Toward a Unified Theory of Immediate Reasoning in Soar

Thad A. Polk, Allen Newell, and Richard L. Lewis
School of Computer Science
Carnegie Mellon University

Abstract

Soar is an architecture for general intelligence that has been proposed as a unified theory of human cognition (UTC) (Newell, 1989) and has been shown to be capable of supporting a wide range of intelligent behavior (Laird, Newell & Rosenbloom, 1987; Steier *et al.*, 1987). Polk & Newell (1988) showed that a Soar theory could account for human data in syllogistic reasoning. In this paper, we begin to generalize this theory into a unified theory of immediate reasoning based on Soar and some assumptions about subjects' representation and knowledge. The theory, embodied in a Soar system (IR-Soar), posits three basic problem spaces (**comprehend**, **test-proposition**, and **build-proposition**) that construct annotated models and extract knowledge from them, learn (via chunking) from experience and use an attention mechanism to guide search. Acquiring task specific knowledge is modeled with the **comprehend** space, thus reducing the degrees of freedom available to fit data. The theory explains the qualitative phenomena in four immediate reasoning tasks and accounts for an individual's responses in syllogistic reasoning. It represents a first step toward a unified theory of immediate reasoning and moves Soar another step closer to being a unified theory of all of cognition.

IMMEDIATE REASONING TASKS

An immediate reasoning task involves extracting implicit information from a given situation within a few tens of seconds. The examples addressed here are relational reasoning, categorical syllogisms, the Wason selection task, and conditional reasoning. Typically, they involve testing the validity of a statement about the situation or generating a new statement about it. The situation, and often the task instructions, are novel and require comprehension. Usually, but not invariably, they are presented verbally. All the specific knowledge required to perform the task is available in the situation and the instructions and need not be consistent with other knowledge about the world (hence the task can be about unlikely or imaginary states of affairs).

THE SOAR THEORY OF IMMEDIATE REASONING

The Soar theory of immediate reasoning makes the following assumptions (elaborated below):

1. **Problem spaces.** All tasks, routine or difficult, are formulated as search in problem spaces. Behavior is always occurring in some problem space.
2. **Recognition memory.** All long-term knowledge is held in an associative recognition memory (realized as a production system).
3. **Decision cycle.** All available knowledge about the acceptability and desirability of problem spaces, states, or operators for any role in the current total context is accumulated, and the best choice made within the acceptable alternatives.
4. **Impasse driven subgoals.** Incomplete or conflicting knowledge at a decision cycle produces an

impasse. The architecture creates a subgoal to resolve the impasse. Cascaded impasses create a subgoal hierarchy.

5. **Chunking.** The experience in resolving impasses continually becomes new knowledge in recognition memory, in the form of chunks (constructed productions).
6. **Annotated models.** Problem space states are annotated models whose structure corresponds to that of the situation they represent.
7. **Focus of attention.** Attention can be focused on a small number of model objects. Operators are triggered by objects in the focus. When no operators are triggered, an impasse occurs and attention operators add other objects to the focus. Matching and related objects are added first.
8. **Model manipulation spaces.** Immediate reasoning occurs by heuristic search in model manipulation spaces that support comprehension, proposition construction, and proposition testing.
9. **Distribution of errors.** The main sources of errors are interpretation, carefulness and independent knowledge.

The first five assumptions are part of the Soar architecture. Annotated models and attention embody a discipline that is used for modeling cognition (and may become part of the architecture). The last two assumptions are specific to immediate reasoning.

A Soar system consists of a collection of problem spaces with states and operators. At each step during problem solving, the recognition memory brings all relevant knowledge to bear and the decision cycle determines how to proceed. An impasse arises if the decision cycle is unable to make a unique choice. This leads to the creation of a subgoal to resolve the impasse. Upon resolving the impasse, a chunk that summarizes the relevant problem solving is added to recognition memory, obviating the need for similar problem solving in the future.

