Beyond ISA: Structures for Plausible Inference in Semantic Networks

COINS Technical Report 88-20

Paul R. Cohen and **Cynthia L. Loiselle**

Experimental Knowledge Systems Laboratory
Department of Computer and Information Science
University of Massachusetts
Amherst, MA

Abstract

We present a method for automatically deriving plausible inference rules from relations in a knowledge base. We describe two empirical studies of these rules. First, we derived approximately 300 plausible inference rules, generated over 3000 specific inferences, and presented them to human subjects to discover which rules were plausible. The second study tested the hypothesis that the plausibility of these rules can be predicted by whether they obey a kind of transitivity. The paper analyzes four sources of variance in subjects' judgments, and concludes that relatively little knowledge is needed to achieve moderately accurate predictions of these judgments.

We are indebted to Carole Beal, David Day, and Adele Howe for their comments on drafts of this paper, to Carole Beal for her help with the statistical analysis, and Evan Smith for his assistance with this project.

This research was sponsored by ONR University Research Initiative grant N00014-86-K-1764 and by a gift from Tektronix.

1 Introduction

Can cough syrup make people drunk? Our favorite brand can, because it contains alcohol. If you didn't already know that cough syrup is intoxicating, you could infer it from two specific propositions—cough syrup contains alcohol and alcohol is intoxicating—and from a general plausible inference rule:

Rule 1

x	CONTAINS	y,	and
y	CAUSES	z	
x	CAUSES	z	

Other familiar rules of plausible inference include property inheritance (e.g., cats have five toes, Ginger is a cat, so Ginger has five toes) and causal abduction (e.g., fires cause smoke, so if you see smoke, look for a fire).

Rules like these have two roles that we expect to become increasingly important in coming years. First, they support *graceful degradation* of performance at the boundaries of our knowledge. A *brittle* knowledge system that doesn't know explicitly whether cough syrup makes you drunk won't offer a plausible answer—it simply won't answer the question [10,9,6]. Graceful degradation depends on general knowledge, which we formulate as as plausible inference rules such as Rule 1, to make up for a lack of specific knowledge. Second, we expect plausible inference to reduce the effort of building knowledge bases, because knowledge engineers needn't state explicitly those propositions that can be plausibly inferred. Property inheritance, for example, relieves us from having to state explicitly that each member of a class has each property of that class [2]. Rules like property inheritance and Rule 1 obviously are needed to build "mega-frame" knowledge bases [9].

Rule 1 has the same structure as property inheritance over ISA links, and can serve the same purposes, that is, supporting graceful degradation and knowledge engineering. We have developed a simple method for deriving such rules from the relations in a knowledge base, and we have shown how to differentiate plausible ones from implausible ones based on their underlying "deep structure."

This paper describes two empirical studies of these rules. Both depend on a moderately large knowledge base that we developed for the GRANT project [3,4,8]. The GRANT KB contains roughly 4500 nodes linked by 9 relations and their inverses. In the first study we derived approximately 300 plausible inference rules from these relations. Then we generated over 3000 specific inferences by replacing the variables in the rules with concepts from the GRANT KB, and presented them to human subjects to discover which syntactically permissible rules were plausible (Sec. 2). The second study tested the hypothesis that the plausibility of these rules can be predicted by whether they obey a kind of transitivity (Sec. 2.5). We will begin by describing these studies, hypotheses, and results. Then we will discuss the role of knowledge in assessing the plausibility of inferences.

2 Experiment 1: Identifying Plausible Rules

In this section we describe how to use the structure of property inheritance to produce many other plausible inference rules, and how we determined the plausibility of these rules.

2.1 Background

Property inheritance over ISA links can be written

$$\begin{array}{llll} n_1 & \text{ISA} & n_2, & \text{and} \\ n_2 & \text{R} & n_3 & \\ n_1 & \text{R} & n_3 & \end{array}$$

where the relation R between n_2 and n_3 is viewed as a property of n_2. For example, if a canary is a bird and bird HAS-COMPONENT wings, then canary HAS-COMPONENT wings (Fig. 1.a). Here, R is HAS-COMPONENT and the inherited property is "HAS-COMPONENT wings." Many plausible inference rules have this structure, but inherit over links other than ISA. For example, in the "cough syrup" inference, above, cough syrup inherits the "CAUSES intoxication" property over the CONTAINS relation:

cough syrup	HAS-COMPONENT	alcohol,	and
alcohol	CAUSES	intoxication	
cough syrup	CAUSES	intoxication	

Figure 1.b shows two other examples. They have the same premises but different conclusions. One premise is "storm HAS-COMPONENT cloud" (and, equivalently, "cloud COMPONENT-OF storm"); the other is "cloud MECHANISM-OF rain" (and, equivalently, "rain HAS-MECHANISM cloud"). But the conclusions are "storm MECHANISM-OF rain" and "rain COMPONENT-OF storm," respectively.

