A Computational Model of Human Abductive Skill and its Acquisition

Todd R. Johnson[1], Josef Krems[2] and Nasir K. Amra[1]

[1]Laboratory for Knowledge-Based Medical Systems
Department of Pathology
The Ohio State University
Columbus, Ohio 43210
johnson.25@osu.edu, amra@med.ohio-state.edu

[2]Institut für Psychologie
Universität Regensburg
Regensburg, Germany
josef.krems@rpss3.psychologie.uni-regensburg.de

Abstract

Abduction is the process of constructing a plausible explanation for a set of observations. It is the fundamental type of reasoning in many complex tasks such as scientific discovery and diagnosis. This paper presents a mental-model theory of human abductive skill and its acquisition in which abduction is viewed as the sequential comprehension and integration of data into a single situation model. Comprehension and integration are accomplished using satisficing search of multiple problem spaces. The model has been implemented in Soar and has been tested by comparing its predictions to those of human subjects. The experimental results show that the model can account for several important behavioral regularities, including power-law speed-up, how the order of data presentation affects a response, deviation of responses from probability theory, and how the task and domain characteristics affect a person's response.

Introduction

Abduction is the process of determining the best explanation for a set of observations (Josephson & Josephson, 1994). It is the fundamental type of reasoning in many complex tasks such as scientific discovery and diagnosis as well as everyday tasks like story comprehension and natural language understanding. The focus of this paper is on how people solve multicausal abduction tasks and how their skill changes with experience. A multicausal abduction task is one in which the explanation consists of a conjunction of causal factors. For example, the best explanation of a patient's symptoms and test results might be a set of simultaneously occurring diseases.

Research in several areas related to abduction suggest that people employ a number of nonnormative heuristics when solving problems involving explanations. For example, when evaluating explanations people tend to ignore base rates and overvalue confirming evidence or evidence that is similar in form to the hypothesis being evaluated (Tversky and Kahneman, 1982; Schustack & Sternberg, 1981). Downing, Sternberg and Ross (1985) found that subjects rated multicausal explanations based on the strongest unicausal factor in the explanation, modified according to the representativeness of the explanation to the evidence. Research on belief updating from many domains reveals that the order in which data is processed can affect a person's belief in a hypothesis (for a review see Hogarth and Einhorn, 1992). Hogarth and Einhorn (1992) have shown that order effects depend on task characteristics such as the complexity and number of items being processed.

Several cognitive models of various subtasks of abductive reasoning have been proposed; however, these models either do not offer the details needed to build process models for multicausal abduction, or fail to consider the sequential nature of the task. Researchers studying scientific discovery have proposed that people reason using coordinated search through experiment and hypothesis spaces. This view is exemplified by Klahr and Dunbar's (1988) model of Scientific Discovery as Dual Search (SDDS) (see also Dunbar & Klahr, 1989). Their theory provides a general explanation for how search in the hypothesis and experiment spaces interacts. SDDS defines three roles for experiments—exploring, hypothesis testing, and hypothesis refinement—and indicates how these roles affect the developing hypothesis. SDDS, however, does not provide detailed models of the subtasks of abduction, such as how hypotheses are generated or how evidence is integrated to select a hypothesis. Thus, while SDDS appears to describe human abductive reasoning at an abstract level, it does not make detailed predictions of human behavior. To more adequately account for human behavior, SDDS must be extended to include details of the problem spaces and the search processes for all of the subtasks of abduction.