The states in problem spaces are represented as *annotated models*. A *model* is a representation that satisfies the *structure correspondence condition*: parts, properties, and relations in the model (model elements) correspond to parts, properties, and relations in the represented situation, without completeness (Johnson-Laird, 1983). By exploiting the correspondence condition, processing of models can be match-like and efficient. The price paid is limited expressibility (e.g., models cannot directly represent disjunction or universal quantification). Arbitrary propositions can be represented, but only indirectly, by building a model of a proposition — a model interpretable as an abstract proposition, rather than a concrete object. Some expressibility can be regained without losing efficiency by attaching *annotations* to model elements. An annotation asserts a variant interpretation for the element to which it is attached, but is local to that element and does not admit unbounded processing (e.g., **optional** means that the model element *may* correspond to an element in the situation, but not necessarily).

Problem space states maintain a *focus of attention* that points to a small set of model objects. An operator is proposed when attention is focused on model objects that match its proposal conditions. When no operators are proposed, an impasse occurs and the system searches for a focus of attention that triggers one. Objects that share properties with a current focus of attention or are linked by a relation to one are tried first (others are implicitly assumed to be less relevant). When attention focuses on an object that triggers an operator, the impasse is resolved and problem solving continues.

Immediate reasoning occurs by heuristic search in *model manipulation spaces* (**comprehend**, **build-proposition**, and **test-proposition**). These spaces provide the basic capabilities necessary for immediate reasoning tasks, namely, constructing representations and generating and testing conclusions (Johnson-Laird, 1988). We assume that normal adults possess these spaces before they are confronted with these tasks. All of these problem spaces use the attention mechanism described above. **Comprehend** reads language and generates models that correspond to situations. It produces a model both of what is described (a *situation model*) and of the linguistic structure of the utterance itself (an *utterance model*). **Build-proposition** searches the space of possible propositions until it finds a proposition that is consistent with the situation model and that satisfies any added constraints in the goal test (e.g., its subject is "fork"). It works by combining properties and relations of model objects into constructed propositions. If attention is focused on an existing proposition, the attention mechanism biases the problem solving toward using parts of it. As a result, constructed propositions tend to be similar to existing propositions on which attention is focused. **Test-proposition** tests models of propositions against models of situations to see if they are valid. It does so by searching for objects in the situation model that correspond to those described in the proposition, and checking if the proposition is true of them. A proposition is considered true or false only if the situation model explicitly confirms or denies the proposition in question (i.e., there are objects in the situation model that correspond to the subject and object of the proposition that are (not) related in the way specified by the proposition). If a proposition is about an object(s) that does not match anything in the situation model, the proposition is considered irrelevant. If a proposition is about an object(s) that does appear in the situation model, but is neither explicitly confirmed nor denied, the proposition is considered relevant but unknown.

Individual subjects respond quite differently from each other in many immediate reasoning tasks. The theory predicts that these differences arise mainly from four sources: (1) the interpretation of certain words and phrases (e.g., quantifiers, connectives), (2) the care taken during reasoning (e.g., completeness of search, testing candidate solutions), (3) knowledge from sources outside the task (such as familiarity with the subject matter), and (4) the order in which attention is focused on model objects. We propose that most errors arise from interpretation mistakes (failing to consider all of the implicit ramifications of the premises or making unwarranted assumptions), incomplete search for conclusions (including the generation of other models if necessary), and less frequently from the inappropriate use of independent knowledge. This predicts that better subjects will interpret premises more completely and correctly or will search more exhaustively for a conclusion. Immediate reasoning tasks are difficult to the extent that they present opportunities for these errors.

ACQUIRING TASKS FROM INSTRUCTIONS

Immediate reasoning is so intimately involved in acquiring knowledge, both of the situation to be reasoned about and the task to be performed, that a theory of immediate reasoning needs to include a theory of acquisition. A companion paper (Lewis, Newell & Polk, 1989) describes NL-BI-Soar, a Soar system that acquires tasks from simple natural language utterances. NL-BI-Soar provides the **comprehend** problem space for IR-Soar, producing both the situation model and the utterance model. It also comprehends the instructions for these tasks. This leads to the creation of a problem space that is unique to the task, whose operators make use of the pre-existing spaces, **comprehend**, **test-proposition** and **build-proposition**. It is usual in cognitive theories for this structuring of the task to be posited by the theory — to be, in effect, added degrees of freedom in fitting the theory to the data. In the Soar