This illustrates that each pair of relations can produce two plausible inference rules that have the same structure as property inheritance over ISA links. For relations R1, R2 these rules are:

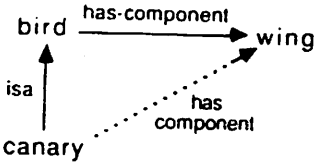
Rule 2

$$\begin{array}{llll} n_1 & \text{R1} & n_2, & \text{and} \\ n_2 & \text{R2} & n_3 & \\ n_1 & \text{R2} & n_3 & \end{array}$$

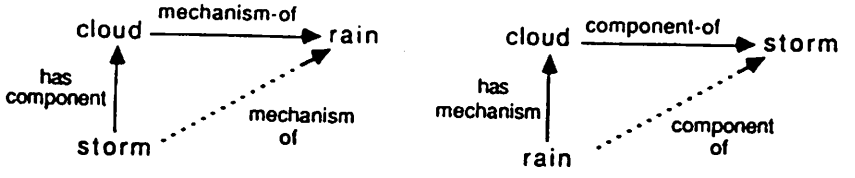
and

Rule 3

$$\begin{array}{llll} n_3 & \text{R2-INV} & n_2, & \text{and} \\ n_2 & \text{R1-INV} & n_1 & \\ \hline n_3 & \text{R1-INV} & n_1 & \end{array}$$



a. Property inheritance over ISA links



b. Inferences that have the structure of property inheritance

Figure 1: Inference from property inheritance and structurally-identical rules

Figure 1.b shows these alternatives for R1 = HAS-COMPONENT, R2 = MECHANISM-OF, n_1 = storm, n_2 = cloud, and n_3 = rain.

Figure 1 introduces the notation we will use throughout. Rules are represented as triangles formed from three concepts and three relations. The legs of the triangle represent premises, and are always drawn as solid lines. The hypotenuse represents the conclusion and is always drawn as a dashed line.

Rules can be chained by letting the conclusion of one serve as a premise for another. Figure 2 shows how the conclusion of a *first generation* inference, "storm MECHANISM-OF rain," serves as the premise of a *second generation* inference, which has the conclusion "storm HAS-PRODUCT runoff."

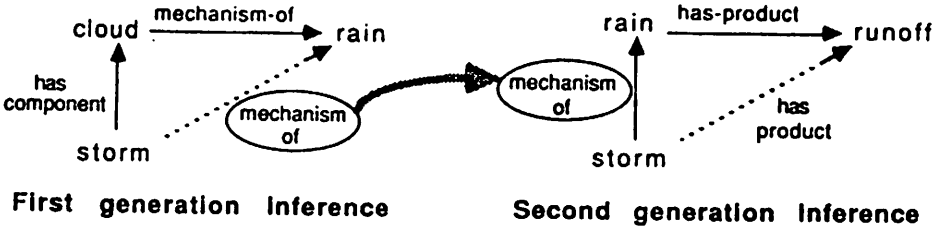


Figure 2: Second-generation inference

Since each pair of relations produces two rules, a knowledge base constructed from N relations will produce $(N^2 + N)/2$ pairs of relations (including relations paired with themselves) and an equal number of rules. The GRANT KB is constructed from nine relations and their inverses, so $(18^2 + 18)/2 = 342$ were generated.

Experiment 1 had two goals. One was to generate all possible rules for the GRANT KB and to determine which of them produce plausible conclusions. The other was to find out how the plausibility of conclusions is affected by chaining these rules. Applying

roughly 300 rules to the GRANT KB (as we describe below), produced thousands of first generation inferences and over 200,000 second-generation inferences. We expected very few of these to be plausible; but, if we could discover or predict the plausible ones, then we would have a powerful method to reduce the effort of constructing large knowledge bases.

2.2 Design

To determine whether the rules produce plausible conclusions, we first *instantiate* them with specific concepts, then present them to human subjects to judge.

We derived 315 rules from the GRANT KB.[1] For each we produced 10 *test items* (five first generation items and five second generation items) by the following method:

Each rule is based on two relations. For each pair, say HAS-COMPONENT and MECHANISM-OF, we search the GRANT KB for triples of nodes n_1,n_2,n_3 that are connected by these relations (i.e., n_1 is connected to n_2 by HAS-COMPONENT, and n_2 is connected to n_3 by MECHANISM-OF). Each triple represents a pair of premises from which *two* inferences can be drawn (see Rules 2 and 3, above). For instance, storm, cloud, and rain instantiate n_1, n_2, and n_3, respectively in Figure 1.b, yielding the conclusions "storm MECHANISM-OF rain" and "rain COMPONENT-OF storm."

