Integration of Anomalous Data in Multicausal Explanations

Josef Krems
University of Regensburg
Department of Psychology
D-93053 Potsdam, Germany
krems@psychologie.uni-regensburg.de

Todd R. Johnson
Division of Medical Informatics and
Center for Cognitive Science
The Ohio State University
Columbus, Ohio 43210
tj@medinfo.ohio-state.edu

Abstract

This paper describes and evaluates a computational model of anomalous data integration. This model makes use of three factors: entrenchment of the current theory (the amount of data explained), the relative probability of the contradictory explanations (based on conditional probabilities as part of the domain-knowledge), and the availability of alternative explanations based on learning. In an experimental study we found that the entrenchment of a theory and the availability and likelihood of an alternative explanation influenced solution speed and the correctness of inferred causal explanations. However, in detail, the single levels of both factors were not clearly distinguishable and did not follow the predictions. These findings suggest that entrenchment itself is not a major factor in determining the difficulty of a task. Instead, we hypothesize that task difficulty is dominated by a person's ability to construct an alternative explanation of a given situation, a factor that is only indirectly related to entrenchment.

Introduction

Integrating anomalous data with an existing theory or explanation is an essential subtask in scientific discovery, diagnostic reasoning and in everyday problem solving such as story understanding. In this paper we focus on the integration of anomalous data into an existing multicausal explanation for a set of observations. In its simplest form, a causal inference has the following form: Given knowledge that *A* causes *B*, upon observing *B*, *A* is hypothesized as the explanation for *B*. This is a kind of abductive inference (Josephson & Josephson, 1994). In multicausal abductive tasks the explanation is composed of multiple causal hypotheses, which together explain the observations. An anomaly occurs when new evidence contradicts the existing explanation. The general problem then is to decide how to modify the multicausal explanation, so that all evidence, including the new observation, is explained. We have designed and implemented a mental model based theory of abduction in Soar (see Johnson, Krems & Amra, 1994, for details) for which we have proposed a mechanism of anomalous data interpretation. This paper describes this mechanism and also presents results of an experimental study in which the cognitive plausibility of the mechanism was evaluated.

Multicausal Explanations and Anomalies

In abductive reasoning, an anomaly occurs whenever new evidence contradicts one or more hypotheses in the existing multicausal explanation for previously given evidence. New evidence can contradict the existing explanation in one of two ways: 1) The new evidence is logically inconsistent with the existing explanation, such that there is no way to explain the new evidence without modifying the explanation; or 2) The hypothesis chosen to explain the new evidence contradicts the existing explanation, but a different hypothesis for the new evidence is consistent with the explanation.

When an anomaly occurs, the reasoner must either modify the old explanation so that it is valid for both the new and pre-existing evidence, or select a different hypothesis for the new data so that the new hypothesis is consistent with the explanation for the old evidence. Our major research question is to clarify the factors that affect this decision and the processes used to make the decision.

Previous studies of the interpretation of anomalous data provide evidence on the role of various factors, such as entrenchment of a theory, the availability and likelihood of an alternative explanation, and a subject's background knowledge. Chinn and Brewer (1993) argue that the entrenchment of a theory is one of the characteristics of an individuals current beliefs that influence how a person responds to anomalies. One way theories are entrenched is due to the amount of evidence they explain. Applied to abductive reasoning this should mean that theory-preserving responses should covariate with the amount of data already explained by the current theory. The literature on the confirmation bias (e.g., Klayman & Ha, 1987; Krems, 1994) as well as studies by Burbules and Linn (1988) also indicate that the availability and likelihood of an alternative hypothesis can influence a person's response to anomalous data.