As another example, Thagard's (1989) theory of explanatory coherence (TEC) captures our intuitive concept of why one theory is preferred over another; however, the theory largely ignores the sequential nature of abduction. Thagard proposed that people prefer theories that best cohere. TEC and the corresponding process model implementation (ECHO) define coherence (and incoherence) in terms of principles that relate hypotheses to other propositions. For example, a hypothesis coheres with the data that it explains and also with analogous explanations. However, ECHO ignores the sequential nature of abduction because it assumes that people can determine the coherence between all propositions (data and explanatory factors) in parallel. Although this might be possible for problems involving a small number of propositions, it seems unlikely for complex problems like those found in diagnosis or scientific discovery. In addition, since ECHO determines coherence in parallel, it cannot account for

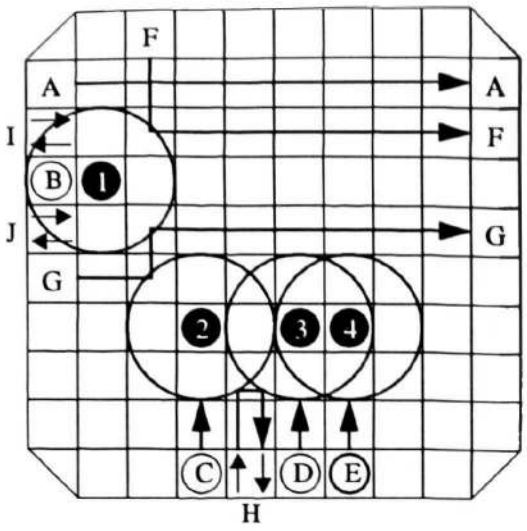


Figure 1: The Black Box with 4 atoms and the paths of several light rays visible.

order effects.

In our research on abductive reasoning, we have taken an approach that is similar to the dual space approach of Klahr and Dunbar and to the mental model approach of Johnson-Laird and Byrne (1991). Johnson-Laird and Byrne have shown that many regularities of human deductive reasoning can be explained by a mental model-based theory. Their theory assumes that people solve deductive tasks by constructing and modifying concrete mental models of a situation, and by searching for alternative models. We have taken a similar view of abductive reasoning. In this paper, we describe a mental model-based theory of human abductive skill and its acquisition that has been implemented in Soar (Laird, et al., 1987). We then show how this theory can account for several regularities seen in human abductive reasoning.

The Experimental Task: Black Box

To further explore human abductive problem solving we have begun to focus on abductive tasks in which the function and structure (F/S) of a device are known and the goal is to determine some hidden state of the device given indirect evidence of that state. As a testbed for our studies we are using a simple game called Black Box (BBX) in which players must locate four atoms hidden in a box by shooting light rays into the box and observing where the rays exit the box. The task was selected because it is easy to understand and yet involves complex abductive problem solving similar to that done in many real-world tasks.

The BBX device is shown in Figure 1. Each atom (labeled 1-4) has a field of influence (shown as a larger circle around the atom). These fields deflect or absorb light rays (according to certain laws) as illustrated in the figure. If a ray directly hits an atom, it is absorbed, and the ray's input cell is marked with a circle (Rays B, C, D and E); if a ray enters and exits at the same location (Rays I, J and H), that location is marked with double arrows (this is called a reflection); otherwise, the locations at which the ray enters and exits the box are marked with a unique symbol (Rays A, F and G, marked with letters).

The Abductive Process Model

The model we describe here is based on Abd-Soar (Johnson and Smith, 1991), a satisficing Artificial Intelligence framework for building abductive systems. Abd-Soar is one of a series of abductive models stemming from the original satisficing technique of Josephson and his colleagues (1987). Although Abd-Soar is designed to capture a wide range of human expert knowledge and to exhibit flexible behavior similar to that exhibited by human experts, the theory's behavior has never been compared, in detail, to human behavior. In addition, many details of the abductive problem-solving process were left unspecified in Abd-Soar. The theory presented here extensively modifies and extends Abd-Soar to account for human abductive behavior.

