SANDIA REPORT

SAND2012-8196
Unlimited Release
Printed September 2012

Combining Deduction and Abduction: Toward an Integrated Theory of Information Processing

David Stracuzzi

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2012-8196
Unlimited Release
Printed September 2012

Combining Deduction and Abduction: Toward an Integrated Theory of Information Processing

David J. Stracuzzi

Cognitive Modeling Department
Sandia National Laboratories
P.O. Box 5800 MS1327
Albuquerque, NM 87185-1327
djstrac@sandia.gov

Abstract

Inference techniques play a central role in many cognitive systems. They transform low-level observations of the environment into high-level, actionable knowledge which then gets used by mechanisms that drive action, problem-solving, and learning. This paper presents an initial effort at combining results from AI and psychology into a pragmatic and scalable computational reasoning system. Our approach combines a numeric notion of plausibility with first-order logic to produce an incremental inference engine that is guided by heuristics derived from the psychological literature.

In this report, we illustrate core ideas with detailed examples and pseudocode. We also discuss the advantages of the approach with respect to cognitive systems, along with the properties of problems for which the approach is best suited. Finally, we outline several key directions for future research.

Contents

1 Introduction 7
2 Knowledge Representation 8
2.1 Beliefs and Working Memory 8
2.2 Rules and Long-Term Memory 9
3 Inference Patterns and Confidence Propagation 11
3.1 Inference Patterns 11
3.2 Evidence Updates 13
4 Inference Engine and Heuristics 15
4.1 Heuristics and Biases 15
4.2 Inference Procedures 17
4.3 An Example 18
5 Discussion 20
6 Future Work 21
7 Concluding Remarks 23
References 24

Figures

1 Sample contents of working memory. 13
2 Coherence graph after three inference cycles. 18

Tables

1 Long-term memory contents for the alarm example. 10
2 Inference patterns and associated confidence updates. 12

1 Introduction

Reasoning plays a fundamental role in cognition and intelligence. It supports decision making, action, problem-solving, and learning by converting observations of the surrounding environment into a high-level representation of the current situation. Cognitive systems aimed at large and complex environments therefore require the support of a computationally efficient and scalable inference system to provide these interpretations. A central problem in computational inference concerns brittleness. Many inference techniques work well under a handful of well-defined conditions, but few can function outside their intended purpose. Each new appliction therefore requires careful revision of existing inference methods.

The artificial intelligence and psychological communities have both studied reasoning extensively. Artificial intelligence focuses on developing both powerful logic-based and uncertainty-based computational mechanisms for reasoning under a variety of circumstances. However, the resulting systems typically rely on proving or on carefully estimating probabilities for large numbers of beliefs at great computational expense. Resulting systems tend to ignore plausible but unsound conclusions, to process large batches of data instead of individual observations, and often do not scale well. Conversely, psychologists focus on describing at a high level the mechanisms that humans use for reasoning under specific circumstances. Most mechanisms therefore lack sufficient description for a computational implementation, or are too specific to apply to a general cognitive systems.

This report presents initial work on the Plausible Logic Inference Engine (PLIE), which seeks to combine results from both artificial intelligence and psychology into an incremental, pragmatic, and scalable computational reasoning system. Our approach includes three key elements. First, it combines first-order logic with uncertainty based on a notion of plausibility to provide the system with a flexible knowledge representation. Second, the system includes an inference mechanism that integrates deductive and abductive methods to provide a more robust inference capability than either method can provide alone. Importantly, the inference mechanism also supports incremental update of individual beliefs. Finally, PLIE includes a guidance mechanism based on biases identified in the psychology literature to determine which inferences the system should pursue.

Our presentation begins by describing the knowledge representation and specific inference patterns applied by the inference engine. Next we discuss the inference process along with the heuristics used it guide it. We illustrate key ideas throughout using an expanded version of Pearl's (1988) burglar alarm example. We also discuss the relationship of our approach to cognitive systems in general throughout the presentation. Later, we outline the relationship of our proposed techniques to other inference methods and discuss both roadblocks and next steps toward realizing the stated goals. Some of these include constructing appropriate testbeds, extensions to analogical reasoning, concept and structure learning, and metacognitive abilities such as bias learning.

2 Knowledge Representation

Combining first-order logic with uncertainty has long been a goal of artificial intelligence. First-order logic provides a compact representation and efficient inference techniques, but it does not handle the uncertainties present in the real world. Probabilistic methods naturally capture and reason about this uncertainty, but they do not represent or exploit the relational structure of the world.

Recent efforts at combining the two, such as Markov logic (Richardson & Domingos, 2006), have shown success. However, such methods rely heavily on complex statistical computation which has not been shown to scale up. More fundamentally, combined methods such as Markov logic depend heavily on translating first-order knowledge into grounded literals, which forfeits any advantage given by the original, relational representation.

By contrast, psychologists have relatively little to say about mechanisms that combine these two capabilities. While there is evidence that humans maintain and process both relational patterns (schemata for example, Bartlett (1932)) and estimates of confidence in their beliefs (Shynkaruk & Thompson, 2006), few specific mechanisms have been proposed. Broad theories of reasoning based on both logic (Rips, 1994) and probability (Oaksford & Chater, 2007) have been advanced, but little progress has been made toward unifying the two.

This work uses a combination of logic and uncertainty based on plausibility as first described by Polya (1954), who codified the methods used for hypothesizing theorems and guiding the search for the associated proofs. Polya's approach was based on a qualitative notion of accumulating confidence from evidence, and is largely consistent with Dempster-Shafer theory (Shafer, 1976), a mathematical theory of evidence that allows one to combine many sources of evidence into a degree of belief. Our approach replaces the qualitative account of confidence with a quantitative one, but retains the important properties highlighted by Polya. The result is similar to MYCIN (Shortliffe & Buchanan, 1975) and to Friedman's (1981) work on plausible inference.

2.1 Beliefs and Working Memory

PLIE represents knowledge using a combination of predicate logic and uncertainty. A *predicate* represents a generalized concept, or equivalently, a class of environmental situations. Each predicate may include a list of arguments. For example, Alarm(x) represents the situation in which the burglar alarm at location x is sounding.