Relational Reasoning

Instructions	**Relation** Problem Space
1. Read four premises.	1. Read-input [**comprehend**]
2. Then read a statement.	2. Read-input [**comprehend**]
3. If the statement is "true", say "true".	3. Test-prop [**test-proposition**]
4. Then produce a statement ...	4. Make-conclusion [**build-proposition**]
5. ...that relates the fork to the knife	5. [goal-test]

Categorical Syllogisms

Instructions	**Syllogism** Problem Space
1. Read two premises that share a term.	1. Read-input [**comprehend**]
2. Then produce a statement that follows from the premises.	2. Make-conclusion [**build-proposition**]
3. The statement relates the unique terms of the premises.	3. [goal-test]

Wason Selection Task

Instructions	**Wason** Problem Space
1. Examine four cards that have a number on one side and a letter on the other side.	1. Read-input [**comprehend**]
2. Then read a statement.	2. Read-input [**comprehend**]
3. For each card, does deciding if the statement is true require turning over the card?	3. Test-prop [**test-proposition**]

Conditional Reasoning

Instructions	**Conditional** Problem Space
1. Read two premises.	1. Read-input [**comprehend**]
2. Then read a statement.	2. Read-input [**comprehend**]
3.Then decide if the statement is true.	3. Test-prop [**test-proposition**]
	OR
1. Read two premises.	1. Read-input [**comprehend**]
2. Then produce a statement that follows from the premises.	2. Make-conclusion [**build-proposition**]

Figure 1: **Task instructions and the corresponding problem spaces.**

theory, comprising NL-BI-Soar and IR-Soar jointly, these degrees of freedom no longer exist. The instructions do not specify all details of IR-Soar versions (there are still substantial individual differences among subjects), but do add a major constraint.

Figure 1 lists the English instructions and the corresponding operators for each task. The subspace used to implement each operator is given in brackets next to the operator name. As NL-BI-Soar reads the instructions it builds a model of what they describe (i.e., the required behavior). When the described task is attempted, impasses arise and NL-BI-Soar consults the behavior model to determine how to proceed, leading to the construction of the problem space (see Lewis, Newell, and Polk (1989) for details).

RELATIONAL REASONING

Relational reasoning involves deducing implicit relationships between objects given explicit relationships (e.g., 3-term series problems). Figure 2 illustrates a version similar to that in Johnson-Laird (1988). Given a set of premises (Figure 2, left) that describe a spatial configuration of objects, the task is to answer questions or make conclusions about the described situation (Figure 2, right).

Premises
1. A plate is left of a knife.
2. A fork is left of the plate.
3. A jug is above the knife.
4. The fork is below a cup.

Read this statement and say if it is true:
"The cup is left of the jug"
then
Produce a statement that relates the fork to the knife

