ELSEVIER

Cognitive Science 25 (2001) 903–939

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Use of current explanations in multicausal abductive reasoning

Todd R. Johnson[a,]*, Josef F. Krems[b]

[a]*University of Texas Health Science Center at Houston, Houston, TX, USA*
[b]*Chemnitz University of Technology, Chemnitz, Germany*

Abstract

In multicausal abductive tasks a person must explain some findings by assembling a composite hypothesis that consists of one or more elementary hypotheses. If there are *n* elementary hypotheses, there can be up to 2^n composite hypotheses. To constrain the search for hypotheses to explain a new observation, people sometimes use their current explanation—the previous evidence and their present composite hypothesis of that evidence; however, it is unclear when and how the current explanation is used. In addition, although a person's current explanation can narrow the search for a hypothesis, it can also blind the problem solver to alternative, possibly better, explanations. This paper describes a model of multicausal abductive reasoning that makes two predictions regarding the use of the current explanation. The first prediction is that the current explanation is not used to explain new evidence if there is a simple (i.e., nondisjunctive, concrete) hypothesis to account for that evidence. The second prediction is that the current explanation is used when attempting to discriminate among several alternative hypotheses for new evidence. These hypotheses were tested in three experiments. The results are consistent with the second prediction: the current explanation is used when discriminating among alternative hypotheses. However, the first prediction—that the current explanation is not used when a simple hypothesis can account for new data—received only limited support. Participants used the current explanation to constrain their interpretation of new data in 46.5% of all trials. This suggests that context-independent strategies compete with context-dependent ones—an interpretation that is consistent with recent work on strategy selection during problem solving. © 2001 Cognitive Science Society, Inc. All rights reserved.

Keywords: Abduction; Hypothesis generation; Cognitive modeling; Explanations

* Corresponding author. Tel.: +1-713-500-3921; fax: +1-713-500-3929.
E-mail address: todd.r.johnson@uth.tmc.edu (T.R. Johnson).

PII: S0364-0213(01)00059-3

1. Introduction

Abduction is the problem of finding a best explanation for a set of observations (Josephson & Josephson, 1994; Peng & Reggia, 1990). It is an essential feature of many tasks, including medical diagnosis (Feltovich, Johnson, Moller & Swanson, 1984), scientific discovery (Thagard, 1989), and discourse comprehension (Kintsch, 1988). In all such tasks, people explain sets of observations by generating and integrating hypotheses to form a best explanation. Such problems are often quite complex due to the number of possible elementary hypotheses for each observation and the many different ways to combine these hypotheses into an explanation. Despite this complexity, people often manage to solve these problems with a high degree of speed and accuracy. A central issue is how people cope with this complexity.

One important factor for controlling complexity is a person's present theory or current explanation of the situation. A new observation can have many possible explanations if considered separately; however, by considering only hypotheses that are consistent with a person's current explanation of the situation, he or she can greatly reduce the number of hypotheses to consider. For example, when reading a story, the reader's understanding of previous sentences and paragraphs constrains his or her interpretation of succeeding sentences (Kintsch, 1988). Likewise, in diagnosis and scientific discovery, a person's working hypothesis often constrains their interpretation of new observations (Elstein, Shulman & Sprafka, 1978). Pennington and Hastie (1988) have also shown that jurors interpret new evidence in the context of their explanation of existing evidence. In this paper, we refer to this set of available evidence and hypotheses as a person's *current explanation.*

Although the current explanation can limit the number of hypotheses considered to explain new evidence, it can also lead to suboptimal behavior, which occurs when interpreting new data in the context of the current explanation leads a person to overlook better explanations. For example, a person's prior attitude regarding a social issue often biases their interpretation of new evidence (Lord, Ross & Lepper, 1979). Pennington and Hastie (1988) showed that the order in which jurors are given evidence reliably affects their verdict. For instance, if the evidence is given in an order that allows jurors to easily construct a story for the defendant's guilt, jurors tend to return a guilty verdict.

Other research suggests that experts compensate for the limitations of hypothesis-driven reasoning by mixing it with data-driven reasoning. Smith et al. (1991) found that expert blood bank technologists used a mix of hypothesis-driven and data-driven reasoning to interpret tests designed to identify red blood cell antibodies. This task is complex because there are 27 different clinically significant antibodies, patients can have more than one antibody, and each test typically has several hundred data points. The researchers found that experts controlled this complexity by using data-driven reasoning to rule out antibodies and suggest highly plausible antibodies, followed by hypothesis-driven reasoning to build a story about which remaining antibodies could explain the data. The hypothesis-driven reasoning helps the expert cope with the complexity of the task, whereas the data-driven reasoning helps ensure that the expert is not blinded to plausible alternative explanations.

This paper examines the role of the current explanation in the interpretation of new evidence. It does this in the context of a model of abductive reasoning that predicts that the current explanation is not used whenever there is a single concrete hypothesis to account for the new evidence; but is always used when discriminating among alternative hypotheses for new evidence. The model views abduction as the sequential interpretation and integration of new data and hypotheses into a single explanation or situation model. The paper presents three experiments designed to test the model's predictions concerning the use of the current explanation.

2. Abduction

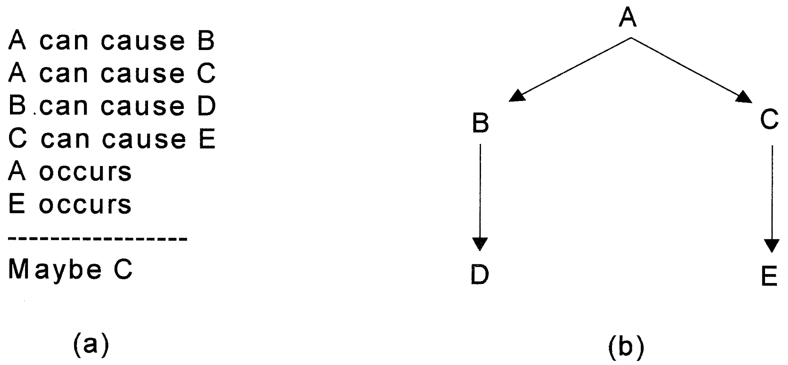
2.1. Simple abduction

In its simplest form, abduction is the process of inferring an explanation for an observation (C. S. Peirce, 1839–1914). Thus, given knowledge *If P then Q,* upon observing *Q, P* can be hypothesized as an explanation for *Q*. Whether *P* is the correct explanation depends on whether there are other explanations for *Q*. If alternative explanations exist, then the most plausible explanation for *Q* must be selected based on other knowledge, such as the relative frequency of occurrence of each explanation.