The basis of our abductive model is a mental model theory that views abduction as the sequential comprehension and integration of data into a single situation model that represents the current best explanation of the data. Although only a single situation model is used, it can contain disjunctive elements. For example, a situation model can contain several possible explanations for a datum. When a new datum is collected the situation model is updated to include the new datum. Next, the new datum must be comprehended to determine what it implies about the situation. The result of comprehension is one or more explanations for the datum. An explanation can be unicausal (a single component cause) or multicausal (a conjunction of component causes). Comprehension can also produce abstract explanations that specify a related class of concrete explanations. If comprehension results in a single explanation that is consistent with the rest of the situation model, then that explanation is assumed to be true. When an explanation is inconsistent with the model an anomaly has occurred and the model must be updated by either finding an alternative explanation for the new datum or by altering an explanation for the old data. When multiple explanations are known to be likely for a datum, one must be selected by considering other data (and possibly collecting new data). This leads to the process of evidence integration.

The remainder of the abductive model is stated in four hypotheses: 1) *The satisficing hypothesis*, that the search for an explanation (or experiment) ends as soon as a single satisfactory explanation (or experiment) is found; 2) *The compilation hypothesis*, that explanations (or experiments) found through search are immediately available to the problem solver when future similar situations arise. In

other words, the results of search are compiled such that the search can be avoided in similar future situations; 3) *The availability hypothesis*, that if only one explanation is directly available and is consistent with the data, it is accepted and used, but if more than one is directly available, the agent must attempt to discriminate. If none are directly available, then the agent must search for an explanation; and 4) *The bounded search hypothesis*, that the search for experiments and explanations is bounded by memory and time constraints.

The abductive model is based on a problem space with 7 operators: *comprehend, refine, discriminate, check, test, resolve-anomaly* and *hypothesize*. The states in this space contain the situation model along with other state information needed to solve the problem. There is no fixed order in which operators are sequenced, rather their sequence is determined at run-time based on the status of each operator's preconditions and on search-control knowledge that prefers one or more operators over others.

Comprehend determines the implications of one or more parts of the situation model. For example, comprehending new data will produce an explanation for that data. Comprehending all of the implications of a single object in a situation model can be a multi-step process, requiring multiple comprehend operators. For example, in the BBX model, *Comprehend* is applied to a ray to produce a path that the ray could have taken through the box. *Comprehend* is then applied to this path to determine the location of atoms that could cause the path. *Comprehend* can also make use of expectations that have been placed in working memory by other operators. For example, *Comprehend* can compare the outcome of an experiment to expectations, thus bypassing the standard comprehension process. The implementation of *Comprehend* depends on the object being comprehended and on the available task knowledge. Hence the number of comprehension steps and the process or processes underlying each step must be derived from an analysis of the task as well as empirical observations of human subjects.

A similar multi-step comprehension process is used in NL-Soar, a system for comprehending natural language (Lewis, 1993). The NL-Soar designers found that this approach increased the generality of acquired comprehension knowledge and also contributed to the explanation of several behavioral regularities in natural language processing.

Refine attempts to refine abstract hypotheses by taking into account explanations in the current situation model. It is the primary mechanism for integrating evidence. This is done by considering hypotheses that have been accepted (such as atoms that have already been placed) as well as hypotheses that are being considered. For example, given an abstract hypothesis that an atom is in a column, *refine* would first check the column to determine whether an atom has already been placed in that column. If so, then it will attempt to use that atom to explain the datum. If not, then refine would check to see if any hypothesized atoms are in the column. If only one is present then this would be used to explain the datum.

Discriminate takes a disjunctive set of explanations and attempts to select one by evaluating each alternative with respect to the situation model. If there is insufficient evidence to select a single explanation, then *discriminate* will attempt to break the tie by collecting data (i.e., by designing and conducting an experiment).

Check determines whether new results (such as new explanations) are consistent with the other parts of the situation model. *Check* annotates the situation model with this information and can also add a certainty annotation to the item being checked.

Test designs and conducts an experiment to either confirm or disconfirm an uncertain item.

