mechanism which combines a model of the world with examples to construct explanations which can be later used to reason about the world [DeJong86, Mitchell86]. Like other knowledge intensive mechanisms, it gains its strength from its explicit world model. Much of EBL's potential as a methodology has not been tapped, however, because the knowledge representation it uses is too restrictive. For example, EBL's restriction to symbol level learning is due to its base representation, *not* an inherent restriction on the EBL paradigm. (Symbol level learning adds no new knowledge but simply makes some knowledge easier to derive [Dietterich86].) EBL's limited use of its training examples is also traceable to its base representation. The flexibility of any architecture depends, in large part, on the representation language it uses, no matter what high level task is being performed. We present a plausible representation and inferencing mechanism as an alternative to classical deduction. The flexibility of this representation is demonstrated by using the Plausible Inferencer as the explanative component for an Explanation Based Learner (PI–EBL). Through its flexible representation PI–EBL overcomes EBL's deficiencies mentioned above, and acquires some advantages associated with both connectionism and traditional induction by greater use of its training examples.

2. Motivation Of Plausible Theories

Deductive frameworks are not sufficient as a knowledge representation mechanism. Their deficiency stems from their precision. Knowledge placed in a deductive framework must be valid—true for all possible examples. This is a problem since much of the knowledge a reasoner must deal with is true for most but not *all* examples; such knowledge cannot be used in a deductive framework. The impossibility of representing real world knowledge in a deductive theory is noted by McCarthy as the qualification problem [Genesereth87, McCarthy69]. The qualification problem states that any universally quantified implication will need a large (infinite) number of preconditions to exclude *all* possible exceptions. The reader may be convinced by considering the implication: $Bird(x) \rightarrow Fly(x)$. Correcting this rule to handle exceptions would require adding $Alive(x)$, $\neg Penguin(x)$, $Sane(x)$, etc. as conjuncts. Of

course such a deductive theory is not practical. The traditional approach to this problem is to translate the original domain into a micro–world which is consistent with the original domain to a fixed level of detail. Because the micro–world does not reflect all details of the real world, it is finitely describable using a deductive theory, thus a deductive theory may be employed. The disadvantage to this approach is that the model is really a model of the micro–world *not* a model of the original domain. The reasoner using such a theory has no recourse if some important portion of the domain is not modeled in the micro–world. Thus, the micro–world must be complex enough to handle the most detailed example encountered by the reasoner. This forces the entire system to represent and reason using this detailed and complex theory for *all* examples, even those which can be handled using a much simpler theory. Adapting the theory to handle one additional special–case example can easily result in many previously tractable problems to becoming inaccessible. Furthermore, simple theories which make different assumptions (based on different micro–worlds) cannot be directly combined because the resulting theory would likely be internally inconsistent. This is a stifling restriction since different micro–worlds (consistent micro–theories) will be useful in reasoning about different problems. Combining inconsistent micro–theories into a *single* consistent micro–world will result in a theory which is too precise, it is too precise because the closure of these micro–theories contain many facts inconsistent with other micro–theories, and the combined theory must be consistent with only *one* of these choices. Because the combined theory is deductive it must specify *exactly* how its component theories interact, even when the interactions are uninteresting parts of the deductive closure. Using a micro–world forces the theory to make predictions (by way of its deductive closure) about hypothetical situations, most of which are not relevant to the goals of the reasoner. No deductive chain of inference no matter how complex may be inconsistent with any other over the entire theory. Requiring the theory to be internally consistent without gaining consistency with the world gains us little power at the great expense of scalablity. A reasoner using deductive knowledge cannot accept *any* knowledge about the world until all the interactions with the *closure* of its current knowledge are precisely specified. This restriction is unacceptable. Avoiding this restriction and its fixed level of detail forces us to abandon the micro–world approach.