Abduction differs from induction and deduction in the following ways. In abduction, one infers an antecedent *P* given a rule *if P then Q* and its consequent *Q*. In contrast, deduction is the process of reasoning forward from the antecedent *P* to the consequent *Q*, given the rule *if P then Q* and *P*. Induction is the process of inferring a rule, *If P then Q,* given information about the relationship between *P* and *Q*. For example, if you notice that your car consistently refuses to start whenever it is below freezing, you can induce the rule: if below freezing then car will not start. Once you have this rule you can use deduction to predict that when the temperature is below freezing your car will not start. Likewise, you can use abduction to conclude that if your car will not start then the temperature *might* be below freezing.[1]

Although abduction is often viewed as reasoning from effect to cause, many real world abduction problems involve finding a hidden state of a system given causal knowledge of the system and indirect evidence of the hidden state. For example, diagnostic tests are often performed by manipulating inputs to a device (e.g., giving specific drugs to a patient) and observing the resulting output (e.g., measuring the degree of absorption of the drug). An abstract example of this type of abduction is shown in Fig. 1. The evidence consists of the beginning of the causal chain (A) and the final result or output (E), allowing us to abduce the hidden state C.

In this form of abduction, a person can generate a hypothesis by reasoning forward (from input to output), backward (from output to input), or both. This hypothesis generation process is easily captured by problem space search, where the states represent device states, the initial state is the device state capturing the input to the device, the goal state is the device state showing the observed output, the operators represent transitions from one device state to another. The current explanation can provide constraints for guiding search through the problem space. Each path from the initial to the goal state is a potential hypothesis.



A can cause B
A can cause C
B can cause D
C can cause E
A occurs
E occurs

Maybe C

Fig. 1. Abductive inference of a hidden state given evidence consisting of the beginning and end of a causal chain. (a) The logical form of the argument. (b) A graphical description of the causal rules shown in 1a.

2.2. Multicausal abduction

In multicausal abduction, the best explanation consists of a set of elementary hypotheses that together comprise the explanation for a set of observations. Fig. 2 illustrates a typical multicausal abductive problem. E1-E4 are evidence (or observations) to be explained. H1-H8 are elementary hypotheses that each explain some portion of the evidence. The arrows indicate which observations are explained by which hypotheses. Elementary hypotheses can explain zero, one, or several observations. For instance, H4 and H5, taken together, explain E2, but H5 by itself cannot explain anything. This is an example of two hypotheses that *together* produce an effect that neither can produce alone. The multicausal abduction problem is to select the set of elementary hypotheses that *best* explains the evidence. In this example, several different subsets of the hypotheses can completely explain the data, including (H1, H2, H7, and H8) and (H1, H4, H5, H6, and H8). Other hypothesis subsets explain portions of the data. However, complete explanations—those that explain all the

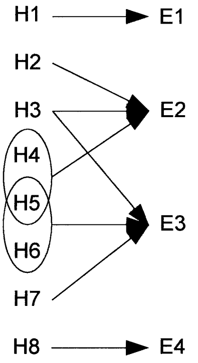


Fig. 2. Multicausal abduction. The problem is to find the best subset of elementary hypotheses (from H1-H8) that together explain the evidence (E1-E4). Arrows point from hypotheses to the evidence they explain. H4 and H5, taken together, explain E2. H5 and H6 together explain E3. Hypotheses H1, H3 and H8 form a multicausal explanation of all the evidence.

observations—are not necessarily the best. Other factors, such as plausibility, parsimony, and consistency must be considered.

The above formulation of abduction assumes that all the relevant causal knowledge is known about how a device works. While this is true of many real-world problems, there are many cases in which this is not true. It is important to note that this paper deals only with hypothesis generation given complete causal knowledge of a device.

There is widespread belief that Bayesian approaches cannot be used for multicausal abduction (see Peng & Reggia, 1990 for a discussion of this issue). Typically, two arguments are given against the use of these approaches. The first is that they require all hypotheses and evidence to be stochastically independent—a requirement that cannot be met in multicausal abduction. The second is that they require a vast number of probabilities—one for each possible compound hypothesis, however, in most cases there is no way to discover or compute these probabilities. Modern Bayesian approaches, however, such as those developed for Bayesian belief networks (e.g., Pearl, 1988), have effectively resolved these problems by exploiting conditional independence. Therefore, for many problems it is now possible to compute the most probable multicausal explanation using a small number of probabilities.

Although modern Bayesian approaches make it possible to find the most probable explanation for many problems and are therefore used routinely, abduction in general is computationally intractable (Bylander, Allemang, Tanner & Josephson,1991). This means that the time required to find the optimal solution to an abductive problem increases exponentially with the size of the problem (the number of elementary hypotheses and observations). Bylander et al. have shown that abduction is only tractable under very restrictive conditions. An immediate conclusion is that when these conditions cannot be met either the problem must be transformed into a simpler problem or considerable domain-specific knowledge must be used to focus the search for a solution. Hence, in many real-world abductive problems the problem-solver must rely on heuristic or satisficing methods to control the complexity.