Although researchers have proposed several models of scientific discovery and abductive reasoning, most do not provide a detailed process model of anomalous data interpretation. For example, Dunbar and Klahr's (1989) model (SDDS, Scientific Discovery as Dual Search) shows how explanations are formed and modified by searching in hypothesis and experiment spaces, but does not provide a

detailed description of what happens when new data contradicts the current explanation. Thagard's (1992) theory of explanatory coherence (TEC) offers an account of how anomalous data affects the strength (or coherence) of new and existing hypotheses; however, it does not offer a theory for how people use belief changes to decide how to modify the explanation so that it can account for both the new and old data. TEC does imply, however, that a person would attempt to retain the most coherent hypotheses. Thus, it seems reasonable for a model based on TEC to search for alternative hypotheses to replace the less coherent, contradictory hypotheses. Theories based on case-based explanation generation (Schank, Riesbeck and Kass, 1994) emphasize the role of prior experience in explaining anomalies. Read and Cesa (1991) showed that expectation failures are important cues for retrieving relevant memories of previous anomalies. However, little is known about the process of explanation modification.

A Computational Model

Basic Features of the Model

In previous work, we developed a mental model based theory of abduction and implemented it in Soar (Newell, 1990). For details of the model see Johnson, Krems and Amra (1994). We view abduction as the sequential comprehension and integration of data into a single situation model that represents the current best explanation of the data. Suppose that a new datum is available. First, the situation model is updated to include the new datum. Next, the new datum is comprehended, i.e., knowledge is brought to bear to determine what the new datum implies about the situation. Comprehension results in one or more explanations for the datum, where each explanation consists of one or more hypotheses together with the data they explain. If the generated explanation is inconsistent with any hypotheses or data in the existing situation 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.

Processing Anomalous Data

The model responds to anomalous data by rejecting all but one of the anomalous explanations and then constructing alternative explanations for the data left unexplained by the rejection(s). It does this using a limited lookahead search to determine which explanation is the best to keep. Suppose that explanations, *e1* and *e2*, for two data, *d1* and *d2*, are inconsistent. In the lookahead search it first selects one of the explanations, say *e1*, and rejects it. Then it searches for an alternative explanation for *d1* and evaluates the resulting situation model. Next it returns to the original anomalous situation, rejects *e2*, searches for an alternative to explain *d2* and evaluates the results.

The model then rejects the explanation whose rejection resulted in the best situation model (where *best* is defined as the model that explains the most data with the fewest number of explanatory components). For example, if rejecting *e1* results in a better alternative explanation (than that found by rejecting *e2*), then *e1* will be rejected and the alternative explanation for *e1* will be used.

If an alternative explanation for one of the data items cannot be found (either because none exists or because processing limitations prevent adequate search), then the explanation for that datum will be retained and the explanation for the other datum will be modified. If rejecting *e1* and *e2* result in equally good situation models, then the difference between the probabilities (if known) that *e1* explains *d1* versus that *e2* explains *d2* is used to break the tie. The probability (or frequency) that a given set of evidence is explained by a certain set of causes is not calculated by the model but is considered to be part of the domain knowledge. A number of studies reveal that people can implicitly acquire such frequency of occurrence information and then use it during decision making (see Hasher & Zacks, 1984).

Thus, to decide which explanation to reject, the model makes use of three factors: entrenchment of the current theory (the amount of data explained), the relative probability of the contradictory explanations (based on conditional probabilities as part of the domain knowledge), and the availability of alternative explanations.

We assume that the availability of alternative explanations depends on situation-specific knowledge and the amount of time spent searching for an alternative. The more often a situation is faced in which the existing explanation is replaced by an alternative, the more likely it is that the person has generated an appropriate alternative for that explanation. This means that availability of potential explanations should increase with problem solving experience. It also means that subjects' confidence in their explanations should correlate with the relative frequencies that the explanations were correct for a set of data.

The Task

To explore human abductive problem solving we use a task called Black Box (BBX). In this task, four atoms are hidden in a box (an 8 x 8 matrix) and the player's goal is to discover their locations by shooting rays into the box (the subjects are trained on these rules prior to introducing them to the abductive task). 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