Non–monotonic theories provide a mechanism for reasoning with incomplete knowledge [McCarthy86, Reiter80]. McCarthy's *Abnormal*(x) predicates allows one set of rules which derives a conclusion to override another rule. Like deductive theories, however, the sentences entailed by a nonmotonic theory are still precisely defined without any examples needed to disambiguate inconsistencies. The whole notion of circumscription makes precise the closure of a non–monotonic theory. In order for a representation to be an effective alternative to the micro–world approach, it must incorporate a means of trading the complexity/precision of a theory against its faithfulness to the world. This means the reasoner cannot close knowledge base (it cannot assume it has complete knowledge of the domain). Instead the world must be left open. Inferred knowledge should be used as *suggestions* to be empirically verified, rather than theorems with are necessarily true.

3. Representation Of Plausible Theories

Plausible domain theories are syntactically similar to a set of implications. Semantically they are quite different, however, since the theories are not deductive. The theories must *suggest* possible relationships between concepts without *entailing* them. Semantically the plausible domain theory may be interpreted as deductive implications with missing preconditions. The missing preconditions are collectively called the *implicit context* of the implication. The implicit context is the set of additional constraints sufficient to guarantee that the plausible implication deductively holds. Thus we refer to plausible implications as *influents* to emphasize the idea that the consequent is influenced, but not entailed, by the preconditions. Asserting an influent specifies that its preconditions are relevant to determining its consequent in some contexts. An influent also specifies the direction of influence between the preconditions and consequent, and they may be positively or negatively related. These influents provide the basis for representing PI–EBL domain theories.

Influent: $P \succ Q$

Meaning: $I \wedge P \rightarrow Q$

(I: Implicit Context)

Figure 1: Semantics of an Influent.

4. Plausible Inferencing

To build explanations from a plausible domain theory (a set of influents) we need an inference mechanism. Influents may be combined with other influents in the same way deductive implications are chained together. Unifying a precondition of one rule with the consequent of another provides the first influent inference mechanism. Chaining influents in this fashion may result in explanations with important preconditions not explicitly represented since each influent has an implicit context which also must be satisfied. Other chainings which derive this conclusion or its negation will also have missing preconditions based on the implicit contexts of its influents. The missing preconditions of the other explanations are likely to be very different, thus together these chainings could provide an explanation which is predictive over a greater range of examples. This is the motivation for the second inference mechanism, the join, which allows combining several influents with the same consequent (or its negation) into an *influent set*. The influent set contains no information regarding the applicability of its component influent chains but it does guarantee

that there is some combination of these chains which correctly predicts the conclusion of the influent set over the relevant problem distribution. Unlike the chaining inference mechanism, the join mechanism only partially specifies the function mapping preconditions to conclusions. *Plausible explanations* are generated by repeated application of these two inference mechanisms. Because of the influent sets within a plausible explanation it cannot be used to directly make predictions about its conclusion. The meaning of a plausible explanation is analogous to the semantics of Stuart Russell's determinations [Grosof89], both serve as a specification of the set of determining factors for some value in the domain. A plausible explanation specifies that its leaf preconditions determine its consequent, without completely specifying the mapping. Unlike Russell's determinations, however, there is no guarantee that a plausible explanation actually *does* determine its consequent because nothing is known about its implicit contexts. So a plausible explanation must be empirically verified by checking that it is consistent with examples from the domain before it can be accepted as a plausible theorem. Many different explanations can be conjectured for the same observed behavior. Ideally, the most plausible explanations are generated first. If these are empirically rejected (ie. the observed data contradicts them) the next most plausible explanation is hypothesized and so on until some hypothesis is empirically confirmed. Generation of plausible *theorems* is more complex than its deductive counterpart since is combines both analytical and empirical constraints.