Researchers have proposed several different heuristics for determining the best explanation. These are often based on the concept of parsimony—that the best explanation is the simplest. Two of the most prevalent metrics for parsimony are minimum cardinality and irredundant covers. Minimum cardinality states that the best explanation is the one that explains all the data with the fewest number of elementary hypotheses. Although this results in parsimonious explanations, Peng and Reggia (1990) have shown that under most real-world situations “minimum cardinality would risk overlooking the most probable hypothesis and is generally not a reasonable criterion to adopt to limit hypothesis generation” (p. 120). Irredundant covers are complete explanations of the data that do not contain a proper subset of elementary hypotheses that also completely account for the data. Minimal covers are by definition irredundant covers, however there can be many more irredundant covers for a set of data. The set of irredundant covers will always include the most probable complete explanation, however, Bayesian analysis tells us that a complete explanation may not always be the most probable.

The discussion thus far has ignored two fundamental aspects of many abductive tasks. The first deals with the generation of hypotheses for individual observations. In the example above, we assumed that potential hypotheses had been enumerated prior to solving the problem, and that the problem simply involved selecting the best set of those hypotheses. However, in many tasks, hypotheses must be generated from underlying causal knowledge of the domain. For example, in medical diagnosis an essential subtask of problem solving is to generate diagnostic hypotheses given evidence from a patient. Typically, single effects can have many causes (a single cause therefore is sufficient but not necessary) and some effects might only be caused by a conjunction of individual conditions. Because there are often many possible hypotheses for a given observation, such that generating and then representing all of them would quickly overload working memory, only a subset of all possible hypotheses can be considered. If this subset is not sufficient for explaining the data, alternatives must be generated. Consider trying to explain the cause of a headache, a common symptom of many diseases. Generating all possible disorders would quickly overload working memory, hence a physician must limit consideration to a few common disorders. If these disorders fail to provide a best explanation, then the doctor must generate and consider alternatives. Indeed, various researchers have found that the number of diagnostic hypotheses generated is small, ranging from three to six (Feltovich et al., 1984; Groen & Patel, 1988; Krems, 1994). These studies also show that the more experienced a person becomes the smaller the number of diagnostic hypotheses considered and the higher the ability to flexibly modify causal explanations.

Another feature of many abductive tasks is that evidence is often revealed sequentially, unlike the example in Fig. 2 in which all evidence was assumed to be known from the start. In these situations the problem-solver must integrate new data and elementary hypotheses into their current explanation. In some cases, such as story comprehension, the order of evidence is largely outside the problem solver's control, but in many tasks, such as scientific discovery and diagnosis, the problem solver must actively design and conduct experiments to collect additional data.

3. A mental-model based theory of abductive reasoning

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. The theory has six central features. First, only one overall situation model (or current explanation) is kept. It contains both the evidence and hypotheses. This is consistent with Dunbar and Klahr's (1989) model of scientific discovery, and with other studies of human problem solving showing that people use progressive deepening to explore a problem space (Newell & Simon, 1972). However, it is inconsistent with many normative models of abduction, which often keep (or find) multiple overall explanations. The difference between psychological and normative models is a potential source of human error that we can use to test the model, as shown later in the experimental section.

The second feature is that the situation model consists of a conjunction of concrete, abstract, and disjunctive hypotheses. A concrete hypothesis is a single, specific hypothesis. An abstract hypothesis represents a class of related concrete hypotheses. For instance, liver disease is an abstract hypothesis in medical diagnosis, because it encapsulates a number of specific alternative disorders. Disjunctive hypotheses are separate alternative hypotheses for one or more data. For instance, a disjunctive hypothesis for chest pain might include myocardial infarction and indigestion. The use of abstract and disjunctive hypotheses allows a single situation model to compactly represent a range of possible situations. This type of representation is similar to Peng and Reggia's generator set, which compactly represents all possible explanations (Peng & Reggia, 1990). However, unlike Peng & Reggia's model, our model does not keep multiple situation models.

Third, evidence is obtained and processed incrementally. The model sequentially obtains and integrates new evidence into the situation model. In some cases, the evidence is processed without reference to the current explanation. Exceptions are noted below.

Fourth, the current explanation is not used if a simple hypothesis can account for the new evidence and that hypothesis can be added to the current explanation without exceeding some limit on the complexity of the explanation. A simple hypothesis is one that is concrete, not abstract or disjunctive. If only abstract or disjunctive hypotheses are available, the current explanation is used to discriminate among them, as described below. Even if a simple hypothesis is available, it can only be used if it does not exceed some limit on complexity of the overall explanation. This limit is often domain-specific. For instance, in the experimental task used in this paper, the participants are sometimes told that their explanations cannot exceed a specific number of elementary hypotheses.

Fifth, the current explanation is always used when evaluating competing hypotheses. This is done when deciding among alternative hypotheses in either a disjunctive hypothesis or an abstract hypothesis. Since abstract hypotheses and disjunctive hypotheses represent a set of competing hypotheses, the same technique is used in both cases. Competing hypotheses are evaluated by finding the intersection of two sets: (1) the competing hypotheses for the new evidence; and (2) the hypotheses proposed or accepted for previous evidence. Hypotheses that are not in this intersection are eliminated from further consideration. If there is only one hypothesis in the intersection, then it is accepted. If there are more than one, then either an additional criterion is used to select from among the hypotheses or additional evidence is acquired and then used to further constrain the hypothesis set. When the intersection of old and new hypotheses is empty the current explanation is of no use for discriminating among the new hypotheses. We refer to this process as *evidence integration.* The results of the process depend on the method by which the intersection is computed, and this in turn depends on the task, its representation, and how well a person can retain the representations needed to find the intersection. For example, if the hypothesis sets are large, it is unlikely that a person will correctly find the complete intersection; however, if the sets are small or highly structured, then the person might fare better. Later we present evidence suggesting that people adopt heuristics for finding the intersection.