Most pairs of relations in the GRANT KB yield dozens of n_1,n_2,n_3 triples. We randomly select five, and their conclusions, to be first generation test items. However, we add the conclusions of *all* the triples to the GRANT KB.

This procedure is repeated to generate second generation test items, with the added condition that one premise of each second generation item must be a conclusion that was produced during the previous search (though not necessarily the conclusion of a first generation test item).

In all, the 315 rules yield a data set of 3116 test items, of which roughly half are first generation and half are second generation items.[2]

2.3 Procedure

Items in the data set were presented to human subjects by a computer program. Subjects were asked first to indicate whether both premises were acceptable, one or both were unacceptable, or they did not understand one or both premises. Next, the conclusion was shown and subjects were asked to judge whether it followed or did not follow from the premises, or else to indicate that they did not understand the conclusion. Each item was seen by two subjects. Following a practice session with 20 items (none of which was in the data set), each subject judged approximately 700 items

[1] Pruning duplicates reduces the original 342 rules to 315.

[2] We don't have 3150 items because, for some rules, the GRANT KB yielded fewer than five first generation instances.

from the data set. This took about five hours, distributed over three or four self-paced sessions.

2.4 Results

Since the premises of the test items came from an existing knowledge base we expected that most would be judged acceptable. This is in fact the case: 82% percent of first generation premises and 63% of second generation premises were judged to be acceptable. The following results pertain only to those items.

Each rule is represented in the data set by five first generation items and five second generation items, and each item was seen by two subjects. Thus, 10 judgments are made of the items in each generation of each rule. Two *plausibility scores* for a rule, ranging from 0 to 10, are equal to the sum of the number of items that subjects judged plausible for each generation of each rule. The mean plausibility score, over the 315 rules, for first generation items is 4.18 (var. = 6.92), and the corresponding statistic for second generation items is 3.17 (var. = 4.88). Both are significantly different from chance and from each other at the $p < .01$ level. The fact that both are *below* chance means that the preponderance of rules are not plausible. Given this, one would expect chaining of inferences to produce increasingly-implausible conclusions. This is supported by the evidence that second generation inferences are significantly less plausible than first generation ones. Subjects judged approximately 50% of the rules to have plausibility scores between 3 and 7 (of a possible 10); they judged the rest of the rules to be predominantly plausible or implausible.

2.5 Discussion

While these results indicate that many rules generate predominantly plausible conclusions, and many others are predominantly implausible, they do not tell us how to predict which will be plausible and which will not. We wanted to find a small set of common characteristics of rules on which to base these predictions. Furthermore, we wanted these characteristics to depend only on the relations in the rules, not on the nodes or any other exogenous factors.

We discovered two common aspects of relations. Some relations, such as HAS-COMPONENT have a *hierarchical* interpretation. Others, such as CAUSES, can be interpreted as *temporal* relations. Lastly, relations such as MECHANISM-OF can have both hierarchical and temporal interpretations: in "n_1 MECHANISM-OF n_2," n_2 may be a process that hierarchically subsumes the mechanism n_1, or n_1 may be an object or process that exists or is required prior to achieving n_2. Table 1 lists the *deep relations* that correspond to all 18 *surface* relations. Each deep relation has a h (hierarchical) or t (temporal) interpretation, or both. Expressing rules in terms of these deep relations reduces the set of 315 surface rules to 95 unique *deep structures*.

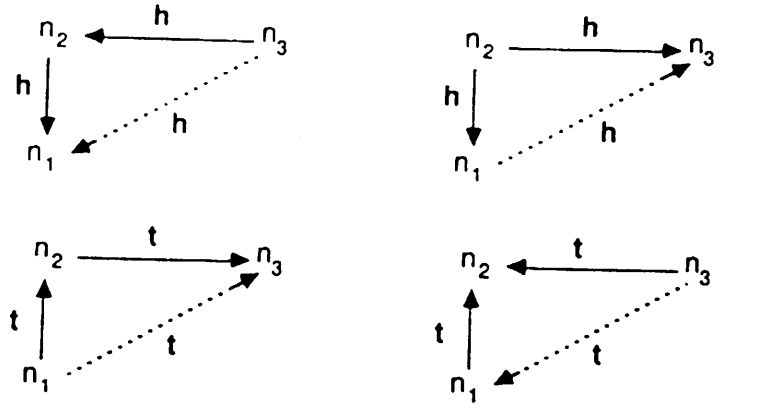
Surface relation	Deep structure	Surface relation	Deep structure
CAUSES	t →	CAUSED-BY	← t
COMPONENT-OF	← h	HAS-COMPONENT	h →
FOCUS-OF	← h	HAS-FOCUS	h →
MECHANISM-OF	t → h ←	HAS-MECHANISM	← t h →
PRODUCT-OF	← t h →	HAS-PRODUCT	t → h ←
PURPOSE-OF	← t t →	HAS-PURPOSE	→ t ← t
SETTING-OF	h →	SETTING	← h
SUBJECT-OF	← h	SUBJECT	h →
SUBFIELD-OF	← h	HAS-SUBFIELD	h →