In order to inductively learn a plausible theorem our reasoner must accept and reject potential explanations based on examples. This is challenging since the influents that makeup the explanations can be neither confirmed nor rejected based on empirical evidence since we can never know if their implicit context is satisfied. As yet we have no restriction that a plausible explanation accurately predicts the world, it is this restriction which makes plausible explanations testable. Thus we define a plausible theorem to be a plausible explanation which contains influent sets which satisfies the three properties shown in figure 2 relating it to the set of observations. As shown in the figure an

Let O be the set of observations.
Let A and D be the preconditions of the affirming and denying influents in the influent set I.
Let c be the conclusion of the influent set I.
Let $P \equiv A \cup D$.
Sat(predicate, observation) is true iff the predicate was satisfied in the particular observation.

Def: I is *deterministic* on O iff

$$\forall o_i, o_j \in \mathbf{O} \quad \left(\forall p \in P \quad Sat(p, o_i) = Sat(p, o_j)\right) \rightarrow \left(Sat(c, o_i) = Sat(c, o_j)\right)$$

Def: I is *predictive* on O iff

$$\neg \exists o_i, o_j \in \mathbf{O}; A' \subseteq A$$
$$\left(\forall p \in P \backslash A' \quad Sat(p, o_i) = Sat(p, o_j)\right) \wedge$$
$$\left(\forall a \in A' \quad \neg Sat(a, o_i) \wedge Sat(a, o_j)\right) \wedge Sat(c, o_i) \wedge \neg Sat(c, o_j)$$

$$\neg \exists o_i, o_j \in \mathbf{O}; D' \subseteq D$$
$$\left(\forall p \in P \backslash D' \quad Sat(p, o_i) = Sat(p, o_j)\right) \wedge$$
$$\left(\forall a \in D' \quad \neg Sat(a, o_i) \wedge Sat(a, o_j)\right) \wedge \neg Sat(c, o_i) \wedge Sat(c, o_j)$$

Def: I is *relevant* on O iff

$$\forall p \in P \quad \exists o_i, o_j \in \mathbf{O}$$
$$\left(\forall r \in P \backslash p \quad Sat(r, o_i) = Sat(r, o_j)\right) \wedge$$
$$\left(Sat(p, o_i) \neq Sat(p, o_j)\right) \wedge \left(Sat(c, o_i) \neq Sat(c, o_j)\right)$$

Figure 2. Consistency properties of an influent set.

influent set is deterministic if its consequent is completely determined by its preconditions with respect to the set of observations. There cannot be two observations which agree on the truth value of all of the preconditions but do not agree on the truth value of the consequent. This is the same constraint which Stuart Russell's determinations have. The second consistency property is predictiveness. When an influent is needed to distinguish between two observations, the influence it has on the consequent must be in the predicted direction. If two observations differ only in the truth values of the preconditions of several positive influents then the consequent must be true when those preconditions are true. This constraint ensures that positive values assigned to the preconditions can only have a posi-

tive effect on the consequent. The analogous negative relationship must also hold for negative influents. The final consistency property is relevance. An influent set is relevant over a set of observations if the preconditions of each influent are necessary for uniquely determining the value of the consequent. In effect, no influent superfluous to the set of observations is permitted; each influent must be the determining factor in some circumstance otherwise it could be eliminated from the explanation with no effect on the learned function. Relevance is required since it disallows plausible theorems which are correct in the sense that their preconditions determine their consequent, but are useless since they have so many irrelevant influents. Any explanation that satisfies these three properties determines a function which can be used to model the domain. This is PI–EBL's ultimate goal. Thus the PI–EBL approach is a search for plausible theorems—explanations which are empirically shown to be deterministic, predictive, and relevant.