Finally, changes to the situation model, either through evidence collection, addition of new hypotheses, or changes to existing hypotheses do not force a total reevaluation of the situation model. This differs from many other models of abduction, such as ECHO (Thagard, 1989) and Bayesian networks, which update the entire network of evidence and hypotheses anytime new evidence is acquired. Our model uses local consistency checks (at the location

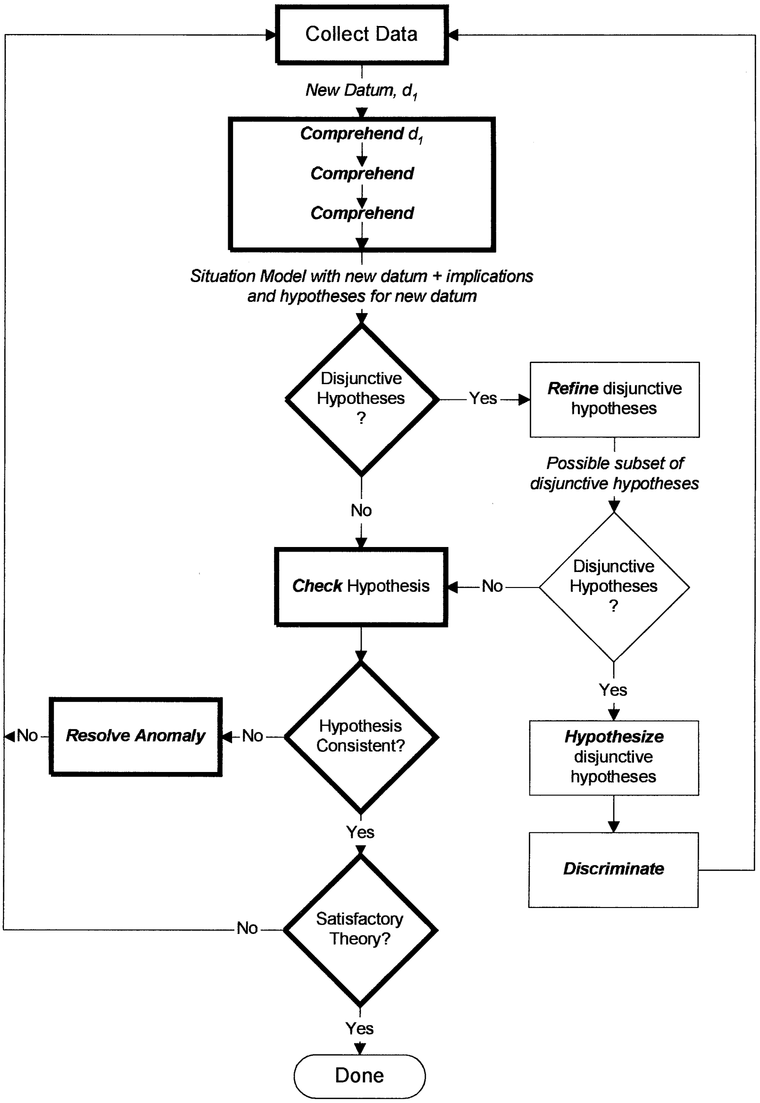


Fig. 3. Flowchart of the model's most common problem solving steps.

of the change), rather than global consistency checks. This strategy predicts specific types of errors that we can use to test the model. For instance, it predicts that a person might not realize that a newly added hypothesis changes the explanatory coverage of a previously accepted hypothesis.

The model's basic problem solving steps are shown in Fig. 3. Suppose that a new observation is available. First, the situation model is updated to include this new information. Next, the new observation is comprehended to determine what it implies about the situation. Comprehension results in one or more hypotheses for the observation. As described above, these hypotheses can be either concrete, abstract, or disjunctive.

If a single concrete hypothesis is generated to explain a new observation, and the hypothesis is consistent with the rest of the situation model, then that explanation is incorporated into the current situation model and problem solving continues. If an abstract or disjunctive hypothesis is generated, the model attempts to refine the hypothesis by using other information in the situation model to discriminate among the alternatives. If this does not result in a single concrete hypothesis, the model then sets itself the task of discriminating among the alternative hypotheses by collecting additional data. Once the explanation is refined to a single concrete hypothesis, it checks the hypothesis to ensure that it is consistent with the rest of the situation model. When a new explanation is inconsistent with the current explanation an anomaly has occurred and the situation model must be modified by either finding an alternative explanation for the new observation or by altering an explanation for the previous observations. Finally, the model determines whether the entire explanation (as expressed in the situation model) is satisfactory. If so, the model stops, otherwise, it will continue to collect data. The metric for a satisfactory theory is, in part, task dependent. For example, in some of the experiments described later, the task specifies that the situation model must contain exactly four elementary hypotheses.

The model was implemented in Soar (Laird, Newell & Rosenbloom, 1987; Newell, 1990) as a set of problem-spaces. The model presented here is based, in part, on the model presented by Johnson and Smith (1991), which was based on a mechanism first proposed by Josephson, Chandrasekaran, Smith, and Tanner (1987). The main space is the abductive problem space with seven operators: *collect-data, comprehend, discriminate, refine, check, test* and *resolve-anomaly.* The states in this space contain the situation model along with other state information needed to solve the problem. The desired state must satisfy general abductive criteria for a best explanation as well as additional domain criteria. The general abductive criteria specify that all the data are explained, that there are no inconsistencies, no redundant explanatory components and that each explanatory component has a high degree of certainty. These criteria are treated as constraints by the model, because it is not always possible (or even desirable) to satisfy all of them.

The operators are defined as follows. *Collect-data* is used to gather data to construct an initial hypothesis or to extend the current situation model. For example *collect-data* would be used at the start of a task before any data has been collected. *Collect-data* is implemented in a subspace that designs and conducts an experiment to generate new data.

Comprehend takes one or more parts of the situation model and determines their implications. Comprehending all of the implications of a single object in a situation model can be a multistep process, requiring multiple comprehend operators. For example, comprehending new data will produce one or more hypotheses to explain that data. Comprehending a hypothesis might in turn result in ruling out other possible hypotheses, or in adding implications of that hypothesis to the situation model.

Refine takes an abstract hypothesis or a disjunctive set of hypotheses and evaluates the alternatives with respect to the situation model. This is the process of evidence integration. It is accomplished using the intersection mechanism described earlier (see the fifth model assumption at the beginning of this section). Specifically, it finds the intersection of two sets: (1) the competing hypotheses for the new evidence; and (2) the hypotheses proposed or accepted for previous evidence. Hypotheses that are not in this intersection are eliminated from further consideration. If there is insufficient evidence to select a single hypothesis, then *discriminate* is used to break the tie by collecting data (i.e., by designing and conducting an experiment).