Table 1: Surface relations and corresponding deep relations

More importantly, we identified a characteristic of deep structures, called *transitivity*, which seemed to explain why some rules were plausible and others implausible. Figure 3 shows two transitive structures and two intransitive ones. The transitive deep structures represent the rules: "If n_1 CAUSES n_2, and n_2 CAUSES n_3, then n_1 CAUSES n_3," and "If n_1 COMPONENT-OF n_2, and n_2 COMPONENT-OF n_3, then n_1 COMPONENT-OF n_2." We call these structures transitive because the premises imply an ordering between n_1 and n_3 that, to be preserved, *requires* a particular ordering between n_1 and n_3 in the conclusion (n_1 to n_3 in one rule and n_3 to n_1 in the other). In contrast, the intransitive structures do not require any ordering on nodes in the conclusion. In one, the premises indicate no hierarchical ordering between n_1 and n_3, only that n_2 is hierarchically-superior to both. Similarly, in the other intransitive rule, n_1 and n_3 are both temporally-prior to n_2, but no ordering is implied between them and, thus, required in the conclusion.[3] The mean plausibility score for transitive rules was 8.94 (out of 20; var. = 16.83), and for intransitive rules, 5.89 (var. = 14.46). Again, the preponderance of these rules are judged implausible, but these values are significantly different ($p < .01$), and provide strong post-hoc evidence that transitivity is a factor.

Transitivity is clear when surface relations map to deep relations whose h and t elements point in just one direction. But the surface relations HAS-MECHANISM and PURPOSE-OF have deep relations where t and h point in opposite directions. Therefore, rules that are transitive under one interpretation of these relations are necessarily intransitive under the other. For example, the structure in Figure 4.a may be transitive or intransitive. We call structures like this *ambiguous*.

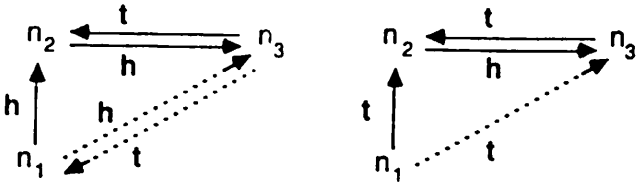
[3]The term transitivity refers to the form of the deep structure, and does not imply mathematical transitivity.



a. Example transitive deep structures

b. Example intransitive deep structures

Figure 3: Transitive and intransitive deep structures



a. Consistent h interpretation

b. Consistent t interpretation

Figure 4: Ambiguous deep structures

Although our data suggested that transitivity predicts the plausibility of rules with unambiguous structures, the results were less clear for ambiguous ones. All ambiguous structures have transitive interpretations, but we knew from our data that not all the corresponding rules were plausible. We hypothesized a characteristic of interpretations, called *consistency*, that might discriminate plausible ambiguous rules from implausible ones. A structure has a consistent interpretation when its deep relations all have the same interpretation, either **h** or **t**. For example, Figure 4.a has a consistent interpretation in which all its deep links can be interpreted as **h**. Moreover, this **h** interpretation is transitive. Figure 4.b has a consistent **t** interpretation, but it is intransitive; and the interpretations of the deep relations that make Figure 4.b transitive are inconsistent (**t**, **t**, and **h**).

3 Experiment 2: Plausible inference as transitivity

At the end of Experiment 1, we had formed the hypotheses that transitivity predicts plausibility, and that consistency determines the interpretation (transitive or intransitive) of ambiguous structures. Experiment 2 tests these hypotheses.

3.1 Design

Experiment 2 focused on ten relations from Experiment 1: CAUSES, COMPONENT-OF, MECHANISM-OF, PRODUCT-OF, PURPOSE-OF and their inverses. (The other relations replicate deep relations and occurred relatively infrequently in the knowledge base.) Since each of these surface relations has a unique corresponding deep relation, the 95 rules they generate map to 95 different deep structures. From these, we chose 56 structures (and thus, rules) as a representative sample.[4] We generated 10 first generation test items for each of the 56 rules, following the procedure described in Experiment 1.

3.2 Procedure

Fourteen subjects each viewed all the test items. Items were presented as in Experiment 1.