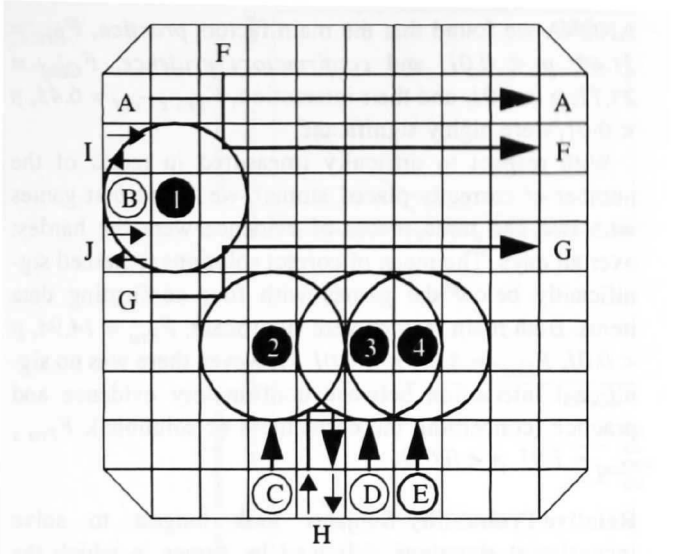

Figure 1: The Black Box with four atoms and the paths of several light rays visible.

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 player does not actually see the path that the ray follows; hence, the path must be inferred.

We selected Black Box for three primary reasons. First, it shares many features with real-world abductive tasks such as device diagnosis and medical test interpretation. These similarities include: 1) Additional data must be collected based on the current working hypothesis; 2) The data can be decomposed into subsets such that the data in a subset can be explained by the same hypothesis; and 3) A single datum can require multiple individual hypotheses to explain. Second, Black Box is easy to understand—subjects easily learn the rules of the task within one hour of training. Third, one of the major problems with many studies of abductive reasoning (such as those done in medical domains or natural scientific domains) is the difficulty in controlling for background knowledge differences between subjects. By using a simple domain like Black Box we can ensure that all subjects have the same knowledge of the device and that no additional external knowledge is given to the subjects.

Anomalous Data Interpretation in BlackBox

A typical example of an anomaly in Black Box and two ways to resolve it are illustrated in Figure 2. In Figure 2a the subject sees Ray A and places Atom 1. Next, Ray B is shot and the subject assumes that the ray actually traveled straight through the box as shown in Figure 2a. This explanation is anomalous, however, with the explanation for Ray A, because Atom 1 will cause Ray B to turn to the right. The typical response, at this point is to assume that Atom 1 is incorrect and to generate an alternative explanation for Ray A. Figure 2b shows one possible alternative in which Atom 1 is removed and Ray A is explained using three different atoms. Figure 2c, however, shows a completely different possibility in which Ray A is still explained by Atom 1, but Ray B is explained by an alternative configuration. Thus, the existence of an anomaly depends on how the data is initially explained.

Entrenchment in this task refers to the number of rays accounted for by a certain atom. The model predicts that the higher this number is before anomalous data are seen the harder it is to give up the current explanation. Relative probability indicates which of the two contradictory explanations is more probable. Specifically, relative probability is the ratio of probabilities between the two contradictory explanations—for example, between the probability that Ray A is explained by the path shown in Figure 2a and the probability that Ray B is explained by the straight path in Figure 2a. The model predicts that the less probable explanation has a greater chance of being modified. Finally, the model predicts that availability of an explanation will increase with level of practice. Therefore, level of practice should affect a person's responses to anomalous data.