Discriminate attempts to select one concrete hypothesis from among alternative hypotheses by collecting additional data, generating hypotheses for that data (using comprehend and refine), and then reapplying *Refine* to the alternative hypotheses that need to be discriminated. Details of this process are illustrated in the example given later in this section.

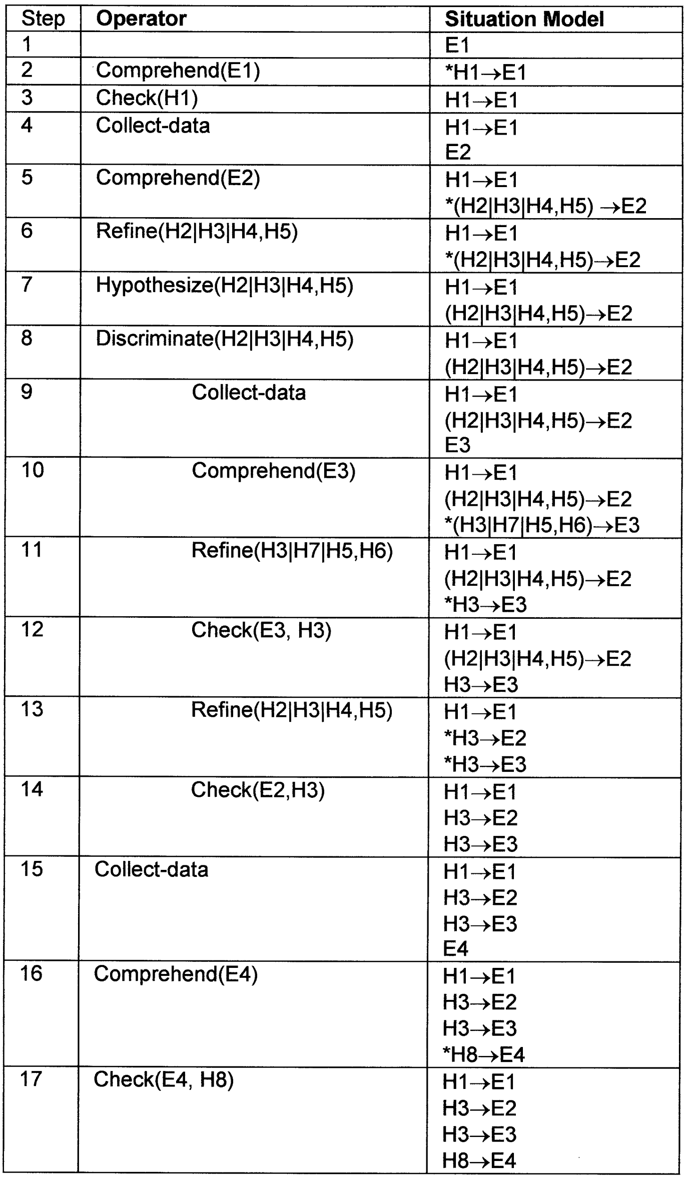
Check takes new results (such as new hypotheses) and determines whether they are consistent with the other parts of the situation model. *Check* annotates the situation model with this information and can also add a certainty annotation to the item being checked. The certainty of an object in a situation model is either *unknown, uncertain, certain*, or *inconsistent.* This is similar to the certainty scheme used by Newell and Simon (1972) to model human problem solving on cryptarithmetic. The use of this scheme in RedSoar (Johnson et al., 1991) and LiverSoar (Bayazitoglu, Smith & Johnson, 1992; Smith, Bayazitoglu, Johnson, Johnson & Amra, 1995), two diagnostic systems, suggests that it is sufficiently powerful for complex problem-solving systems. Klahr and Dunbar (1988) also use a similar scheme to rate hypotheses in their model of scientific discovery.

Test takes some uncertain item (such as a hypothesis) in the situation model and designs and conducts an experiment to either confirm or disconfirm the item

Resolve-anomaly takes anomalous parts of the situation model, such as two contradictory hypotheses, and determines which should be rejected. Once only one item remains, *resolve-anomaly* is finished and *comprehend* must be reapplied to generate alternative explanations. If there is insufficient information to resolve the anomaly, then *discriminate* must be used to make a selection (by collecting additional data).

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 search-control knowledge that prefers one or more operators over others. The search-control knowledge can be sensitive to the contents of the situation model, hence the particular sequence of operators is determined dynamically based on the current situation.

To better understand how the model works, consider the hypotheses and evidence shown in Fig. 2. Fig. 4 shows the steps that the model takes to explain the data. To begin assume that the model has the causal knowledge shown in Fig. 2 and that it is incrementally given data on request. The example begins when the model is given E1 (Step 1). The model first comprehends the new evidence, resulting in the situation model shown in Step 2. The asterisk next to H1 indicates that it has been proposed, but not yet accepted. In Step 3, H1 is checked to ensure it explains E1. Once checked, it is accepted (which is indicated by removing the asterisk). Additional data are then collected (Step 4) and comprehended (Step 5), resulting in new evidence E2 and a disjunctive set of three hypotheses (H2 or H3 or H4, H5), which indicates that E2 can be explained by either H2, H3, or H4 and H5 taken together. In Step 6, the disjunctive hypothesis is refined, but the intersection between it and the other hypotheses in the situation model is empty, so the disjunctive hypothesis is accepted (Step 7) and then *discriminate* is applied to it (Step 8). New evidence, E3, is collected in Step 9, then comprehended (Step 10) to produce a second disjunctive hypothesis (H3 or H7 or H5,H6). Step 11 refines the second disjunctive hypothesis by finding the intersection with the