3.3 Results

Our hypothesis is that transitivity, as determined by the consistent interpretation of the deep structure, predicts plausibility. Eight rules are composed of surface relations that have just one deep interpretation (CAUSES, CAUSED-BY, HAS-COMPONENT, COMPONENT-OF; see Fig. 5). With these we can analyze the effects of transitivity and consistency on plausibility in rules with single interpretations. A two-way analysis of

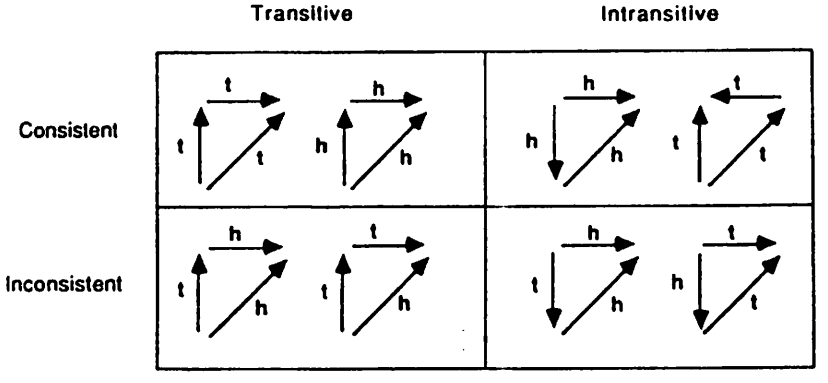


Figure 5: Single interpretation deep structures

variance found a significant main effect of transitivity ($p < .001$) and a significant transitivity × consistency interaction ($p < .001$), but no main effect of consistency ($p > .2$),

[4]Rules generated from a single surface relation and its inverse always map to one transitive and two intransitive deep structures. Our sample included the transitive structure and one of the intransitive structures (chosen randomly). Pairs of non-identical relations and their inverses form four transitive and four intransitive rules. Our sample included two transitive and two intransitive rules from each of these sets.

confirming that transitivity predicts the plausibility of these rules. A graph of the

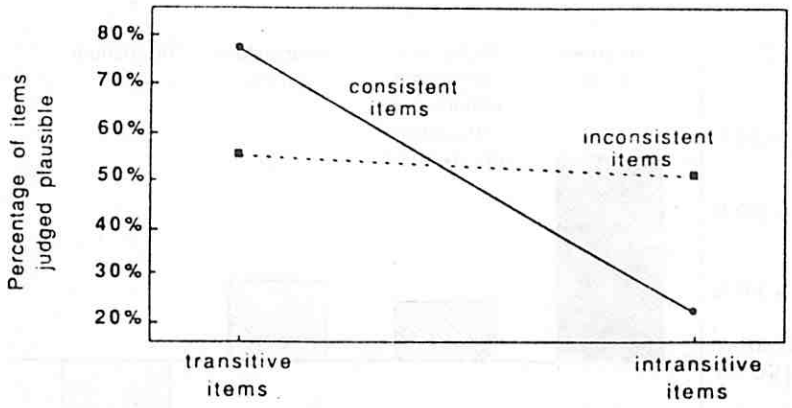


Figure 6: Transitivity × consistency analysis

means (Fig. 6) suggests that we cannot predict the plausibility of rules that have no consistent interpretation, because the mean plausibility score for these rules is roughly five out of 10 (i.e., at chance) irrespective of whether the rule is transitive. Figure 7 compares the mean plausibility scores of transitive, intransitive, and inconsistent rules to chance performance; transitive and intransitive inconsistent items are collapsed into one category.

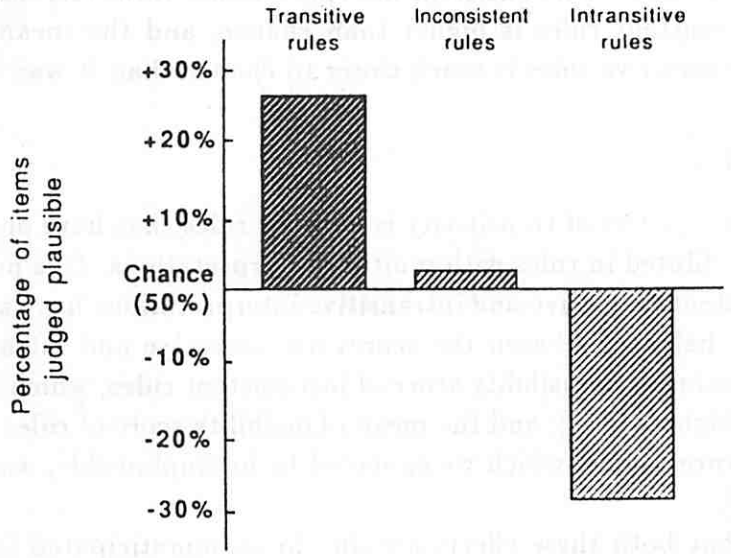


Figure 7: Plausibility scores for rules with single interpretations, expressed as deviations from chance performance

Analyzing all our rules in terms of these categories yields 18 that have consistent transitive interpretations, 20 consistent intransitive rules, 8 inconsistent rules, and

4 rules that have both transitive and intransitive consistent interpretations.[5] The histogram for all rules (including the eight analyzed earlier) is presented in Figure 8.