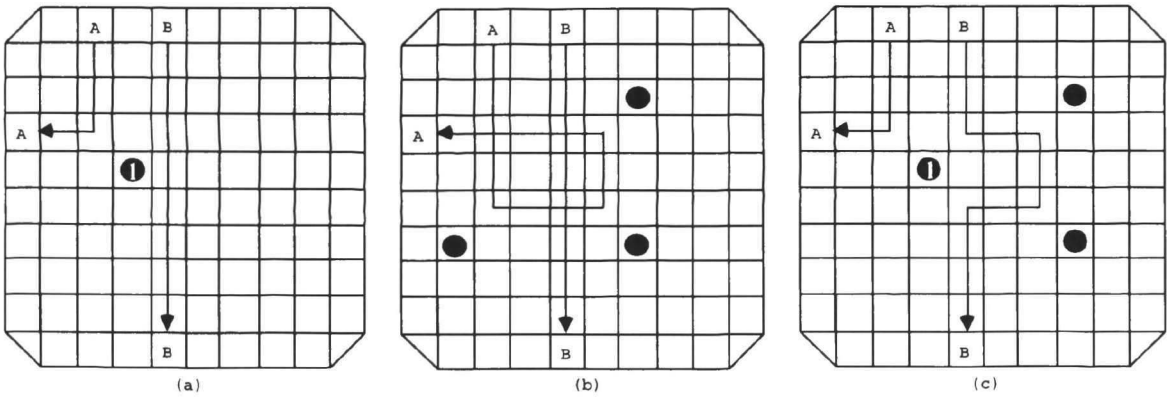

Figure 2: (a) illustrates an anomalous situation; (b) and (c) illustrate two ways to resolve the anomaly.

Experimental Evaluation of the Predictions

The predictions described in the last section were investigated in an experimental study. Ten undergraduate students at the University of Regensburg played on five consecutive days a total number of 185 games (25 training games and 12 test games every day).

Design and Procedure

A 5 (level of practice) x 3 (confirmatory evidence) x 4 (relative probability) within-subjects design was used. The factor *level of practice* has five levels: one for each day of training. *Confirmatory evidence* was measured in terms of the number of ray markers an explanation (one or more atoms) accounts for. This factor varied from 2 to 4. *Relative probability* refers to the ratio of the probability of the old versus the new explanation. This factor had four levels: equal, new explanation is less likely than the old explanation, new explanation is more likely, and new datum absolutely contradicts the explanation for the old data, i.e., the new datum cannot be explained without modifying the existing explanation. The relative probabilities are based on the frequencies that were computed for all possible combinations of ray patterns and atoms in BBX (e.g., we know that 86.7% of all absorptions are explained by a single atom). Combining these two factors results in 12 different combinations.

Every session consisted of a training phase followed by a test phase. In the training phase, subjects were trained on 25 randomly generated games. In the test phase, subjects were presented with 12 critical cases (one for each of the above mentioned combinations) containing anomalies. The subjects' task in the training and the test games was to develop an explanation of rays by placing atoms. In a modified version of the original BBX game, subjects could place atoms, remove atoms or ask for new data. New data was requested by clicking on a button, labeled "More Data," which highlighted one of the perimeter cells of the Black Box matrix. This told the subject where the ray would be shot into the box. Clicking on this perimeter cell revealed the outcome of the ray shot. Thus the data was presented sequentially as it would be if the subjects had actually shot the rays themselves. By not allowing the subjects to select their own ray shots, we could use predefined games and therefore better control the situations presented to the subjects during the trials. After placing an atom, the subjects gave a confidence judgment that the location was correct. The confidence scale consisted of seven categories (between guessing and certain). All atom placements, removals and data requests were time-stamped and recorded.

Results

Amount of evidence. On average, it took subjects less time to solve games with two or four pieces of confirming evidence than games with three data items (see Figure 3). In an ANOVA we found that the main factors *practice*, $F_{Pra} = 31.44$, $p < 0.01$, and *confirmatory evidence*, $F_{Conf} = 23.71$, $p < 0.01$, and their interaction, $F_{Pra \times Conf} = 6.43$, $p < 0.01$, were highly significant.

With respect to difficulty (measured in terms of the number of correctly placed atoms), we found that games with two and three pieces of evidence were the hardest over all days. The mean of correct solutions remained significantly below the games with four confirming data items. Both main factors were significant, $F_{Pra} = 14.94$, $p < 0.01$, $F_{Conf} = 8.21$, $p < 0.01$. However, there was no significant interaction between confirmatory evidence and practice (concerning the correctness of solutions), $F_{Pra \times Conf} = 1.91$, $p < 0.071$.

Relative Probability. Subjects took longest to solve inconsistent situations, followed by games in which the new explanation was more likely or as likely as the existing one. Games in which the old explanation was more likely were comparatively quick to play, $F_{Relpr} = 7.7$, $p < 0.001$. Solution time decreased constantly with level of practice, $F_{Pra} = 19.2$, $p < 0.000$, $F_{Pra \times Relpr} = 1.13$, $p < 0.34$. The mean of correct solutions was lowest with inconsistent situations, followed by equal and then by new and old, $F = 10.5$, $p < 0.01$. New and old switched after the first two trial sessions (see Figure 3). There was a general improvement between the training sessions, $F_{Pra} = 14.93$, $p < 0.01$ With respect to correctness, no interaction between the factors *relative probability* and *level of practice* could be found, $F_{Pra \times Relpr} = 0.67$, $p = 0.8$.