Step	Operator	Situation Model
1		E1
2	Comprehend(E1)	*H1→E1
3	Check(H1)	H1→E1
4	Collect-data	H1→E1 E2
5	Comprehend(E2)	H1→E1 *(H2|H3|H4,H5) →E2
6	Refine(H2|H3|H4,H5)	H1→E1 *(H2|H3|H4,H5)→E2
7	Hypothesize(H2|H3|H4,H5)	H1→E1 (H2|H3|H4,H5)→E2
8	Discriminate(H2|H3|H4,H5)	H1→E1 (H2|H3|H4,H5)→E2
9	Collect-data	H1→E1 (H2|H3|H4,H5)→E2 E3
10	Comprehend(E3)	H1→E1 (H2|H3|H4,H5)→E2 *(H3|H7|H5,H6)→E3
11	Refine(H3|H7|H5,H6)	H1→E1 (H2|H3|H4,H5)→E2 *H3→E3
12	Check(E3, H3)	H1→E1 (H2|H3|H4,H5)→E2 H3→E3
13	Refine(H2|H3|H4,H5)	H1→E1 *H3→E2 *H3→E3
14	Check(E2,H3)	H1→E1 H3→E2 H3→E3
15	Collect-data	H1→E1 H3→E2 H3→E3 E4
16	Comprehend(E4)	H1→E1 H3→E2 H3→E3 *H8→E4
17	Check(E4, H8)	H1→E1 H3→E2 H3→E3 H8→E4

Fig. 4. A trace of the model when solving an abductive problem based on the causal knowledge shown in Fig. 2.

rest of the hypotheses in the situation model. This leaves H3 as the only candidate. Since the other hypotheses have been eliminated for E3, H3 is checked (Step 12) and accepted. Since the model is still in the process of trying to discriminate among the alternative hypotheses for E2 (from Step 8), it once again attempts to refine the disjunctive hypothesis for E2 (Step 13). H3 now intersects with the disjunctive hypothesis, so it is selected and then checked (Step 14). Steps 15 through 16 show the collection of the final evidence and how it is explained.

This example illustrates the important theoretical points of the model. First, it shows that only one overall explanation is kept. Second, it shows that some disjunctive hypotheses are kept, but these are internal to the overall explanation, and represent alternative hypotheses for a single evidence item (see the disjunctive sets in Steps 5 through 12). Third, the model processes evidence incrementally, independent of the existing explanation, which is used only during the refine stage (see Steps 6, 11, and 13).

The example shows how the model strives to keep the situation model as simple as possible by keeping track of as few disjunctive hypotheses as possible, and quickly simplifying the model whenever possible. This is consistent with the coherence approach to belief revision which emphasizes consistency and conservatism and claims that belief revision should keep the resulting belief system consistent and should only make a minimal change to the current belief system (Gardenfors, 1992).

This example also suggests several sources of error. First, the model assumes that explanatory coverage increases monotonically with the addition of new hypotheses. This is clearly not always true, because adding an additional hypothesis could alter what an existing hypothesis explains. Second, the model is biased to producing locally parsimonious explanations, but these explanations might not be globally parsimonious. The explanations are locally parsimonious because the intersection technique in the Refine stage finds the most parsimonious explanation for the evidence being processed, but can ignore much existing evidence and is not automatically updated in light of new evidence. In addition, focusing on parsimony at the local level can blind the model to situations in which the correct explanation is not locally parsimonious. In fact, we observed all of these types of errors in human reasoning, as we describe in the following section.

Due to the complexity and number of features of the model, it is difficult to test the model as a whole. Instead, it is necessary to analyze specific features that are worthy of testing and test each individually. The focus of this paper is on the use of the current explanation in evidence interpretation. As such, it examines the last three features of the model: 1) how the nature of the evidence determines whether it is processed with or without reference to the current explanation; 2) that the current explanation is always used to discriminate among alternative hypotheses; and 3) that changes to the situation model do not force a total reevaluation of the explanation. The first two features are unique claims of the model, whereas the last feature is important for capturing the satisficing nature of human reasoning.

4. Applying the model to an abductive task

4.1. The task: black box

To investigate our hypotheses and model we use a task in which the function and structure of a device is known and the goal is to determine the hidden state of the device given indirect evidence of that state. In order to see how people deal with complexity and how they use the current explanation to generate and integrate hypotheses we developed an experimental paradigm that allows individuals to first learn and then apply a device mechanism for explaining observations by a combination of causes. The paradigm also allows us to verify

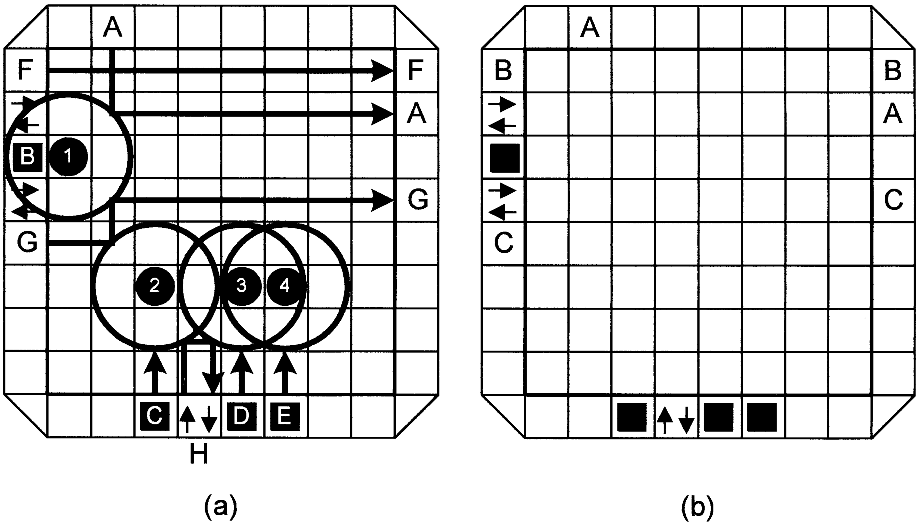


Fig. 5. (a) The Black Box with four atoms and the paths of several light rays visible. (b) The Black Box as it would appear to an individual trying to locate the atoms: neither the hidden atoms, nor the ray paths are visible.