Beliefs represent specific groundings or instances of a predicate such that each variable in the predicate's argument list is bound to some domain constant. For example, Alarm(BOB) indicates a belief that the alarm at Bob's house has sounded. Observations similarly correspond to grounded predicates, but come from the perceptual process rather than the inference process. This work does not speculate as to the nature of perception; we assume that some process produces the needed results.

Each belief, b, also has a plausibility score composed of the difference in evidence for and against the belief, $\mathrm{Pl}(b) = E^+(b) - E^-(b)$, where $0 \leq E^+(b), E^-(b) \leq 1$. The methods for calculating these evidence values are described later. Plausibility scores lie in the interval $[-1, 1]$ such that $\mathrm{Pl}(b) = 1$ indicates a strong belief in b while $\mathrm{Pl}(b) = -1$ indicates a strong belief against b. Importantly, the plausibility of a belief is tied to the plausibility of its logical negation, $\mathrm{Pl}(b) = -\mathrm{Pl}(\neg b)$, preventing contradictions. $\mathrm{Pl}(b) = 0$ confounds the cases in which there is no evidence in either direction (default state of all beliefs) and in which the evidence for and against a belief cancel each other. From a decision-making point of view, these two cases may be reasonably viewed as equivalent, but for cases in which distinguishing between no evidence and conflicting evidence is important, the evidence values can be used directly.

PLIE stores both beliefs and observations in its *working memory* as nodes in a directed graph. The edges of the graph represent derivational information, and stem from the inference process. Specifically, a directed edge from belief b_1 to b_2 indicates that b_1 formed part of the evidence used to conclude (or explain) b_2. Thus, observations have only outward directed edges, while all inferred beliefs have at least one edge directed inward.

2.2 Rules and Long-Term Memory

Graph edges derive from the *rules* stored in *long-term memory*. Rules represent the relationships among predicates that can be used to derive new beliefs. From this perspective, the contents of working memory may be viewed as activated structures from long-term memory. This is consistent with Cowan's (1988) view that working memory is an extension of long-term memory, and that the number of activated long-term structures is not limited, although the number of items in the focus of attention is limited.

Rules in PLIE take the form

$$p_1(\cdot),\ p_2(\cdot),\ \cdots,\ p_m(\cdot)\ \Rrightarrow q_1(\cdot),\ q_2(\cdot),\ \cdots,\ q_n(\cdot)\ ,$$

where $p_i(\cdot)$ and $q_j(\cdot)$ represent predicates, and $\Rrightarrow$ denotes *threshold-implication* (Stracuzzi & Könik, 2008). Threshold-implication defines the consequent to be true if the linear-threshold function, $\sum_{i=0}^{m} v_i \mathrm{Pl}(p_i) > 0$, is satisfied given that v_i represents the weights associated with each term in the antecedent (w_i for the consequent), and $\mathrm{Pl}(p_0) = \mathrm{Pl}(q_0) = 1$ by definition. Note that $p_i(\cdot)$ and $q_i(\cdot)$ represent general predicates while p_i and q_i denote the specific groundings used to instantiate a rule, which PLIE identifies by pattern matching from working memory. In this work, both sides of the implication represent linear threshold functions, so when the antecedent function is satisfied, the system updates beliefs in an effort to satisfy the consequent function.

Each implication also has two associated parameters that represent rule strength. For a rule $\mathsf{A} \Rrightarrow \mathsf{B}$, the parameter $\alpha \approx \Pr(\mathsf{B}|\mathsf{A})$ represents the reliability of the rule when viewed deductively, meaning that belief in A implies a belief in B. Similarly, $\beta \approx \Pr(\mathsf{A}|\mathsf{B})$ represents the reliability of the rule when viewed abductively, meaning that belief in B allows the assumption of A. Here, $0 \leq \alpha, \beta \leq 1$, though the two are not complimentary. In practice these parameters will initially take default values, which the system can later modify based on experience.

(1)	$\mathsf{Burglar}(x_1) \overset{\alpha=0.95}{\underset{\beta=0.40}{\Rightarrow}} \mathsf{Alarm}(x_1)$
(2)	$\mathsf{Earthquake}(x_2) \overset{\alpha=0.30}{\underset{\beta=0.60}{\Rightarrow}} \mathsf{Alarm}(x_2)$
(3)	$\mathsf{Alarm}(x_3) \wedge \mathsf{Neighbor}(y_3, x_3) \wedge \mathsf{Quiet}(y_3) \overset{\alpha=0.90}{\underset{\beta=0.05}{\Rightarrow}} \mathsf{Calls}(y_3, x_3)$
(4)	$\mathsf{Alarm}(x_4) \wedge \mathsf{Neighbor}(y_4, x_4) \wedge \neg\mathsf{Quiet}(y_4) \overset{\alpha=0.70}{\underset{\beta=0.5}{\Rightarrow}} \mathsf{Calls}(y_4, x_4)$
(5)	$\mathsf{PhoneRings}(x_5) \wedge \mathsf{Neighbor}(y_5, x_5) \wedge \mathsf{Quiet}(y_5) \overset{\alpha=0.10}{\underset{\beta=0.95}{\Rightarrow}} \mathsf{Calls}(y_5, x_5)$
(6)	$\mathsf{SecurityCo}(y_6, x_6) \wedge \mathsf{Calls}(y_6, x_6) \overset{\alpha=1.0}{\underset{\beta=1.0}{\Rightarrow}} \mathsf{Burglar}(x_6)$

Table 1. Long-term memory contents for the alarm example.