5. Reasoning With PI–EBL

A fault found with traditional EBL is its limited use of training examples. Examples are much more central to PI–EBL learning. The PI–EBL approach has aspects of traditional induction, but it is not a hybrid system in the sense that it can be easily decomposed into these separate components. Because of the homogeneity of the approach, aspects of both approaches are intertwined in its learning. The plausible inferencer is used to explain the value of the predicate of interest in the system's observations. Observations are also used to verify or reject hypotheses generated by the inferencer. This is done in a generate and test fashion. If an observation presented to the system causes failure of the current hypothesis to be deterministic, predictive, or relevant over the set of observed examples, then the hypothesis is rejected and the inferencer is used to generate another plausible explanation. As plausible explanations are being generated and monitored for correctness the observed examples are used for yet another purpose. For every influent set in the plausible explanation there is a function mapping the the preconditions of the influent set to its consequent. Observations are used to constrain these functions. This use of the observations is more similar to traditional induction since they are used to learn a function with fixed attributes (preconditions of the explanation). Even here, however, the explanative structure is relevant since it constrains the function being induced; it must be consistent with the plausible explanation.

The plausible inferencer builds explanations of the observation in much the same way that traditional EBL explains the observations it is given. Unlike EBL, the PI–EBL approach admits multiple incompatible explanations of the same goal concept. Because of their implicit context, the predictiveness of these explanations will vary greatly, which places great importance on the *order* that the plausible inferencer uses in generating its explanations. Since these explanations bias PI–EBL's learning, constraints on their order provide an ordering bias for the entire PI–EBL system. The a priori chance that an explanation is a theorem can be approximated in several ways. The size of the explanation is useful, larger explanations generally must satisfy a larger implicit context so they are more likely to fail. Another promising estimator of predictiveness can be derived by combining the predetermined predictiveness of each of the influents in the explanation.

6. Learning Example

We present a simple example of the PI–EBL learning algorithm to supplement its description and as a concrete example of its advantages over traditional EBL.

$$Clawed(x) \succ Bird(x)$$
$$Feathered(x) \succ Bird(x)$$
$$Immobile(x) \succ Fly(x)$$
$$Dead(x) \succ Immobile(x)$$
$$Bird(x) \succ Fly(x)$$

Figure 3: Plausible Domain Theory

Figure 3 shows a plausible domain theory which is used to predict which objects fly. The theory is plausible since it draws contradictory conclusions about immobile birds. Each of the influents are plausible in the sense that they have many exceptions. We will demonstrate how an agent can use the remainder of these influents to rectify one of the missing preconditions in the $Bird(x) \succ Fly(x)$ influent. For reasons external to the PI–EBL algorithm assume the reasoner needs to learn which objects are likely to fly. The first example seen by the agent is a clawed, feathered bird which can fly. This is explained using: $Clawed(x) \succ Bird(x) \succ Fly(x)$. This explanation is rejected when the agent sees a clawed, hairy bear which cannot fly. The next most a priori plausible explanation chosen is consistent with the examples seen: $Feathered(x) \succ Bird(x) \succ Fly(x)$. Even this explanation is insufficient in explaining a hunters trophy: a dead bird with feathers and claws which cannot fly. As before, the explanation is rejected and the next most a priori plausible explanation is generated. Figure 4 shows the next explanation, it has an influent set com-

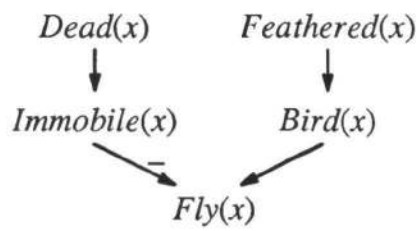

Figure 4: Plausible Explanation

posed of two influences at its top level. There are two functions mapping deadness and featheredness to the ability to fly which are consistent with this plausible explanation. One predicts dead, feathered objects fly, and the second function predicts that they do not. Both are consistent with the theory but only the first is consistent with the hunter's trophy so it is accepted. There are actually sixteen functions which accept two boolean variables, but only two of the sixteen functions would cause the explanation to be deterministic, predictive, and relevant over examples consistent with the function. If the explanation is insufficient to cover some example (it is not deterministic over the examples) then a new explanation is generated, and so on, until enough examples are seen that are consistent with this explanation. At this point it is accepted as a plausible theorem along with the classification function mapping the leaves of the explanation to its consequent.