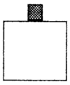
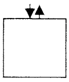
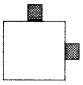
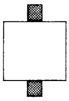
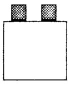
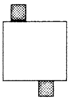
the complexity of the observations as well as the complexity of the causal patterns. The task is easy to understand but still admits a large number of possible hypotheses, which makes it a difficult task. In the task, called Black Box,[2] participants must locate *n* (normally 4 or 5) atoms hidden in a box by shooting light rays into the box and observing where the rays exit the box.

The Black Box consists of an 8 × 8 matrix in which the atoms are hidden, and 32 perimeter squares from which light rays can be shot into the box (see Fig. 5a). Each atom (labeled 1–4) has a field of influence (shown in the figure as a larger circle around the atom). These fields deflect or absorb light rays according to two primary laws. First, if a ray hits a field of influence head-on, it is absorbed (Rays B, C, D and E). If a ray is absorbed, the ray's input cell is marked with a black square. Second, if a ray hits a field of influence at an angle it is reflected 90 degrees away from the atom (Rays A and G). 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). A reflection occurs whenever an atom is located on the edge of the box to either side of the input square (Rays I and J), or when a ray's direction of travel is reversed because it approaches the space between two atoms that are one cell apart (Ray H). If a ray enters and exits different locations, those locations are marked with matching symbols (Rays A, F and G, marked with letters).

In the Black Box task, the atoms are hidden in the box and the participant's goal is to discover their location using the fewest number of light rays. The participant sees only the entrance and exit of the ray, not the path that the ray follows; hence, the path must be inferred (see Fig. 5b). To record his or her hypotheses regarding atom locations, the participant can place atom markers on the grid. When the participant believes that all atoms have been correctly marked he or she ends the game and the hidden atoms are revealed.

We selected Black Box for four primary reasons: First, it shares several features with

Table 1
Frequency of ray pattern classes and explanations

	Ray pattern classes					
	Absorption	Reflection	L	Straight	U	Zig-Zag
	59.77%	18.68%	11.82%	7.62%	1.32%	0.79%
Number of Atoms Causing Pattern						
0	N/A	N/A	N/A	99.62%	N/A	N/A
1	86.90%	89.91%	97.48%	N/A	N/A	N/A
2	11.16%	8.70%	N/A	N/A	99.06%	98.59%
3	1.75%	1.26%	2.46%	0.36%	N/A	N/A
4	0.19%	0.13%	0.06%	0.02%	0.94%	1.41%

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) Data collection can be given a cost, thereby encouraging participants to minimize data collection; 3) A single hypothesis can explain more than one data item; 4) A single datum can require multiple individual hypotheses to explain; and 5) Abstract hypotheses can be formed during problem solving.

Second, Black Box is easy to understand. All participants in the experimental studies were able to learn the rules in under an hour. Third, one of the major problems with the study of abductive reasoning (such as studies done in medical domains or natural scientific domains) is the difficulty in controlling for knowledge differences among individuals. By using a simple domain like Black Box we can ensure that all participants have the same domain knowledge and that no additional external knowledge is given to the individuals.

To assist in the analysis of Black Box problem solving, we distinguish between light ray input-output patterns, light ray paths, and the atoms that produce the paths. When a light ray is shot into the Black Box it produces a particular input-output pattern. These patterns are the data that must be explained by the participant. We classify these patterns as shown at the top of Table 1. Each class corresponds to a unique input-output relation without respect to how the ray actually travels inside the box. The patterns called L, Straight, U and Zig-Zig get their names from the shape of the most frequent ray path that explains the pattern. The percentage beneath each ray pattern indicates the relative frequency of occurrence of the pattern for all possible 4-atom configurations.

The internal path of the ray is classified by the number of atoms that support the path. The remainder of the table lists the relative frequency of occurrence for each path class. N/A indicates that there is no possible path to produce the ray pattern with the given number of atoms. For example, 59.77% of all ray shots (over all 4-atom configurations) are absorptions. Of these, 86.9% are caused by a single atom, 11.16% by two atoms (meaning that the ray is reflected by one atom and then absorbed by a second atom), 1.75% by three atoms, and 0.19% by four atoms.

Note that only the L and Straight ray patterns can be readily explained by simple concrete hypotheses—the remaining ray types require abstract or disjunctive hypotheses. For instance, an L pattern precisely specifies a specific path, supported by a single specific atom. Likewise, a Straight can be explained by a specific path that goes straight through the box, meaning that there are no atoms along that path. In contrast, the simplest hypothesis for a Zig-Zag pattern is any two atoms arranged in the configuration needed to produce the pattern, hence additional evidence is needed to further constrain the hypothesis. Since the proposed cognitive model assumes that the current explanation is not used if there is a simple hypothesis to explain new evidence, L and Straight rays should never make use of the current explanation, whereas the remaining ray types should always make use of the current explanation.

4.2. Case study

To apply the general model to Black Box, the domain-independent model outlined above must be supplemented by domain knowledge. To do this, we conducted a case study using Black Box (see Johnson, Krems & Amra, 1994, for further details).

In the case study, five Ohio State University students solved a series of 170 randomly constructed cases in up to five two-hour sessions. One had significant prior experience with Black Box and is referred to as the "expert" or E1. The others, called "novices," had no previous experience with Black Box. The participants were told to speak aloud while solving the tasks. All sessions were videotaped for later analysis.

This study revealed several regularities. First, novices immediately generated abstract hypotheses, confirming our assumption that such hypotheses play a role in Black Box. For example, individuals immediately produced hypotheses like "There is an atom somewhere in column 3."

Second, the case study revealed that the most commonly generated explanation for each ray type (when given no current explanation) was typically the one involving the fewest number of atoms (see Table 1). However, people regularly generated three alternative explanations for a reflection: a single atom on either side of the reflection, or a pair of atoms located along the row or column of the ray. These are illustrated in Fig. 6. When a participant sees the reflection (Ray A) he or she produces three explanations as illustrated by statements like "This means that there is either an atom here [1] or here [2], or a pair of atoms somewhere along the column [3–9]." This seems to violate the principle of parsimony, because participants simultaneously entertain both a one- and two-atom explanation. However, as we discuss later in this section, people are more concerned about the parsimony of the path a ray travels, rather than the number of atoms needed to produce the path.