Table 1 shows the rules for the burglar alarm example. For the sake of clarity, they are expressed using Boolean connectives rather than in threshold logic. For example, PLIE could represent the antecedent of rule (6) as

$$1 \cdot \mathrm{Pl}\big(\mathsf{SecurityCo(y, x)}\big) + 1 \cdot \mathrm{Pl}\big(\mathsf{Calls(y, x)}\big) - 1.5 > 0 \; .$$

Threshold logic provides several important advantages. First, it provides a more powerful representation than Horn clauses, including conjunctions, disjunctions, and many other simple functions such as $p_1 \wedge (p_2 \vee p_3)$. Second, the underlying numeric representation (the weights) also provides a natural way to handle numeric (as opposed to symbolic) domain constants. Finally, with respect to the long-term goals of this work, threshold logic provides a strong foundation for rule learning. Specifically, the weights associated with each rule term creates a continuous relationship among possible rule definitions, which provides for a much finer level of control than typically available in relational representations. Stracuzzi and Könik (2008) demonstrate how to learn a set of hierarchically structured predicate definitions based on threshold logic functions from sparse, supervised examples.

3 Inference Patterns and Confidence Propagation

Researchers in artificial intelligence have considered inference methods extensively, though often in a piecemeal fashion. As noted, the integration of logical and statistical inference has only recently come to the fore with methods such as Markov Logic (Richardson & Domingos, 2006) and Bayesian Logic Programs (Kersting & De Raedt, 2005). The integration of multiple inference patterns, such as deduction, abduction, and analogy has received even less attention. For example, Kate and Mooney (2009) applied the deductive inference engine included with the Markov logic framework abductively. However, the two methods rely on different rule structures so that the user must decide which method to apply before encoding the domain rules.

In this report, we consider the combination of deduction and abduction. The utility of combining the two follows from their complementary nature. Deduction produces sound arguments (in which the conclusion necessarily holds given the evidence), but can only restate existing knowledge in different terms. Abduction adds new information to the system by hypothesizing (unsound) explanations for existing knowledge, which deductive methods can then further expand upon. The resulting system should therefore reason more broadly than one based on either method alone.

3.1 Inference Patterns

PLIE relies on four distinct inference patterns for constructing new beliefs. Each pattern, shown in Table 2, applies to a specific situation depending on the available evidence. For example, *modus ponens* deductively concludes the consequents (q_j) of a rule given that the antecedents (p_i) hold. In the context of threshold-implication, this means that the evidence equation, $\sum_{i=0}^{m} v_i \mathrm{Pl}(p_i) > 0$, must be satisfied by pattern matching beliefs from working memory to the predicates in the rule antecedent. When successful, PLIE concludes that $\sum_{j=0}^{n} w_j \mathrm{Pl}(q_j) > 0$ must also hold. In practice, this means that the system increases the plausibility in each of the rule's consequent beliefs, q_i. The other inference patterns operate analogously depending on which beliefs get used as evidence.

The four inference patterns have differing characteristics. Two of them, modus ponens and modus tollens, are logically sound, deductive rules. A third, denying the antecedent, is also deductive in nature though not logically sound because it ignores the possibility that a conclusion may be drawn in multiple ways. Nevertheless, logically unsound inference patterns are still useful for accumulating evidence about beliefs (Polya, 1954) and for considering beliefs that, while not provable, may still accurately describe an agent's environment. Finally, affirming the consequent serves as the basis for abduction, which is also not logically sound for similar reasons. Notice also that two of the patterns deal with inference from positive evidence (plausibilities sufficiently greater than zero), while two of them deal with negated evidence (plausibilities less than zero). This lets PLIE handle a broad array of situations and rule structures.

Consider the following application of the inference patterns to the rules shown in Table 1. For now, we consider only the symbolic application of patterns, but return in the next section to describe the plausibility updates. Suppose that working memory initially contains the observations

Pattern	*Symbolic Form*	*Evidence Update Equation*
Modus Ponens • *Deductive* • *Increases evidence for* q_i	$p_1(\cdot),\cdots,p_m(\cdot) \Rrightarrow q_1(\cdot),\cdots,q_n(\cdot)$ $\frac{\sum_{i=0}^{m} v_i \mathrm{Pl}(p_i) > 0}{\sum_{j=0}^{n} w_j \mathrm{Pl}(q_j) > 0}$	$\Delta E^{+}(q_j) = \alpha \frac{\sum_{i \in S} v_i \mathrm{Pl}(p_i)}{\sum_{i \in S} \lvert v_j \rvert}$
Modus Tollens • *Deductive* • *Increases evidence against* p_i	$p_1(\cdot),\cdots,p_m(\cdot) \Rrightarrow q_1(\cdot),\cdots,q_n(\cdot)$ $\frac{\sum_{j=0}^{n} w_j \mathrm{Pl}(q_j) < 0}{\sum_{i=0}^{m} v_i \mathrm{Pl}(p_i) < 0}$	$\Delta E^{-}(p_i) = -\alpha \frac{\sum_{j \in S} w_j \mathrm{Pl}(q_j)}{\sum_{j \in S} \lvert w_j \rvert}$
Denying the Antecedent • *Deductive* • *Increases evidence against* q_i	$p_1(\cdot),\cdots,p_m(\cdot) \Rrightarrow q_1(\cdot),\cdots,q_n(\cdot)$ $\frac{\sum_{i=0}^{m} v_i \mathrm{Pl}(p_i) < 0}{\sum_{j=0}^{n} w_j \mathrm{Pl}(q_j) < 0}$	$\Delta E^{-}(q_j) = -\alpha \frac{\sum_{i \in S} v_i \mathrm{Pl}(p_i)}{\sum_{i \in S} \lvert v_j \rvert}$
Accepting the Consequent • *Abductive* • *Increases evidence for* p_i	$p_1(\cdot),\cdots,p_m(\cdot) \Rrightarrow q_1(\cdot),\cdots,q_n(\cdot)$ $\frac{\sum_{j=0}^{n} w_j \mathrm{Pl}(q_j) > 0}{\sum_{i=0}^{m} v_i \mathrm{Pl}(p_i) > 0}$	$\Delta E^{+}(p_i) = \beta \frac{\sum_{j \in S} w_j \mathrm{Pl}(q_j)}{\sum_{j \in S} \lvert w_j \rvert}$

Table 2. Inference patterns and associated confidence updates. For the symbolic forms, note that the implication (first line) represents a rule from long-term memory, the inequality above the horizontal bar represents the evidence required for the pattern to apply, and the inequality below the bar represents the conclusion drawn given the pattern, the rule, and the evidence.