Confidence-Rating. The model predicts that anomalous situations will lead subjects to search for alternative explanations, hence with practice on anomalous situations people should become more aware of alternatives. This means that the confidence ratings should increase for explanations that rarely proved to be wrong, but decrease for explanations that were frequently wrong due to anomalous data. The first assumption could be verified, $\chi^2 = 28.94$, $p < 0.00$. For the second hypothesis, however, statistical tests showed that the categorical variables *confidence* (7 categories) and *level of practice* (5 days) are independent, $\chi^2 = 13.24$, $p > 0.05$. This means that subjects stabilized their judgment in cases where the current explanation remained correct but they did not become "more careful" or "more uncertain" for anomalous situations. Even after having seen a set of counterexamples that made it necessary to give up a current explanation, subjects continued placing these atoms with a degree of confidence that did not change based on experience.

Discussion and Conclusions

The amount of evidence and the likelihood of an alternative interpretation clearly influenced the modification of explanations. This is, in general, consistent with the find-

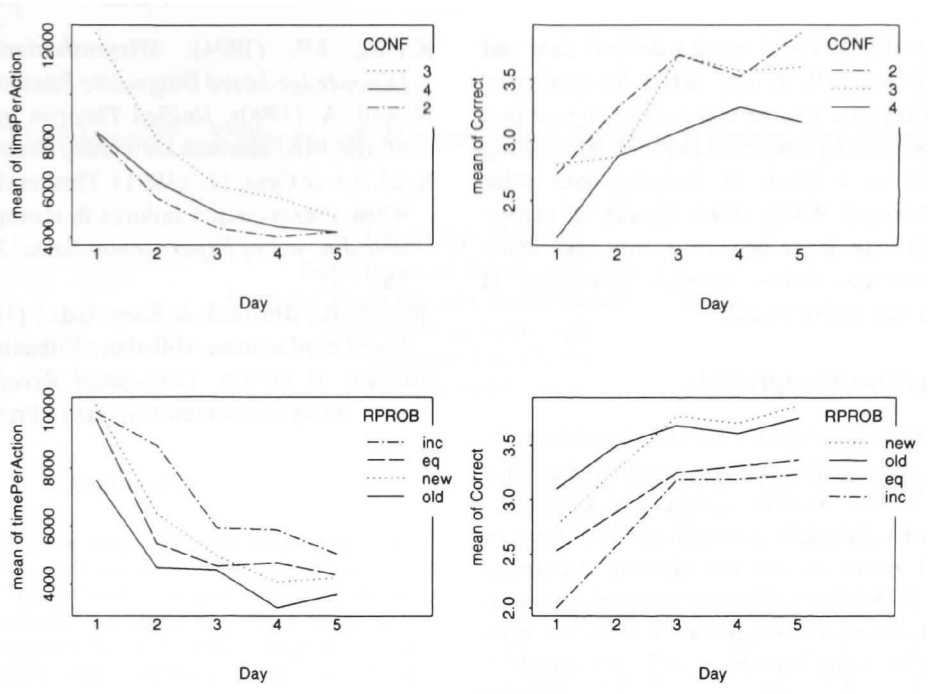

Figure 3: Improvement across level of practice (day) dependent on the confirmatory evidence of the existing explanation (above) and the relative probability of the alternatives (bottom). The graphs on the left show improvement in terms of mean time per action (ms), where an action consisted of placing or removing an atom or asking for more data. The right-most graphs show improvement in terms of the number of atoms correctly placed. [CONF—confirmatory evidence, RPROB—relative probability with inc—inconsistent, eq—equal.]

ings of Chinn and Brewer (1993) and also with the literature on cognitive biases (e.g. Klayman & Ha, 1987). However, in detail, the single levels of both factors were not clearly distinguishable and did not follow the predictions.