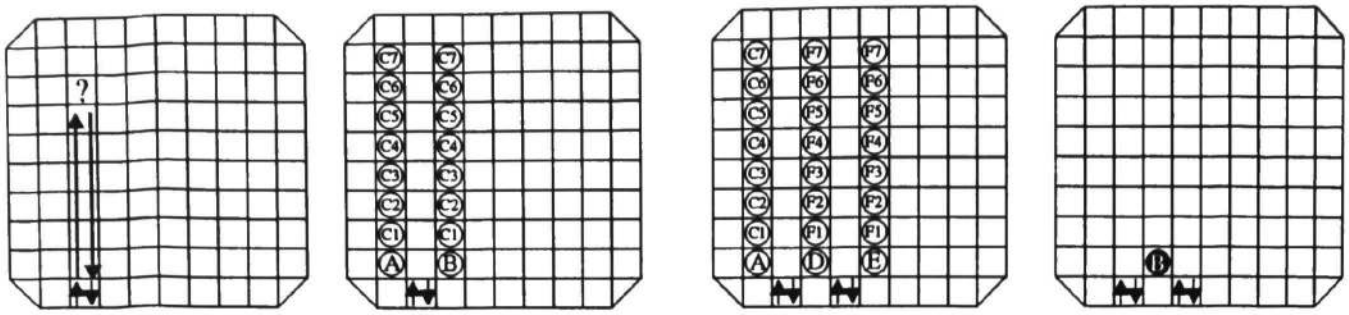
Resolve-anomaly takes anomalous parts of the situation model, such as two contradictory explanations, and makes appropriate changes to the situation model. A theory of anomalous data interpretation is given in Krems and Johnson (1994).

Hypothesize adds a disjunctive set of hypotheses to the situation model as explicit hypothesized components. This operator, in conjunction with *refine*, provides the primary evidence integration mechanism in the model. An evidence integration example is given below, following the next section.

Implementation and Example: Applying the Model to Black Box

The model described above has been applied to BBX and implemented in Soar. In BBX there are two comprehend operators. *Comprehend ray-shot* produces a path, possibly abstract, that explains how a ray travels from the input to the output cell. This is done using satisficing search through a space with operators that trace a path from one cell to another. *Comprehend path* then determines the atoms that are needed to support the path. It does this by reasoning backward over the rules of ray travel.

Figure 2 illustrates how the model works. The model begins by requesting the first datum. Upon seeing the result (a) the model applies *comprehend* to the data which produces an abstract path indicating that the ray went into the box, then turned 180 degrees and exited at the same point (also shown in 3a). *Comprehend* is then applied to this path (b) resulting in 3 hypotheses: 2 single atom hypotheses, *A* and *B*, and 1 two-atom hypothesis, *C*. The latter hypothesis is abstract, in that it specifies that a pair of atoms can be located at any position along the column. Abstract hypotheses are initially represented propositionally in the model. For example, C is represented as an atom pair that can occur anywhere along the third column. A nonpropositional model-based representation would add explicit concrete elements for each atom pair to the model.



(a) (b) (c) (d)

Figure 2: The use of evidence integration. Atom B (3d) is placed after the second reflection is seen in 3c.

Since more than one explanation has been generated, *refine* is applied to the hypotheses. *Refine* is unable to select an explanation, so *hypothesize* is applied, which adds the hypothesized atoms to the situation model as explicit objects.

Next *discriminate* is applied. This simply updates the state context to indicate that discrimination is being done. This allows the model to collect more data (c), another reflection. This second ray is comprehended just as the first, resulting in a similar set of three hypotheses: D, E and F, also shown in 3c. Following this, *refine* is applied to the new ray's hypotheses. Since no atoms have been placed, the model checks to see if an existing hypothesized atom will explain the new ray. It finds that *B* is the only hypothesized atom that can explain the new ray, so it updates the explanation for the ray to indicate that *B* explains the ray. It then places an atom at location *B* (d) and checks (using *check*) to see if the second ray is actually explained by simulating the ray shot and comparing the outcome of the simulation to the actual outcome. Since the second ray is explained, the model shifts attention back to the first ray. Because the situation model has changed, *refine* is reapplied to the first ray. This time, *refine* sees that the newly placed atom is consistent with the hypothesis for the first ray, so it updates the situation model to indicate that the first ray is explained by the atom. It then checks, by simulating, to ensure that the first ray has been explained. Since it is, the model is then free to collect additional data.