Figure 2: **Relational reasoning task (after Johnson-Laird, 1988).**

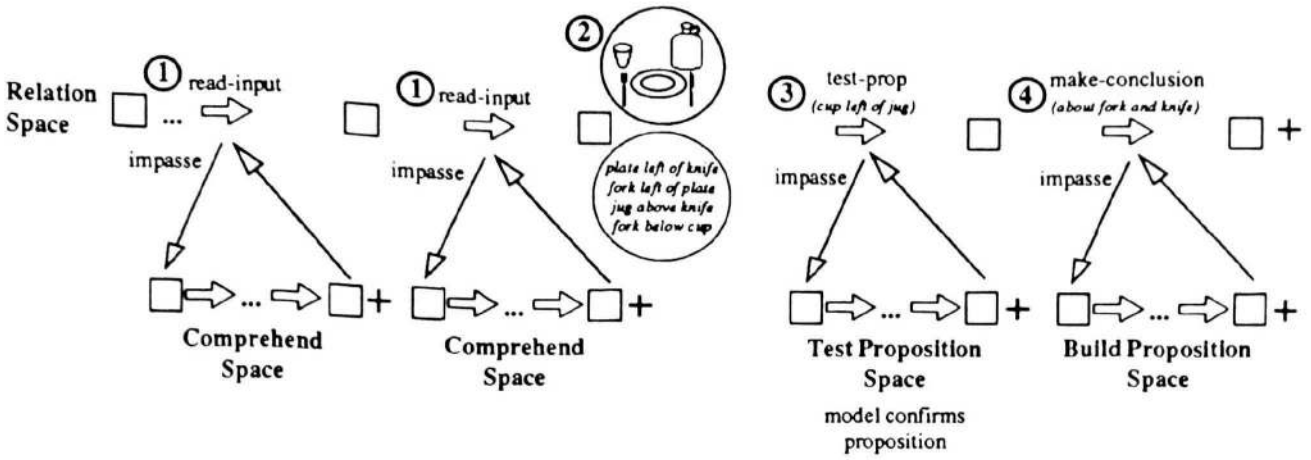


Figure 3: **Behavior of IR-Soar on the relational reasoning task.**

Reading the instructions for this task (Figure 1, top left) leads to a model of the required behavior. The objects in this behavior model are actions that need to be performed for this task. When the task is attempted, NL-BI-Soar consults this behavior model and evokes the operators listed in the figure, instantiating them with the appropriate arguments and goal tests.

Figure 3 illustrates the system's behavior on this task. (1) After acquiring the task from the instructions, the system starts in **relation** and applies read-input, implemented in **comprehend**, to each of the premises describing the situation. (2) This results in an initial model of the situation as well as a model of the premises (the utterance model). (3) The third instruction triggers the test-prop operator for the proposition "The cup is left of the jug". This operator is implemented in **test-proposition**. Since the situation model contains an object with property cup that is related via a left-of relation to an object with property jug, the proposition is considered true. (4) Instructions four and five call for generating a proposition about the fork and knife so make-conclusion is chosen, implemented in **build-proposition**. **Build-proposition**'s initial state is focused on a proposition with subject fork and object knife but no relation. Attending to the proposition's fork leads to focusing on the fork in the situation model (which is left of the situation's knife). This leads to constructing the proposition "A fork is left of a knife".

The theory predicts the same relative difficulty of problems of this type as Johnson-Laird (1988). It predicts that problems that have an unambiguous interpretation (i.e., admit only a single model) will be the easiest since they do not present opportunities for interpretational errors (assumption nine). Further, since a single model cannot represent disjunction (assumption six), realizing that a relation holds in some situations while not in others requires using multiple models in searching for a conclusion. Hence, problems without valid conclusions will be the hardest since they invite incomplete search (assumption nine). Ambiguous problems that support a valid conclusion will be of intermediate difficulty since conclusions based on considering only a single model may be correct. The percentage of correct responses for each of these problem types confirms these predictions (70%, 8%, and 46% correct, respectively). Many relational reasoning studies have focused on response latencies (Huttenlocher, 1968) and we have not yet addressed this data. The emphasis here is on accounting for major phenomena from many different tasks rather than explaining a single task in its entirety. Eventually we

Premise 1 : No archers are bowlers
Premise 2 : Some bowlers are clowns

Conclusion : Some clowns are not archers

(classified as EabIbc Oca)

A : All a are b
I : Some a are b
E : No a are b
O : Some a are not b

#1 ab #2 bc (EabIbc)	#1 ba #2 bc (AbaObc)
#1 ab #2 cb (OabAcb)	#1 ba #2 cb (IbaEcb)

Figure 4: **Syllogism task.**

expect deep coverage in all of them.

CATEGORICAL SYLLOGISMS

Syllogisms are reasoning tasks consisting of two premises and a conclusion (Figure 4, left). Each premise relates two sets of objects (a and b) in one of four ways (Figure 4, middle), and they share a common set (*bowlers*). A conclusion states a relation between the two sets of objects that are not common (the end-terms, *archers* and *clowns*) or that no valid conclusion exists. The three terms a,b,c can occur in four different orders, called *figures* (Figure 4, right, examples in parentheses), producing 64 distinct premise pairs.