Quiet(JOHN) and Calls(JOHN, BOB), each with plausibilities of 1.0. The system can then apply rules (3) and (5) with accepting the consequent to conclude (or assume) Alarm(BOB) (rule 3), PhoneRings(BOB) (rule 5), and Neighbor(JOHN, BOB) (both rules). Figure 1 shows the coherence graph that would result if the system applied both rules. Notice that Calls(JOHN, BOB) explains Quiet(JOHN). The opposite is not true in this case because the antecedents to rules (3) and (5) are not satisfied, which prevents PLIE from applying the rules with modus ponens.

Similarly, rule (4) does not apply in this case. Although the consequent matches with the observation Calls(JOHN, BOB) in working memory, the antecedent term ¬Quiet(JOHN) contradicts the observation Quiet(JOHN). For now, we assume that PLIE cannot change the plausibility of observations (it trusts its senses). However, the assumption is not fundamental, and may be dropped in later versions of the work. Rule (6) also does not apply because, in the context of working memory contents, it fits the modus ponens pattern, which requires that the antecedent be satisfied.

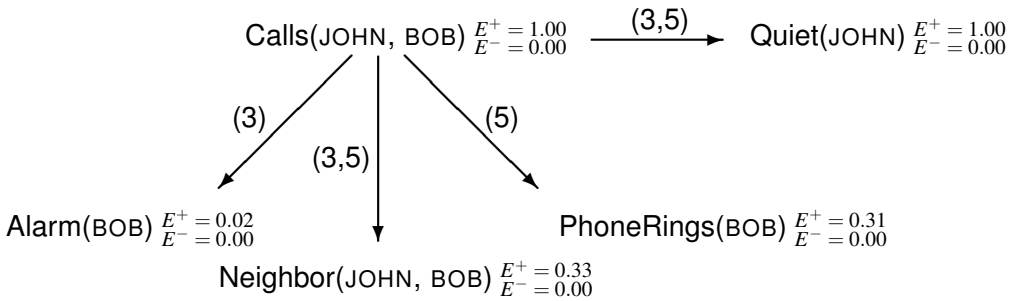

Figure 1. Contents of working memory generated by applying the inference patterns and evidence updates from Table 2 to the long-term memory contents in Table 1. Edges indicate derivation, while edge labels indicate the rules used to derive the beliefs.

3.2 Evidence Updates

As discussed earlier, plausibility quantifies the system's belief in specific aspects of the environment. PLIE represents and propagates any uncertainty associated with its rules and observations through the inference process. This is particularly important in the context of abduction, which produces logically unsound assumptions. Some of these may be reliable, while others may not. Quantifying this uncertainty lets PLIE distinguish between these two cases and focus on working with more plausible conclusions during subsequent inference steps. This is an essential aspect of maintaining tractability in complex domains.

The update equations shown in Table 2 implement this by first combining the plausibility values used to trigger the rule, and then dividing it among the conclusions. The summation factor in the equations determines the amount of evidence that gets propagated by performing a weighted sum over the plausibility scores associated with the matched beliefs. However, recall that threshold logic can represent disjunctive (or partially disjunctive (as in $p_1 \wedge (p_2 \vee p_3)$) rule bodies. This means that not all of the terms in the evidence must be satisfied for the rule to apply.

A belief b satisfies a rule term if $\text{sign}(w_b \text{Pl}(b))$ is positive for the inference patterns that increase confidence in conclusions (b contributes to exceeding the threshold), and negative for patterns that decrease confidence in conclusions (b contributes to *not* exceeding the threshold). We define S as the set of beliefs used to instantiate the *satisfied* terms in the evidence of a rule application. For example, suppose $p_1 \wedge (p_2 \vee p_3)$ represents the evidence for an inference step and that PLIE has p_1 and p_3 in working memory with positive confidence, and p_2 with negative confidence. Applying modus ponens would produce $S = \{p_1, p_3\}$, while denying the antecedent would produce $S = \{p_2\}$ (and fail).

Before PLIE propagates the accumulated update value to the conclusions, the value first gets scaled by the reliability of the applied rule (α or β) as shown in Table 2. This lends greater confidence to conclusions drawn from reliable rules. Note also the sign of the evidence update,

which is negated for inference patterns that increase evidence against a belief. In these cases, the weighted sum of the evidence values will be negative because the sign of the weighted plausibilities is negative. Evidence values are always positive however, so the sign needs to be flipped.

Finally, the evidence score for a belief q_i is updated by

$$E_{t+1}(q_i) = (1 - \lambda \hat{w}_i)E_t(q_i) + \lambda \hat{w}_i \Delta E_t(q_i) \tag{1}$$

where $\hat{w}_i = \frac{|w_i|}{\|w\|}$ represents the normalized weight magnitude associated with q_i in the rule, and $0 < \lambda < 1$ represents a recency parameter (discussed below). The equation for updating p_i is analogous. Note that we omitted the positive and negative superscripts from E; Table 2 shows which evidence value gets updated by a given inference pattern. However, if the q_i term is negated in the rule ($w_i < 0$), then the update gets switched from E^+ to E^- or vice versa. This follows from the plausibility relationship $\mathrm{Pl}(b) = -\mathrm{Pl}(\neg b)$ between beliefs and their negations.

Equation 1 supports a form of the primacy-recency bias (Ebbinghaus, 1913) by scaling the existing and updated evidence scores. PLIE implements primacy by setting $\lambda = 1$ on the first evidence update for each belief. This causes the initial evidence for a belief to have an outsized effect compared to subsequent evidence (typically $\lambda < 1$, which scales down the impact of new evidence). New evidence will eventually dominate the older evidence (a recency effect), but this may require a sequence of several updates. Setting λ near 1 accelerates this process by increasing the impact of new evidence, while λ near 0 extends the primacy effect. Psychologists view these as memory biases, but this work assumes that biases in how beliefs and their supporting evidence get stored impact the reasoning process. Along similar lines, rule terms with relatively large weight magnitudes receive higher impact from new evidence than those with low weight magnitude, as these contribute the most to satisfying the rule.