The above example illustrates several important features of the model-based theory. First, because of the satisficing nature of hypothesis generation, the model doesn't need to consider all possible explanations for a given datum. Second, although the system uses satisficing search, the example shows that the multi-stage comprehension process can make use of abstract hypotheses to generate a class of possible explanations for a datum.

Third, it illustrates how a complex abductive problem can be solved by the sequential application of relatively simple local reasoning processes that bring to bear different bodies of knowledge. The results of some of these processes feed into other processes, while some processes, such as *check*, independently check the results of others. Smith, et al. (1991) found that expert technologists use a similar technique to cope with the complexities of blood typing, an abductive task requiring the interpretation of a large set of test data. This is in sharp contrast to theories that assume that the entire set of data must be reevaluated each time a new datum is received or that assume that evidence can be brought to bear in parallel. Finally, evidence integration is done not by counting the data explained and not explained nor by combining probabilities, but simply by checking the model for previously hypothesized atoms that overlap with those hypothesized for the current datum.

Evaluation of the Model: Regularities Met

The plausibility of the model can be evaluated by comparing it to the behavior of subjects in abductive reasoning tasks. Here, we briefly review some phenomena discovered in previous studies as well as in our own studies based on the BBX task (see Johnson et al., 1993).

Order Effects

As noted earlier, the order of data presentation sometimes affects a person's belief in an explanation (Hogarth & Einhorn, 1992). In BBX, we found that order of data presentation can affect what components are used in the multicausal explanation (Bogenberger, In preparation). Figure 3 illustrates this effect. When ray A is presented first, subjects normally place Atom 1. If Ray B (entering at B1 and exiting at B2) is then presented, an atom is placed at cell 2. When C is shown next, atom 3 is placed. However, at this point an atom at either cell 4 or 2 could explain B with equal likelihood (p=.49), given all of the data currently available. If the data is reversed, C, B, A, then an atom is placed at cell 4 instead of 2.

According to our model, order effects occur because the simplest consistent explanation for a new datum is immediately added to the situation model and then used to constrain the explanation for succeeding data (through *refine*). By changing the order of data, the context in which explanations are refined is changed, leading to the selection of different explanations. In cases where *refine* is not used,

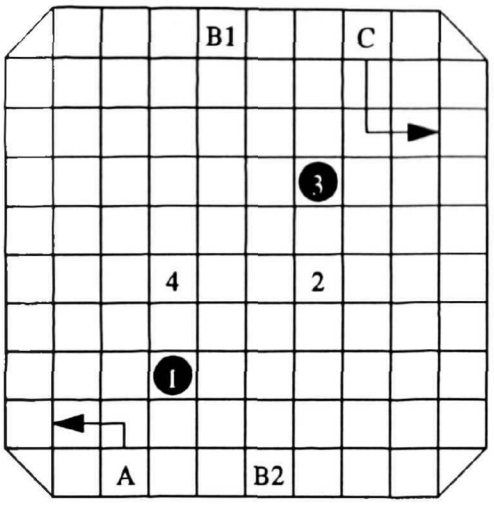


Figure 3: Order of data presentation affects outcome. A, B, C results in placement of atoms at locations 1, 2, and 3 whereas C, B, A results in a placement at locations 1, 3 and 4. Locations 2 and 4 are equally likely ($p = .49$).

no order effects will be seen.