$Bird(x) \succ Fly(x)$ is missing the precondition "not dead". This is rectified by combining it in an explanation dealing with immobility. The missing precondition was "added" by building the composite explanation. If $Bird(x) \succ Fly(x)$ is later used in another explanation its simpler form may be used. This is important since it is quite likely that the dead bird exception will be irrelevant (as will the hundreds of other exceptions). As we discussed earlier the deductive version: $\ldots \wedge \neg Dead(x) \wedge Bird(x) \succ Fly(x)$ would force the system to reason about *all* those exceptions at *all* times. Plausible explanations allow the reasoner to correctly handle these cases without forcing all reasoning to be done at this level. Notice the imprecision in the plausible theory, the theory does not specify which of the two competing sub–explanations would prevail if both had their preconditions satisfied. Of course if that level of detail were known it could easily be specified. The important point is its possible to specify that featheredness relates to flying and deadness relates to flying without precisely specifying *all* possible interactions between the knowledge. The importance of this simplification increases as the theory increases in complexity because the number of possible interactions increases dramatically with complexity, and many of these interactions need not be dealt with by the reasoner since they will never occur in practice. Even if all necessary distinctions were known, expanding a plausible theory into a deductive theory would cause duplication. As it stands now $Dead(x) \succ Immobile(x)$ can be used as an exception to flying, walking, swimming, dancing, etc. If it were written as a deductive theory the same exception would need to be placed explicitly in the preconditions of many rules.

7. Conclusions

The advantages of knowledge intensive approaches are well known. A reasoner using a plausible model instead of a deductive model will retain these advantages, but avoids the unnatural constraints of global consistency and precision of traditional deductive systems. This is very important because the agent may accept and reason with new knowledge without completely understanding all possible interactions with existing knowledge. The PI–EBL approach, because of its representation, allows very simple and imprecise theories to be made precise by learning how the plausible knowledge interacts from actual world observations (examples). In this way the agent can boot–strap itself into specialized domains by combining relevant portions of very general theories, thus building a specialized theory which is adapted empirically.

References

[DeJong86] G. F. DeJong and R. J. Mooney, "Explanation–Based Learning: An Alternative View," *Machine Learning 1*, 2 (April 1986), pp. 145–176.

[Dietterich86] T. G. Dietterich, "Learning at the Knowledge Level," *Machine Learning 1*, 3 (1986), pp. 287–316.

[Genesereth87] M. Genesereth and N. Nilsson, *Logical Foundations of Artificial Intelligence*, Morgan Kaufmann, Palo Alto, CA, 1987.

[Grosof89] B. N. Grosof and S. J. Russell, "Declarative Bias for Structural Domains," *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, NY, June 1989, pp. 480–482.

[McCarthy69] J. McCarthy and P. J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in *Machine Intelligence 4*, B. Meltzer and D. Michie (ed.), Edinburgh University Press, Edinburgh, Scotland, 1969.

[McCarthy86] J. McCarthy, "Applications of Circumscription to Formalizing Common–Sense Knowledge," *Artificial Intelligence 28*, (1986), pp. 89–116.

[Mitchell86] T. M. Mitchell, R. Keller and S. Kedar–Cabelli, "Explanation–Based Generalization: A Unifying View," *Machine Learning 1*, 1 (January 1986), pp. 47–80.

[Reiter80] R. Reiter, "A Logic for Default Reasoning," *Artificial Intelligence 13*, 1–2 (April 1980), pp. 81–113.

[Rumelhart86] D. E. Rumelhart, G. E. Hinton and J. L. McClelland, "A General Framework for Parallel Distributed Processing," in *Parallel Distributed Processing: Explorations in the Micro–Structure of Cognition*, D. E. Rumelhart and J. L. McClelland (ed.), MIT Press, Cambridge, MA, 1986, pp. 46–73.