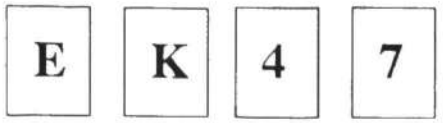
In addition to the basic model manipulation spaces, the task-specific **syllogism** space is used in syllogistic reasoning. Figure 1 shows the correspondence between this problem space and the instructions. This problem space arises directly from the English instructions via NL-BI-Soar. After acquiring the task from the instructions, the system reads both premises and builds a situation model and a model of each of the premises (the utterance model) via **comprehend**. It then attempts to make a conclusion in the **build-proposition** problem space. The attention mechanism biases the form of the constructed conclusion to be similar to that of existing propositions (the premises) (assumptions seven and eight), leading to both the *atmosphere* and *figural* effects. The system may then test the proposition in **test-proposition** and construct additional models, though we have not found this necessary in modeling subjects in the Johnson-Laird & Bara (1984) data.

Polk & Newell (1988) showed how an earlier version of this theory could account for the main trends in group data. Our coverage with the more general theory is almost identical. We have also modeled the individual responses of a randomly chosen subject (subject 16 from Johnson-Laird & Bara (1984)). This subject was modeled by assuming the following processing errors (assumption nine): (1) *all x are y* implies *all y are x* (interpretation), (2) *no x are y* does *not* imply *no y are x* (interpretation), and (3) if neither premise has an end-term as subject, the search is abandoned (carefulness). The focus of attention was treated as a degree of freedom in fitting the subject. For this subject, we were able to predict 55/64 responses (86%).

THE WASON SELECTION TASK

The Wason selection task involves deciding which of four cards (Figure 5, left) *must* be turned over to decide whether or not a particular rule (Figure 5, right) is true (Wason, 1966). This task has been much studied mainly because very few subjects solve it correctly.

Figure 1 shows the top-level **wason** problem space and its correspondence with the instructions. For



Given: Every card has a number on one side and a letter on the other side.
Rule: Every card with an 'E' on one side has a '4' on the other side.

Figure 5: **Wason selection task.**

each of the four cards, this problem space uses the **test-proposition** problem space to try to decide whether it must be turned over. Since the model will not directly answer this question, the system impasses and tries to augment the model. It does so by watching itself decide whether the rule is true while only turning over relevant cards (again using the **test-proposition** problem space). The system will often mistakenly consider cards that do not match the rule to be irrelevant (assumptions seven and eight) and will not select them. The model of deciding whether the rule is true is then inspected to see if the card was in fact turned over, thus resolving the initial impasse of deciding if it must be.

In this task, the cards can be classified into four cases: (1) those that satisfy the antecendent of the rule (the 'E'), (2) those that deny the antecedent of the rule (the 'K'), (3) those that affirm the consequent of the rule (the '4'), and (4) those that deny the consequent of the rule (the '7'). Cards in cases (1) and (4) are the only ones that must be turned over. The theory predicts that, ceteris paribus, cards that do not match the rule will be selected less frequently than those that do (assumptions seven and eight). Evans & Lynch (1973) demonstrated this *matching bias* in an experiment in which they varied the presence of negatives while holding the logical case constant (e.g., they used rules like "Every card with an E on one side does *not* have a '4' on the other side"). In all four logical cases, cards that did not match the rule were selected less frequently than those that did (56% vs. 90%, 6% vs 38%, 19% vs. 54%, and 38% vs. 67%). The standard task is difficult because the correct solution requires overcoming this matching bias to select the '7' (which does not match and hence seems irrelevant) and to reject the '4' (which does match and hence does seem relevant). These mistakes are indeed the two most common made by subjects. A number of other phenomena (e.g., facilitation) arise in variants of this task and the theory has not yet been applied to these.

CONDITIONAL REASONING

Conditional reasoning tasks involve deriving or testing the validity of a conclusion, given a conditional rule and a proposition affirming or denying either the rule's antecedent or consequent (Figure 6).