The mental-model theory produces the same behavior as the subjects. When A is shown, Atom 1 is placed to explain it. If B is shown next, *comprehend* constructs an abstract "zig-zag" path from B1 to B2, then comprehension of this path results in knowledge that there must be atoms in the fourth and sixth columns. *Refine* is then applied to the path where it detects Atom 1, modifies the abstract "zig-zag," so that it uses Atom 1 and specifies that there must be an atom at Cell 2. Since there is now only one explanation for B, an atom is placed at location 2, then B is checked and Datum C is requested. The explanation for C (Atom 3) does not contradict any other explanation so the model requests additional data to locate the fourth atom. An alternative explanation for Ray B is not considered. When the data is presented in reverse order, the model places an atom at location 4.

Deviation from Normative Abduction

The example in Figure 3 illustrates how the mental-model theory predicts deviations from normative abduction. Given the data shown in Figure 3, locations 2 and 4 are equally likely locations for atoms; however, both locations are not considered because the order of data presentation together with the satisficing approach to explanation generation and selection resulted in a model in which all of the data is explained. In general, we expect that explanations produced by the model will sometimes differ from the answer dictated by probability theory because the model selects a best explanation based on ease of search, availability of alternatives, and model-based evidence integration. The model currently does not use, or even know, probabilities.

Task and Domain Effects

The application of the model to BBX illustrates how the task and domain influence the outcome (as contrasted to the outcome that would be produced by a purely syntactic theory). The comprehension of data (to produce an explanation) is a multi-stage process, where the number and nature of the stages depend on the domain and task. Furthermore, the implementation of each stage is determined by domain and task characteristics. For example, we observed that subjects tend to trace rays from input cell to output cell, despite the fact that all rays are bi-directional. This can affect the answer given. If the data in Figure 3 is presented in the order A, C, B, and B is shot in at B1, then subjects tend to place an atom at location 2, because they trace a path from B1 to B2 and notice Atom 1, but not Atom 3. However, if ray B is shot in from B2, then subjects tend to place an atom at location 4, because Ray B is traced from B2 to B1 and Atom 3 is noticed, but not Atom 1. Thus, task-specific processing has a major effect on the abductive conclusions. In the application of the mental-model theory to BBX, we implemented *Refine* so that it would trace from input cell to output, thus our model is able to replicate this behavior. Any model of abductive reasoning that completely abstracts away the properties of a domain would be unable to explain this behavior.

Power-law Speed-up

To test for power-law speed-up we ran the model 5 times on 52 games (each run used a different random ordering of the same games). We then compared the speed-up to 5 human subjects playing the same 52 games. For the simulation, we found that a power-function ($y = x^{-0.2579} \cdot 607.89$) explains 74% of variance. The linear fit is: $y = -4.305x + 404.74$, explaining 57% (substantially less). The mean number of Soar decision cycles (DCs) (over all model runs and games) is 290.66 per game. This drops from 593 DCs to 250 after 52 games. The minimum value was 174.

For the 5 subjects on the same 52 games, we found that a power law ($y = x^{-0.4816} \cdot 192529ms$) explained 50% of variance while a linear fit ($y = -1799x + 104066ms$) only explained 28%. The subjects improved from 263 sec's to 48 sec's per game. They averaged 56.4 sec's per game (for all games). The minimum time was 17.8 sec, the maximum, 332.9 sec.

The Soar architecture specifies that a decision cycle (DC) corresponds to a value in the range of 30 ms to 300 ms (Newell, 1990). With 290 DCs for the model and 56.4 sec for the subjects the simulation is operating at 195ms/DC, which is well within the theoretical range. But the speed-up of the model is much smaller than that of the subjects: 58% compared to 83%. The learning rate (power law coefficient) of the model is approximately half of the

subjects (0.25 compared to 0.48). In general, this means that the model reaches the asymptote more quickly (i.e., the model stops improving earlier), but that after learning the task is approximately of the same complexity for both the model and the subjects. The difference in learning rate likely occurs because the trial time for subjects includes the time to visually scan the screen, move the mouse and click the mouse button. The model does not attempt to simulate these actions whereas with the subjects these actions are getting automated and appear in the speed-up effect. With the model we only see cognitive speed-up, the commands that interact with the BBX display never get faster.