Returning now to Figure 1, notice that the plausibility associated with Alarm(BOB) is very low at 0.04 (assuming that $\lambda = 1$ and the weight on each rule term is 1). This follows from the low β value associated with rule (3), implying that John has a high false-positive rate. The plausibility of PhoneRings(BOB) is much higher, indicating that John's call is a much better indicator of Bob's phone ringing than of his burglar alarm's sounding. Rules (3) and (5) both contribute to the plausibility of Neighbor(JOHN, BOB) (assume that the rules were applied in the order listed in Figure 1).

Upon initial review, the inference patterns and evidence updates may appear complex. However, closer inspection reveals that all of the patterns and equations represent minor variations on a general mechanism depending on whether the inference flows with or against the rule's implication sign, and whether the evidence increases the evidence for or against a belief. This closely resembles Friedman's (1981) work, except that in switching from Boolean to threshold logic, we have collapsed several highly specialized cases into one general case.

4 Inference Engine and Heuristics

The inference patterns described above govern the details of how PLIE combines beliefs with rules to derive new beliefs. Given these patterns, the inference engine first selects beliefs and rules from memory, and then applies the patterns to construct new beliefs and update plausibility scores. Importantly, the incremental application of the inference patterns from Table 2 provides a seamless integration of forward and backward reasoning methods, allowing PLIE to exploit opportunities in reasoning as they become available. A similar idea appears in the planning community as opportunistic planning, or metaplanning (see work by Hayes-Roth & Hayes-Roth, 1979, for an early example). In this section, we describe the details of the inference engine, including the methods used to select beliefs and rules for inference.

4.1 Heuristics and Biases

This work aims to produce a scalable inference engine that, while neither complete nor sound, performs well in practice and provides a foundation for other cognitive abilities. At present, PLIE does not have access to an agent's goals, so our approach focuses on properties of beliefs based on three cognitive biases, and one that is more computational in nature. The cognitive biases specifically impact reasoning and have amassed substantial evidence in their favor. We do not claim to reproduce the exact conditions that lead to these biases as identified in the literature. Instead, we focus on incorporating a generalized view of each bias into our computational reasoning system. Incorporating other biases and heuristics remains a key area of future work.

The first cognitive bias, explanatory coherence (Thagard, 1989), implies a preference for beliefs that "hang together" in the context of explaining observations. Although often associated with abduction, we also apply coherence to deduction, as PLIE may also make unsound deductive inferences.

PLIE implements coherence as a measure of linkage with emphasis on the plausibility of the linked beliefs and their proximity to observations. Let $\text{adj}(b)$ represent the set of beliefs adjacent to b in working memory regardless of edge direction. For a belief $b' \in \text{adj}(b)$, let $\text{rel}(b', b)$ represent the rule reliability (α or β) that applies when the system views the rule that links b with b' as though it were used to derive b from b' (see Table 2). Also let $\delta(b')$ represent the shortest distance from b' to any observation in working memory (0 if b' is an observation).

Given these definitions, we define coherence as

$$\text{coh}(b) = \sum_{b' \in \text{adj}(b)} \frac{\hat{w}_{b'} \, \text{rel}(b', b)}{\delta(b') + 1} \text{Pl}(b') \; .$$

Though related to plausibility in its use of belief weights and rule reliabilities, coherence is a measure of the total evidence in working memory for b, with each piece of evidence weighted by its proximity to an observation. In practice, a unit increment in the coherence measure represents direct evidence from or explanation of one (high plausibility) observation through a reliable rule.

The measure attenuates as evidence gets divided over multiple rule terms, rules become unreliable, observations grow more distant, and evidence terms become less plausible. As a result, coherence is unbounded in magnitude and cannot be computed incrementally. Nevertheless, the coherence measure is computationally simple and calculated locally, so the approach should remain tractable as the size of working memory grows large. Our coherence measure is similar in spirit to that of Ng and Mooney (1990), but operates over individual beliefs rather than proof trees.

The second cognitive bias implements a preference for working with highly plausible beliefs. We view this as a generalization of several specific biases identified in the literature, such as the confirmation and disconfirmation biases (Lord et al., 1979). The former states that people tend to favor information that confirms their preconceptions, while the latter states that people tend to be very critical of information that contradicts them.

PLIE implements these in two ways. First, it ignores beliefs that have low plausibility. In the alarm example, an agent may ignore the contingency in which the alarm has sounded if John's unreliability causes a sufficiently low plausibility score and no other evidence is available to increase it. Second, PLIE prefers rule applications with stronger evidence over those with weaker evidence. The extent to which a rule application exceeds its threshold indicates how strongly the evidence supports the rule application. This favors rule instances with more matched evidence terms, and whose evidence terms have high plausibility.

The preference for working with highly plausible beliefs comes with two caveats. First, the approach may ignore a plausible inference if evidence is spread thinly across many rule applications. In cases for which no one rule application provides sufficient evidence to maintain a belief, and the inference procedure does not apply multiple rules for the belief in a single cycle, the belief will be ignored repeatedly. The extent to which this is acceptable behavior requires empirical evaluation. The second caveat is that the inference procedure could revisit plausible beliefs repeatedly, causing them to become more plausible without any new evidence. To control this, the inference procedure is explicitly prevented from revisiting previously made inferences as indicated by the coherence graph.

The third cognitive bias applies a second form of recency in the inference engine by preferring to expand upon beliefs and observations that have been recently updated or added to working memory. Recall that primacy and recency biases have already been implemented by the plausibility update equations. This gets implemented by having the inference engine select only a small number of beliefs for expansion during any given inference cycle based largely on their the recency.

The computational bias implemented in the inference engine borrows from Fischer et al's (1991) work on essential explanation. Here, the system prefers beliefs that unify with fewer rules in long-term memory. The underlying insight is that choosing from among fewer options means that fewer incorrect choices are possible, and the computational cost of evaluating each option will be lower overall. Like coherence, essential explanation was developed in the context of abductive inference, although PLIE uses it for guiding deduction as well.