Figure 1 shows the correspondence between the top-level problem space and the instructions. For this task, the system comprehends the conditional rule and the proposition. It then either constructs a conclusion or tests one that is given, depending on the instructions (using **build-proposition** or **test-proposition**, respectively). In the absence of other knowledge, the system will consider given

Conditional Rule:	If the letter is 'A' then the number is '4'.
Assumed Proposition:	The number is not '4'.
Derive or Test:	The letter is not 'A'.

Figure 6: **Conditional reasoning task.**

conclusions that do not match the conditional to be less relevant (assumptions seven and eight). When constructing conclusions, the system is similarly biased toward conclusions that match (share one or more terms with) the rule (assumptions seven and eight).

Thus, as in the selection task, the theory predicts a matching bias. For conditional reasoning, this implies that conclusions that do not match the conditional will be less frequently constructed and considered relevant than those that do. Evans (1972) showed that when the logical case was factored out, conclusions whose terms did not match the rule were indeed less likely to be constructed than those that share one or both terms (the percentage of subjects constructing conclusions with zero, one, and two shared terms were 39%, 70%, and 86% respectively). Further, when Evans & Newstead (1977) asked subjects to classify conclusions as 'true', 'false', or 'irrelevant', mismatching conclusions were indeed often considered irrelevant.

CONCLUSION

We have presented a theory of human behavior in immediate reasoning tasks based on Soar. The theory uses model manipulation spaces (**comprehend**, **test-proposition**, and **build-proposition**) to construct and extract knowledge from annotated models and is guided by an attention mechanism. Though not reported on here, it includes a theory of learning (chunking). The theory accounts for qualitative phenomena in multiple immediate reasoning tasks and for detailed individual behavior in syllogistic reasoning. This theory is joined by the Soar subtheory for taking instructions in moving Soar to be a unified theory of cognition that deals in depth with a wide range of psychological phenomena.

Acknowledgements

Thanks to Norma Pribadi for making the intricate figures and to Kathy Swedlow for technical editing. Thanks to Erik Altmann and Shirley Tessler for comment and criticism. This work was supported by the Information Sciences Division, Office of Naval Research, under contract N00014-86-K-0678, and by the Kodak and NSF fellowship programs in which Thad Polk and Richard Lewis, respectively, participate. The views expressed in this paper are those of the authors and do not necessarily reflect those of the supporting agencies. Reproduction in whole or in part is permitted for any purpose of the United States government. Approved for public release; distribution unlimited.

References

Evans, J. S. B. (1972). Interpretation and 'matching bias' in a reasoning task. *Quartlerly Journal of Experimental Psychology, 24(2)*, 193–199.

Evans, J. S. B. and Lynch, J. (1973). Matching bias on the selection task. *British Journal of Psychology, 64*, 391–397.

Evans, J. S. B. and Newstead, S. (1977). Language and reasoning: A study of temporal factors. *Cognition, 5(3)*, 265–283.

Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review, 75(6)*, 550–560.

Johnson-Laird, P. (1988). Reasoning by rule or model? In *Proceedings of the Annual Conference of the Cognitive Science Society*, pages 765–771.

Johnson-Laird, P. and Bara, B. (1984). Syllogistic Inference. *Cognition, 16*, 1–61.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness.* Harvard University Press, Cambridge, Massachusetts.

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

Lewis, R., Newell, A., and Polk, T. (1989). Toward a Soar Theory of Taking Instructions for Immediate Reasoning Tasks. To appear in the Proceedings of the Annual Conference of the Cognitive Science Society, August, 1989.

Newell, A. (1989). *Unified Theories of Cognition.* Harvard University Press, Cambridge, Massachusetts. In press.

Polk, T. A. and Newell, A. (1988). Modeling human syllogistic reasoning in Soar. In *Proceedings of the Annual Conference of the Cognitive Science Society*, pages 181–187.

Steier, D. M., Laird, J. E., Newell, A., Rosenbloom, P. S., Flynn, R. A., Golding, A., Polk, T. A., Shivers, O. G., Unruh, A., and Yost, G. R. (1987). Varieties of learning in Soar: 1987. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 300–311.

Wason, P. C. (1966). Reasoning. In Foss, B. M., editor, *New Horizons in Psychology I*, Penguin, Harmondsworth, England.