Conclusion

The theory of abduction described above shows how the interaction of relatively simple symbolic processes can account for complex behavioral regularities. The model views human abductive behavior not as the imperfect application of formal syntactic laws or probability estimation, but as the process of building and modifying an explicit situation model using satisficing search. This is similar to the mental model theory of human deductive reasoning. Because the model is based on satisficing search, it makes reasonable assumptions about the human cognitive architecture. This is in contrast to models of abductive reasoning that assume massively parallel computation or the ability to remember and accurately combine probabilities. Our model is similar to many search based theories of scientific discovery, such as Klahr and Dunbar's dual space search model, but it offers more details of the problem solving processes as well as process models for evidence integration and skill acquisition.

Acknowledgements

This work has benefitted from comments and suggestions of Jack W. Smith, Kathy Johnson, Ayse Bayazitoglu, Frank Ritter, John Josephson, B. Chandrasekaran and 3 anonymous reviewers. Thomas Rothenfluh created the original help screens for BBX and provided many suggestions during the early stages of the research. Hans Bogenberger programmed the Windows version of BBX and collected the data for the skill acquisition study. This work is supported by a German-American Collaborative Research Grant from the American Council of Learned Societies and the German Academic Exchange Program. Additional support was provided by a Seed Grant from The Ohio State University.

References

Bogenberger, H. (In preparation) Universität Regensburg: unpublished thesis (Diplomarbeit).

Downing, C. J., Sternberg, R. J. & Ross, B. H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General, 114*, 239-263.

Dunbar, K. & Klahr, D. (1989). Developmental differences in scientific discovery processes. In D. Klahr & K. Kotovsky (Ed.), *Complex Information Processing: The Impact of Herbert Simon* (pp. 109-143). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hogarth, R. M. & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Johnson, T.R., Krems, J.F. & Amra, N.K. (1993). A proposed model of human abductive skill and its acquisition. LKBMS Technical Report. Columbus, Oh: The Ohio State University.

Johnson, T. R. & Smith, J. W. (1991). A framework for opportunistic abductive strategies. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, (pp. 760-764). Chicago: Lawrence Erlbaum Associates.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991) *Deduction.* East Sussex: Lawrence Erlbaum Associates.

Josephson, J., Chandrasekaran, B., Smith, J. & Tanner, M. (1987). A mechanism for forming composite explanatory hypotheses. *IEEE Transactions on Systems, Man, and Cybernetics, 17*(3), 445-454.

Josephson, J. & Josephson, S. (Eds.). (1994). *Abduction: Computation, Philosophy, Technology.* Cambridge University Press.

Klahr, D. & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1-48.

Krems, J.F. & Johnson, T.R. (1994). Interpretation of Anomalous data. Technical Report. Laboratory for knowledge-Based Medical Systems. Columbus, Oh: The Ohio State University.

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

Lewis, R. L. (1993) *An Architecturally-based Theory of Human Sentence Comprehension.* Ph.D. Thesis, Computer Science Department, Carnegie-Mellon University.

Newell, A. (1990) *Unified Theories of Cognition.* Cambridge, MA: Harvard University Press.

Schustack, M. W. & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110*, 101-120.

Smith, P., Galdes, D., Fraser, J. M., Miller, T. E., Smith, J. W., Svirbely, J. R., Blazina, J., Kennedy, M., Rudmann, S. & Thomas, D. L. (1991). Coping with the complexities of multiple-solution problems: A case study. *International Journal on Man-Machine Studies, 35*, 429-453.

Thagard, P. (1989). Explanatory Coherence. *Behavioral and Brain Sciences, 12*, 435-502.

Tversky, A. & Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Ed.), *Judgment Under Uncertainty: Heuristics and Biases* (pp. 3-31). Cambridge: Cambridge University Press.