This combination of biases should cause PLIE to reason along a small number of plausible trajectories. Nothing explicitly prevents the system from exploring other lines of inference opened

by new observations or substantial changes to the plausibility of established beliefs. However, it should not explore broadly through highly implausible regions of the belief-space.

4.2 Inference Procedures

PLIE operates in cycles, with each cycle consisting of three steps. First, the system selects a small subset of beliefs from working memory for expansion and identifies the relevant rules from long-term memory. Next, it applies the inference patterns from Table 2 to derive new and update existing beliefs. Finally, it filters the resulting beliefs to remove those with low plausibility or coherence. The remainder of this section details each step in turn.

Each cycle begins with PLIE applying the essential explanations bias to select a small subset of beliefs, W_{sel}, from working memory for expansion. The number of selected beliefs depends on the number of beliefs having the same number of essential explanations (all beliefs with the minimum count are selected). PLIE then identifies the applicable rules from long-term memory given the beliefs in W_{sel}. A rule is applicable if three conditions are met: (1) a belief from W_{sel} matches a term in the rule evidence, (2) the remaining evidence terms can be matched by beliefs in working memory, and (3) the intended inference has not been previously made using the same rule and evidence.

After identifying the set of applicable rules, PLIE then sorts them according to the change in belief coherence that applying the rules would effect. The system prefers larger improvements to coherence over smaller improvements or reductions. Finally, PLIE applies the inference patterns from Table 2 to generate conclusions and update plausibility scores. This may entail multiple rule applications and may generate many conclusions, or multiple updates to a single conclusion. Though the search remains focused on recent, coherent beliefs, this provides a broader search through belief space than if the system used only the single "best" belief in each cycle.

Also note that conclusions drawn abductively by PLIE may contain unbound variables, such as $\mathsf{Neighbor}(\mathsf{y}_3, \mathsf{BOB})$. Here, the system creates a Skolem constant as a place holder. During subsequent inference cycles, it can unify Skolem constants with other constants. For example, this would effectively replace the belief "Bob has some specific, unnamed neighbor" with "Bob's neighbor is John." All instances of a Skolem constant must be replaced with the same domain constant and all unifications must be consistent (must not create a belief that contradicts any existing belief).

After generating conclusions and updating plausibility scores, PLIE filters its results to remove beliefs with very low magnitude plausibility. Specifically, any belief b with $|\mathrm{Pl}(b)| < \tau$ does not get added to (or replaced in) working memory, where τ is an adjustable parameter. This prevents it from creating and storing a large number of beliefs with scores very near to zero. Although PLIE is unlikely to select such beliefs for W_{sel}, they may still get matched as part of the inference process, thereby deriving a potentially large number of new beliefs with low plausibility.

An important side effect of this is that some beliefs may lose their support over time. The plausibility of such a belief does not change in working memory, but the coherence declines, which

biases PLIE against selecting the belief for expansion. Plausibility changes only when the system draws a direct conclusion about a belief.

4.3 An Example

Consider now a larger example based on the rules in Table 1. To begin, suppose that PLIE receives four observations, each with plausibility 1.0: Neighbor(JOHN,BOB), Quiet(JOHN), SecurityCo(XYZ,BOB), and Calls(XYZ,BOB). Initially, the only possible inference applies rule (6) with modus ponens to conclude that there is a burglar at Bob's house, Burglar(BOB), as shown in the coherence graph in Figure 2. The high plausibility of the conclusion follows from the use of both highly plausible observations as evidence, and a highly reliable inference rule ($\alpha = 1.0$).

In the second cycle, PLIE chains forward from Burglar using rule (1) with modus ponens to conclude that the alarm has sounded at Bob's house, Alarm(BOB). This conclusion is also highly plausible for reasons similar to those above. Note however that the coherence of Burglar (1.13) is substantially higher than that of Alarm (0.48) because Burglar links directly to observations.

PLIE has available two applicable rule instances in the third cycle given that W_{sel} contains Alarm(BOB). Rule (2) applies abductively and rule (3) applies with modus ponens. Assuming similar thresholds for the two rules, PLIE would prefer rule (3) over rule (2) based on the slightly larger

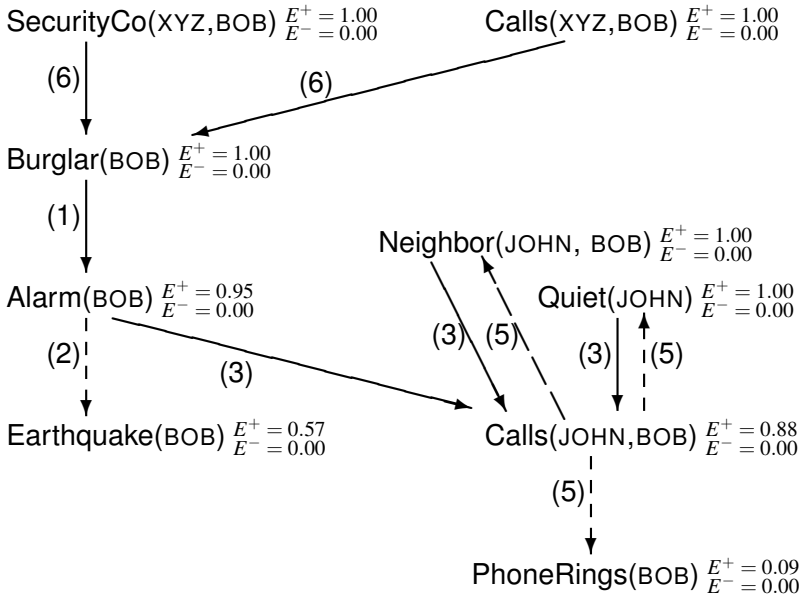

Figure 2. Coherence graph after three inference cycles given the observations Neighbor(JOHN,BOB), Quiet(JOHN), SecurityCo(XYZ,BOB), and Calls(XYZ,BOB). Dashed arrows indicate available inferences not made by PLIE.

plausibility contributions, and therefore coherence improvements, from Neighbor and Quiet (1.0 versus 0.95 from Alarm). PLIE therefore infers Calls(JOHN,BOB) with coherence 1.16 based on its links to two observations. The inference also raises the coherence of Alarm slightly to 0.50. The small increment in this case follows from the low reliability of rule (3) when viewed abductively ($\beta = 0.05$).