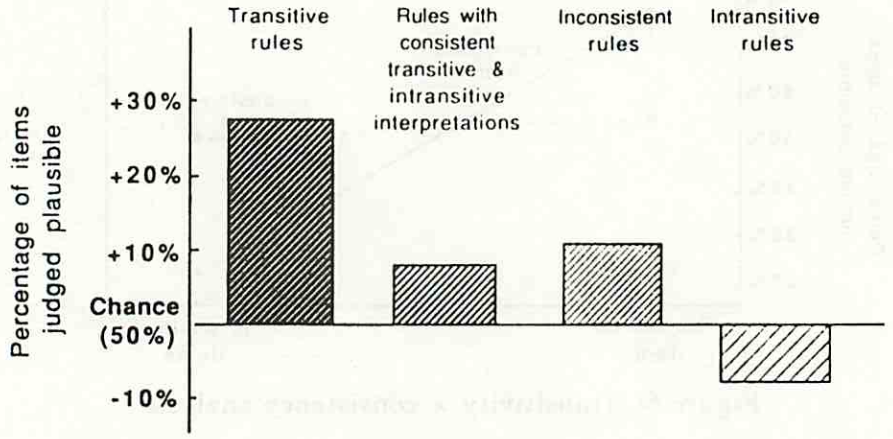


Figure 8: Plausibility scores for all rules, expressed as deviations from chance performance

Although less clear-cut, Figure 8 echoes one of our earlier results: transitivity predicts the plausibility of rules with consistent interpretations. However, the mean plausibility score for inconsistent rules is higher than chance, and the mean plausibility score of consistent intransitive rules is much closer to chance than it was in Figure 7.

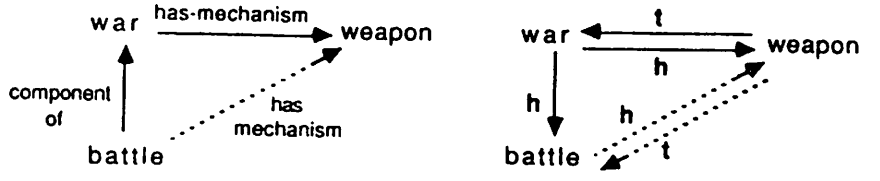
3.4 Discussion

While the predictive power of transitivity is high for rules that have only one interpretation, it becomes diluted in rules with multiple interpretations. It is not surprising that rules with consistent transitive *and* intransitive interpretations have a mean plausibility score roughly halfway between the scores for transitive and intransitive rules (Fig. 8). However, the mean plausibility score of inconsistent rules, which we expected to be at chance, was higher (61%); and the mean plausibility score of rules with consistent intransitive interpretations, which we expected to be implausible, was not as low as we expected (43%).

We hypothesize that both these effects are due to an unanticipated factor that is raising the plausibility of some but not all of these rules. Whereas all our surface rules have the same structure as property inheritance over ISA links, some but not all of the *deep structures* of both the intransitive and inconsistent rules have this form. For

[5] Unfortunately, the test items for the other six rules shared many common premises. This was an unavoidable consequence of our decision to generate test items randomly. Four had consistent transitive interpretations, two had consistent intransitive interpretations.

example, the deep structure for the rule n_1 COMPONENT-OF n_2, n_2 HAS-MECHANISM $n_3 \rightarrow n_1$ HAS-MECHANISM n_3 is intransitive, but its conclusion is often plausible, as illustrated in Figure 9. In this instantiation, battle inherits "HAS-MECHANISM weapon" from war over a COMPONENT-OF relation. We expect rules with this structure to



a. Surface structure b. Deep structure

Figure 9: Surface and deep structures of an intransitive but plausible rule

yield relatively high plausibility ratings even if they are intransitive, because property inheritance is a common and powerful plausible inference rule.

Generalized property inheritance (GPI) is a characteristic of a rule's deep structure, comparable with transitivity:

> If n_1 is related to n_2 by h, and n_2 is related to n_3 by any relation i, then it is plausible to infer that n_1 is related to n_3 by i

This definition does not restrict the direction of h; it can point "up" or "down" from n_1 to n_2, whereas in property inheritance over ISA links, n_1 must be a subclass or instance of n_2, that is, ISA must point "up." We relax this for GPI because it is often plausible to infer that a concept will have properties of those concepts hierarchically-inferior to it.

GPI explains why some intransitive rules have higher than expected plausibility scores. Since some transitive rules are also GPI, we ran a post-hoc transitivity $\times$ GPI analysis of variance, and found main effects of transitivity ($p < .001$) and GPI ($p < .05$), with no interaction effect. Post-hoc tests on the means (Newman-Keuls) found a significant difference between GPI intransitive items and non-GPI intransitive items ($p < .05$), which means that among intransitive rules, GPI differentiates two statistically-distinct classes—relatively plausible and relatively implausible rules. After removing GPI rules, the mean plausibility score of inconsistent rules decreases (Fig. 10). Therefore, GPI provides a post-hoc explanation of why intransitive and inconsistent rules have higher-than-expected plausibility scores. Among transitive items, GPI had no statistically discernible effect. And since there was no interaction between transitivity and GPI, we regard them as independent factors.