Based on these findings we assume that entrenchment itself is not a major factor in determining the difficulty of a task. Instead, we hypothesize that a person's ability to *construct* an alternative explanation of a given situation affects task difficulty. We hypothesize that in a situation with anomalous data a person selects an explanation to modify based on awareness of alternatives for that explanation and on the possibility to explain all conflicting data by an alternative. If such an alternative explanation can be developed, then the simplest version that requires the smallest number of changes to the existing explanation, but that explains the most data, is selected. This is indirectly connected to the amount of confirming data. The greater this number is for the current theory the harder it may be to find an alternative that explains this data as well as the conflicting data. This is consistent with the model we outlined earlier.

Note, the necessity to construct an alternative explanation is a specific feature of the BBX task. Anomalous data in this task cannot just be ignored by rejection or exclusion, in contrast to the task used by Chinn and Brewer (1993). Therefore entrenchment and relative probability are dominated by a search for (and availability of) an alternative explanation. According to our model, search for an explanation ends as soon as a single satisfactory explanation is found. The empirical results on the confidence rating support this assumption since subjects' confidences in their initially constructed explanations remain constant even after seeing anomalous situations in which they had to construct alternative explanations for the same data. Thus, a subject's confidence in an explanation was independent of the relative frequencies that a certain explanation is correct for a pattern of data. This suggests that the confidence in a hypothesis is either dominated by the ease with which a hypothesis can be initially generated from domain knowledge or by parsimony.

The subjects not only got faster, but also achieved better accuracy (see Figure 3). We hypothesize the following explanation based on bounded search and knowledge compilation. According to the bounded search hypothesis, people will only expend a limited amount of effort before terminating a search with failure. These searches are not done in vain, however, because knowledge compilation will compile many of the steps of the terminated search. The next time a search is done, the person will be able to reach the previous point of the search process with less effort, due to the compiled steps. This allows the person to

search further. Eventually they will reach a desired state and generate a solution which will, in turn, affect the quality of their performance. Thus, we assume that people engage in a type of progressive deepening where the depth of succeeding searches is extended as a result of learning over prior searches (Johnson, Zhang & Wang, 1994). Quality of performance improves with experience because it may take many terminated search attempts before enough knowledge is compiled to permit a successful search.

Acknowledgments

This work has benefitted from comments and suggestions by Kathy Johnson. Hans Bogenberger programmed the Windows version of BBX. This work is supported by a German-American Collaborative Research Grant from the American Council of Learned Societies and the German Academic Exchange Program (DAAD). Additional support was provided by a Seed Grant from The Ohio State University and a grant from Vielberth-Stiftung, University of Regensburg.

References

Burbules, N.C. & Linn, M.C. (1988). Response to contradiction: Scientific reasoning during adolescence. *Journal of Educational Psychology, 80*, 67–75.

Chinn, C. A. & Brewer, W. F. (1993). Factors that influence how people respond to anomalous data. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 318–323). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Hasher, L. & Zacks, R. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist,* 39, 1372–1388.

Johnson, T.R., Krems, J.F. & Amra, N.K. (1994). A computational model of human abductive skill and its acquisition. In A. Ram & K. Eiselt (Ed.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society,* (pp. 463–468). Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, T.R., Zhang, J. & Wang, H. (1994). Bottom-up recognition learning: a compilation-based model of limited-lookahead learning. In A. Ram & K. Eiselt (Ed.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 469-474). Lawrence Erlbaum Associates.

Josephson, J.R. & Josephson, S.G. (1994). *Abductive Inference*. Cambridge: University Press.

Klayman, J. & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211–228.

Krems, J.F. (1994). *Wissensbasierte Urteilsbildung [Knowledge-based Diagnostic Reasoning]*. Bern: Huber.

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

Read, S.J. & Cesa, I.L. (1991). This reminds me of the time when...: expectation failures in reminding and explanation. *Journal of Experimental Social Psychology*, 27, 1–25.

Schank, R., Riesbeck & Kass. (eds.) (1994). *Inside Case-based Explanation*. Hillsdale: Erlbaum.

Thagard, P. (1992). *Conceptual Revolutions*. Princeton, New Jersey: Princeton University Press.