To summarize the inference process so far, PLIE has determined that a burglar has very likely struck Bob's house given that the XYZ security company has called him. His alarm system has also plausibly sounded, which implies that John may also call soon. PLIE avoided the inference that an earthquake triggered the alarm, though, even if made, the inference would be both less plausible and less coherent than the burglar explanation. Nevertheless, this does raise a question about whether preferences and coherence are sufficient to avoid or distinguish between multiple, conflicting explanations for a single set of observations. PLIE does not explicitly represent the notion that, while not technically mutually exclusive, rules (1) and (2) should not typically occur together. For now we reserve this as a point for future work.

A second open question relates to a stopping criterion. For cases in which the stream of incoming observations stagnates for a period of time, PLIE would ideally determine a point at which continuing inference is no longer productive. In the current example, PLIE could continue reasoning by applying rule (5) abductively to generate the assumption PhoneRings(BOB). As with Earthquake, this represents an alternate, and less plausible explanation for John calling Bob. If other rules related to events or beliefs stemming from the alarm's sounding or John's call to Bob were present in long-term memory, then the system would tend to follow those paths. In this case, no other inference paths are available, so PLIE returns to considering the low-plausibility and low-coherence lines of reasoning ignored earlier.

5 Discussion

The plausibility-based approach to computational reasoning outlined above represents a distinct departure from traditional proof-based logic. It provides mechanisms for drawing unsound conclusions, and relies on incomplete heuristic search. Although this admits the possibility that the system may draw incorrect conclusions or ignore many correct conclusions, it also represents an attempt at avoiding the intractability associated with many logical inference systems by identifying potentially useful conclusions. This is an important point in the context of cognitive agents whose goal is to perform tasks in complex environments.

From this perspective, adding information about agent goals to PLIE would be a fruitful direction for future work. For example, PLIE could use the relationship of a belief's predicate to goal predicates as an additional guidance heuristic. This respects the approach of exploring from observations (as well as goals) while further sharpening the system's focus on working with beliefs that are relevant to the task or domain at hand.

In the context of symbolic inference, Bridewell and Langley's (2011) AbRA is most similar to PLIE. Although their discussion focuses on abduction, AbRA can also apply rules deductively. Both systems can therefore combine both forward chaining from observations and backward chaining from goals. Their approach does not support inference from negated evidence (modus tollens and denying the antecedent) and does not handle uncertainty in beliefs or rules, however. Friedman's (1981) work does support all four inference patterns identified in Table 2 and uses a numeric confidence measure that is somewhat similar to PLIE's plausibility scores, but relies on brute force computation to derive all possible beliefs.

Production systems such as Soar (Laird et al., 1987) also perform a kind of symbolic inference. Unlike many other inference systems, production systems do not commit to any one reasoning formalism, although the rules can be structured to implement a variety of formalisms. Instead, they take a purely syntactic approach, firing rules that match the current belief state and using other productions to implement preferences while typically ignoring the structural properties of the generated beliefs. In contrast, PLIE commits to a fixed reasoning formalism based on deduction and abduction while its guidance heuristics take the relationships among beliefs into consideration.

PLIE also represents a departure from traditional statistical optimization-based inference techniques. These methods, including Bayesian networks (Pearl, 1985) and more recent efforts like Markov logic (Richardson & Domingos, 2006), attempt to estimate probability distributions associated with one or more query variables. Statistical inference algorithms can require substantial computation, often drawing on information from large portions of the underlying belief network in response to a single query, or evaluating every possible belief simultaneously. Conversely, PLIE updates selected plausibilities incrementally in response to changes in observations. Though clearly less precise than statistical inference methods, PLIE entails a much lighter computational burden. This makes sense in the context of agents acting in worlds where responsiveness often takes precedence over precision.

6 Future Work

One of the major hurdles encountered so far relates to identifying appropriate testbeds for evaluation. Both the artificial intelligence and psychology research communities have developed a number of tasks for use in studying reasoning. Examples include the challenge problems for commonsense reasoning[1] from artificial intelligence, and the well-studied Wason selection task (Wason, 1966) from psychology. The problem with these tasks and others like them is that they were crafted to highlight very limited reasoning capabilities. Many of the reasoning problems used by artificial intelligence researchers were designed with pure deductive or abductive reasoners in mind. Likewise, the Wason selection task was designed to study very simple deductive reasoning in humans. As a result, none of these tasks are appropriate to evaluating a more general reasoning engine such as PLIE.

The examples above suggest that PLIE may demonstrate a number of interesting and useful properties, but a much more extensive evaluation is required. Of specific concern are the system parameters and the inference heuristics. With additional experimentation, the parameters should be set to fixed, domain independent values. Likewise, the impact of the inference heuristics requires further study. We selected plausibility, coherence, and primacy/recency because they are well-known and easy to implement. However, other implementations are possible and many other cognitive biases have been studied. The impact of these and other biases on system performance needs further study.

Given these points, continued development of PLIE requires the context of real-world applications. For example, aerial image analysis and cyber defense both focus on inferring activity and intent from sequential data. However, machine learning techniques have shown limited success at these tasks because context and knowledge of the local environment play a major role in distinguishing between normal and malicious activity. Accounting for this contextual information in a "big data" environment is a potential strength of the plausible logic approach.

We envision a system for collaborative human-computer data analysis in which the automated reasoner exchanges information with the human analyst in a continuous feedback loop. The reasoner provides the computational power needed to insulate the analyst from high data volumes, but also models analyst knowledge and methods to support productive, high-level interactions. The goal is then to improve analyst accuracy and throughput by combining both computational speed with human ability to identify novel situations.

As noted earlier, analogical and inductive reasoning processes will play major roles in producing a system that scales and performs well on real world problems, such as aerial image analysis and cyber defense. Briefly, the role of analogy is to expand the applicability of the rules in long-term memory. By mapping known rules into novel domains, PLIE can leverage a greater proportion of its knowledge when it encounters new situations.