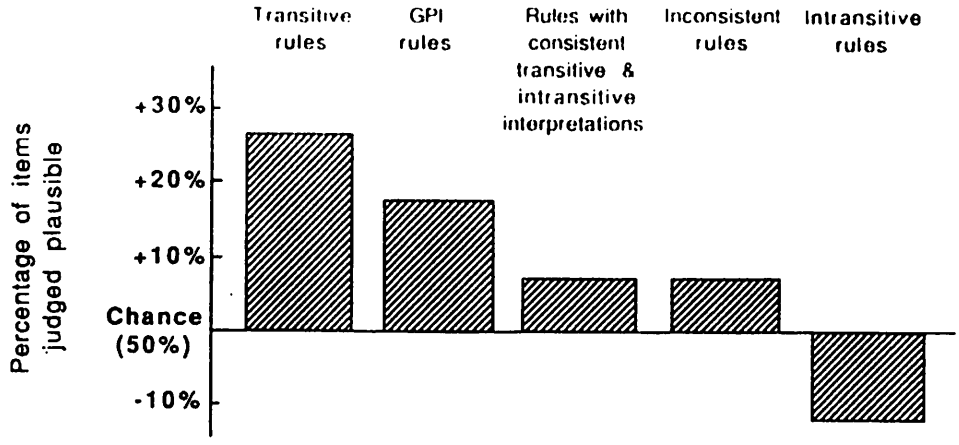


Figure 10: Post-hoc revision of Figure 8, treating GPI rules separately

4 General Discussion—Judging plausibility

In this section we will discuss the factors that contribute to judgments of plausibility. Recall that our goal is to find plausible inference rules that support graceful degradation and help knowledge engineers. Ideally, the agent who uses these rules should not need much knowledge to judge the plausibility of their conclusions. For example, the plausibility of the conclusion of the rule

$$\begin{array}{lll} n_1 & \text{CAUSES} & n_2, \text{ and} \\ n_3 & \text{CONTAINS} & n_1 \\ \hline n_3 & \text{CAUSES} & n_2 \end{array}$$

seems not to depend on the objects that instantiate n_1, n_2, and n_3. In contrast, to judge the plausibility of a conclusion of the rule

$$\begin{array}{lll} n_1 & \text{CAUSES} & n_2, \text{ and} \\ n_2 & & \\ \hline n_1 & & \end{array}$$

we need knowledge about n_1 and n_2 that can tell us how likely n_1 is given n_2.

Several authors have noted the tradeoff between the amount one knows about a plausible inference and and one's confidence in its conclusion [6,5,1,11]; for example, Polya notes that "In order to judge the weight of the evidence, you have to be familiar with the domain; in order to judge the weight with assurance, you have to be an expert in the domain." [11, p. 114].

What knowledge contributes to the plausibility of the conclusions of the rules in Experiments 1 and 2? Said differently, what factors account for the *total variance in judgments of plausibility* (T) among our subjects? We believe T has four additive components:

subject variance—the proportion of T due only to individual differences in subjects' knowledge, experience, motivation, and so on.

item variance—the proportion of T due only to differences in the concepts that instantiate n_1, n_2, n_3 in the rule.

between-rule variance—the proportion of T due only to differences in the surface structures of rules.

deep structure variance—the proportion of T due only to whether deep structures are transitive, intransitive, or GPI structures.

Ideally, deep structure variance should account for the largest component of T. If 100% of T was due to deep structure variance, then transitivity and GPI would be perfect predictors of plausibility. In contrast, if a large fraction of T is due to item variance, then one needs to know the specific instantiation of a rule—the concepts in the test item—to predict its plausibility. Similarly, between-rule variance represents the effect of knowing the surface structure of test items on one's ability to predict their plausibility. Subject variance represents the limit of our ability to predict plausibility.

For transitive and intransitive rules, and to a lesser extent for GPI rules, deep structure variance accounts for a large fraction of T. For all test items with these structural characteristics, our predictions of plausibility will be correct for 77% of transitive items and 68% of GPI items; and our prediction of *im*plausibility will be correct for 62% of intransitive items. Since these numbers are not 100%, the remaining variance in T must be due to the rule, item, and subject factors.