Adding analogical reasoning to PLIE also offers potential for interdisciplinary collaborations. For example, analogical reasoning is also known by the term *transfer* or *transfer learning* in the

[1] http://www-formal.stanford.edu/leora/commonsense/

artificial intelligence community. Based on prior work in this area (Stracuzzi et al., 2011), we hypothesize that successful analogy depends on the beliefs and interpretations provided by deduction and abduction. This view differs from current psychological theories, which often hold that analogy serves as the fundamental basis for all reasoning. This difference in hypotheses supports a human subjects study in which analogical performance is compared between subjects that are asked to construct analogies directly and subjects that apply other forms of reasoning to the task first.

Like analogy, an inductive process will expand and add nuance to the system's existing knowledge. This is particularly important in the context of complex domains such as aerial image analysis and cyber security that may rely heavily on large stores of background knowledge. Inductive modifications to long-term memory include tuning rule strengths (α and β), modifying rule antecedents and consequents, adding new rules based on experience, and organizing the contents of long-term memory to improve the efficiency of reasoning and learning. Together, these should reduce PLIE's reliance on manual knowledge engineering and increase the breadth of its reasoning capability.

PLIE's learning methods will draw heavily on principles established in prior work on hierarchical knowledge base induction (Stracuzzi & Könik, 2008) and on concept learning in computational cognitive architectures (Li et al., 2011). Some of the principles advocated by the work on knowledge base induction have already been incorporated into PLIE's knowledge representation, as discussed at the end of Section 2. Others, such as the importance of knowledge organization must be added to the induction methods. Briefly, hierarchical rule organization helps to reduce learning complexity by reducing the number of candidate rule terms evaluated by the concept learning algorithm. This results in a smaller search. Learning in PLIE then becomes an incremental, bottom-up refinement of representation, which is well suited to the types of streaming data domains discussed above.

A more distant area of future work is metacognitive in nature. Currently, PLIE contains a set of four inference patterns, which may expand as analogical and inductive processes get added. These patterns are fixed, but could be adapted or expanded based on experience. For example, if the abductive pattern regularly produces conclusions that later get removed or refuted, then the system could modify the pattern require evidence with higher plausibility scores. This would increase it's efficiency by finding and exploiting general patterns in the reasoning process.

7 Concluding Remarks

The goal of this research is to combine ideas and results from a broad array of work in both the artificial intelligence and psychological research communities into a pragmatic and computationally tractable inference system. Our approach draws on experiences in both logical and statistical inference while dropping obsessions with formal proof and statistical precision. Likewise, even though our approach takes inspiration from research on human cognition, our objective is simply to identify reasoning heuristics that lead the inference engine to construct and maintain beliefs that are relevant to an agent acting in the environment.

The work reported here summarizes an initial effort at designing a knowledge representation and inference algorithms to achieve the stated goals. The provided examples demonstrate the intended function of the system, but further testing is required to establish that the intended performance holds up across a variety of domains. Nevertheless, combination of logic, plausibility, and heuristics presented above provide a strong foundation on which to experiment with, revise, and expand the core ideas of this work. Moreover, the Plausible Logic Inference Engine shows potential to support a number of interdisciplinary research endeavors, including clear connections to experimental psychology and computational neuroscience.

References

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.

Bridewell, W., & Langley, P. (2011). A computational account of everyday abductive inference. *Proceedings of the Thirty-Third Annual Meeting of the Cogntive Science Society*. Boston, MA: Cogntive Science Society, Inc.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*, 163–191.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University.

Fischer, O., Goel, A., Svirbely, J., & Smith, J. (1991). The role of essential explanation in abduction. *Artificial Intelligence in Medicine*, *3*, 181–191.

Friedman, L. (1981). Extended plausible inference. *Proceedings of the 7th International Joint Conference on Artificial Intelligence* (pp. 487–495). Vancouver, BC, Canada: William Kaufmann.

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science*, *30*, 275–310.

Kate, R. J., & Mooney, R. J. (2009). Probabilistic abduction using Markov logic networks. *Proceedings of the IJCAI-09 Workshop on Plan, Activity and Intent Recognition*. Pasadena, CA: AAAI Press.

Kersting, K., & De Raedt, L. (2005). Bayesian logic programming: Theory and tool. In L. Getoor and B. Taskar (Eds.), *An introduction to statistical relational learning*, 291–321. MIT Press.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

Li, N., Stracuzzi, D. J., & Langley, P. (2011). Improving acquisition of teleoreactive logic programs through representation change. *Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems*. Arlington, VA: AAAI Press.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Ng, H. T., & Mooney, R. J. (1990). On the role of coherence in abductive explanation. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 337–342). Boston, MA: AAAI Press.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings of the Seventh Conference of the Cognitive Science Society*. Irvine, CA.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan-Kaufmann.

Polya, G. (1954). *Mathematics and plausible reasoning, volume ii: Patterns of plausible inference*. Princeton, NJ: Princeton University Press.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Journal of Machine Learning Research*, *62*, 107–136.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medecine. *Mathematical Biosciences*, *23*, 351–379.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory and Cognition*, *34*, 619–632.

Stracuzzi, D., Fern, A., Ali, K., Hess, R., Pinto, J., Li, N., Konik, T., & Shapiro, D. (2011). An application of transfer to american football: From observation of raw video to control in a simulated environment. *AI Magazine*, *32*.

Stracuzzi, D. J., & Könik, T. (2008). A statistical approach to incremental induction of first-order hierarchical knowledge bases. *Proceedings of the Eighteenth International Conference on Inductive Logic Programming, LNAI 5194* (pp. 279–296). Prague, Czech Republic: Springer-Verlag.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*, 435–467.

Wason, P. (1966). Reasoning. In B. Foss and P. Dodwell (Eds.), *New horizons in psychology*. London: Penguin.

DISTRIBUTION:

1 MS 0899 Technical Library, 9536 (electronic copy)

Sandia National Laboratories