We estimated between-rule variance as follows: We ran three one-way analyses of variance, by rule, for transitive, intransitive, and GPI rules. This allowed us to make a rough estimate of the proportion of the variance in each of these categories due to rule (as suggested by [7, p. 485]). Between-rule variance is 16% for transitive rules, 27% for intransitive rules, and 52% for GPI rules. That is, if a rule is transitive, then knowing *which* rule it is provides little additional information about the plausibility of items. However, this knowledge accounts for much of the variance in plausibility scores of intransitive and GPI items.

Given that we know the specific rule, how much of the remaining variance (the *within-rule* variance) is due to the individual item and how much is due to subject differences? When the within-rule variance is low, all the items in the rule received approximately the same plausibility score. Therefore, if you know that an item is an instantiation of one of these rules, knowing *which* instantiation it is does not help you predict its plausibility: Most of the within-rule variance is due to subjects. But when the within-rule variance is not low, it may be due either to item, subjects, or a combination of both. One way a rule can have a high within-rule variance is if the plausibility scores of the items fall at both extremes of the scale. For example, if half the

items were judged plausible by all subjects and the other half were judged implausible by all subjects, then none of the variance is due to subjects and all is due to items. On the other hand, if many items received split plausibility scores (i.e., half the subjects found them plausible, the other half did not) then much or all of the remaining variance is due to subjects. Thus, we can use the number of split plausibility scores as a rough estimate of the variance due to subjects.

Twenty-two of our 50 rules had low within-rule variance. (The distribution of within-rule variances was skewed low, but ranged from 0 to 19 with a mean of 6.5; a variance ≤ 5.0 was "low.") For these rules, subject differences seem to contribute more to the within-rule variance than do item differences. That is, knowing the instantiation of these rules does not improve our predictions of their plausibility over the prediction we can make from structure and rule knowledge. The remaining 28 rules have split plausibility scores on one or more items. Nine of them have two or fewer split plausibility scores, indicating that little of the variance is due to subjects. Knowing the instantiation of these rules *would* improve our predictions of their plausibility over the predictions we can make from structure and rule knowledge. The remaining 19 rules all have between 3 and 6 split plausibility scores, which suggests that more of their variance is due to subject differences.

5 Conclusion

This paper suggests that we can automatically derive plausible inference rules from the relations in knowledge bases and predict judgments of plausibility for the conclusions of these rules. Two structural factors (transitivity or GPI) correctly predict plausibility 77% and 68% of the time. No knowledge is required to apply these criteria. Greater accuracy requires more knowledge, particularly knowledge about the specific rules and the concepts that instantiate them; but because we could not accurately estimate the contribution of individual differences among our subjects to T, we do not know the limit on the accuracy of our predictions.

Our experiments relied on the GRANT KB, which was built for a different purpose. Although our results are limited to this knowledge base, we believe they are more general, because the surface relations in the GRANT KB are common, and because h and t are general semantic components, and because transitivity and GPI are common structural characteristics. But further work is required to *prove* the generality of our results.

Our goal was to develop methods to support graceful degradation and knowledge engineering. Clearly, these purposes are not met if plausible inference rules require masses of knowledge to judge their conclusions. We are very encouraged by the relatively high accuracy of criteria that require no knowledge, and by the fact that our accuracy is higher for plausible rules than for implausible ones.

References

[1] Michelle Baker and Mark H. Burstein. Implementing a model of human plausible reasoning. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 185–188, Milan, Italy, 1987.

[2] Ronald J. Brachman. "I lied about the trees" or, defaults and definitions in knowledge representation. *AI Magazine*, 6(3):80–93, Fall 1985.

[3] Paul R. Cohen, Alvah Davis, David S. Day, Michael Greenberg, Rick Kjeldsen, Sue Lander, and Cindy Loiselle. Representativeness and uncertainty in classification systems. *AI Magazine*, 6(3):136–149, Fall 1985.

[4] Paul R. Cohen and Rick Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management*, 23(4):255–268, 1987.

[5] A. Collins. Fragments of a theory of human plausible reasoning. In D. Waltz, editor, *Theoretical Issues in Natural Language Processing*, University of Illinois, Urbana, IL, 1978.

[6] A. Collins, E. Warnock, N. Aiello, and M. Miller. Reasoning from incomplete knowledge. In D. G. Bobrow and A. Collins, editors, *Representation and Understanding*, Academic Press, New York, 1975.

[7] Willim L. Hays. *Statistics for the Social Sciences*. Holt, Rinehart, and Winston, second edition, 1973.

[8] Rick Kjeldsen and Paul R. Cohen. The evolution and performance of the GRANT system. *IEEE Expert*, 2(2):73–79, Spring 1987.

[9] D. B. Lenat and E. Feigenbaum. On the thresholds of knowledge. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 1173–1182, Milan, Italy, 1987.

[10] D. B. Lenat, M. Prakash, and M. Shepherd. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine*, 6(4):65–85, Winter 1986.

[11] G. Polya. *Patterns of Plausible Inference*. Princeton University Press, Princeton, New Jersey, 1954.