Third, participants did not always explicitly check to see if newly gathered data were already explained by their current explanation. Instead, they often generated a hypothesis to account for the data and then checked to see if it was already part of the current explanation. In other words, they decided where atoms should be located to explain a ray, and then checked to see if the atoms were already there. This led to errors when the ray was already explained by a hypothesis that differed from the one generated. One such error is shown in Fig. 7. The white atom was placed to explain Ray D (Fig. 7a), even though the three previously placed atoms already accounted for Ray D (Fig. 7b).

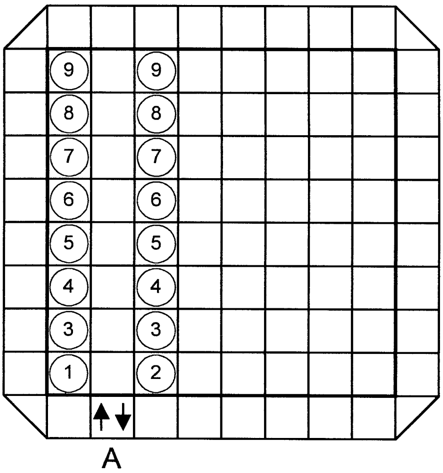


Fig. 6. Three alternative explanations for a reflection. The atom pairs labeled 3 to 9 are part of a single abstract explanation.

Fourth, people clearly reasoned from effect to cause in order to hypothesize atom locations, but this process involved multiple levels of abstraction as illustrated by the following protocol fragment from a novice participant on her third trial. The situation is shown in Fig. 8 by Ray A. Recall that the participant sees the input and output ray markers only, not the actual path of the ray.

[Participant shoots the ray in at A1 and it exits the box at A2.]

"This was bounced here" [Traces the path from A1 to A2 using the mouse]

"So there must be a field of influence here" [Points to cell C]

"which means there must be an atom here." [Places atom 1 to explain L]

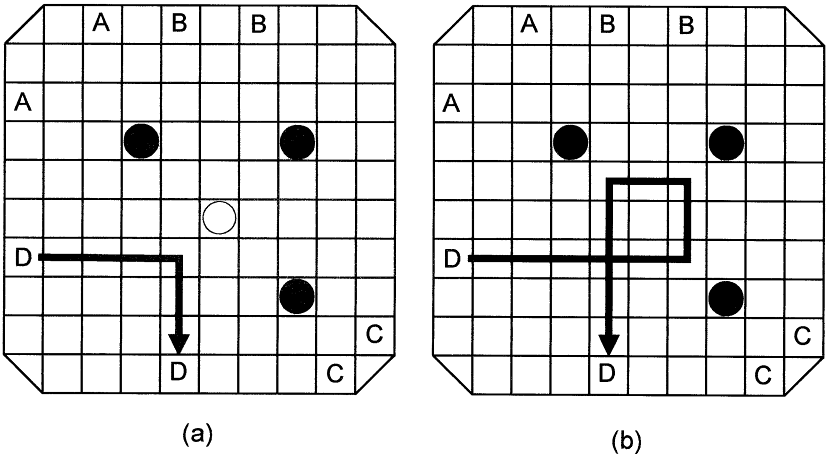


Fig. 7. Ignoring context in explanation generation. (a) Ray D is explained using a new (white) atom. (b) An alternative explanation for Ray D that uses the existing three atoms.

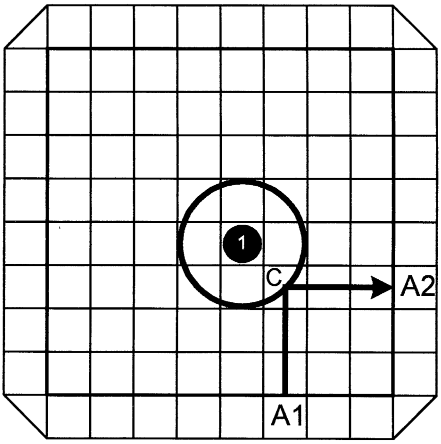


Fig. 8. To explain Ray A (which enters at A1 and exits at A2) participants first hypothesize the ray path shown above, then hypothesize the atom (labeled 1) needed to support this path.

This participant has reasoned backward over the causal knowledge: from the ray result, to a path that the ray could have traveled, to the atom required to support that path. Note that the path shown is not the only way to explain the ray shot, but the participant did not mention any alternatives.

Fifth, the use of the current explanation to generate hypotheses varied with the type of the ray being explained. For some types of rays the current explanation was always used to constrain hypothesis generation, but for other rays it was never used. We have already discussed the example shown in Fig. 7, where the current explanation was completely ignored. When participants saw a straight ray they usually assumed that any atoms lying along or to the side of its path indicated an anomaly. However, when individuals saw a zig-zag, they would use existing atoms to constrain their explanation of it. Absorbed rays provide another interesting example. Given no current explanation, people explained an absorption by hypothesizing a single atom somewhere in the line or column into which the ray was shot (see Ray A in Fig. 9). However, when there was an atom located to either side of the column or row (Ray B in Fig. 9), participants explained the absorption by proposing that there was an atom somewhere along the path that a ray would travel if it were reflected by the existing atoms.[3]

Finally, the case study suggests that evidence integration is a rapid, possibly automatic process. Fig. 10 illustrates the use of evidence integration based on a reflection. The participant first sees the reflection shown in Fig. 10a whereupon he or she considers three explanations labeled 1, 2 and 3–9. As discussed above, a typical participant statement at this point is “This means that there is either an atom here [1] or here [2], or a pair of atoms somewhere along the column [3–9].” Since these are competing hypotheses, the participant must use additional information to determine which to accept. This was usually done by shooting a ray one or two spaces over from the first reflection as shown by the second ray in Fig. 10b. Upon seeing the second reflection (Ray B), participants immediately placed Atom 2. A number of other explanations are possible (e.g., Atoms 1, 2 and 10 could all be

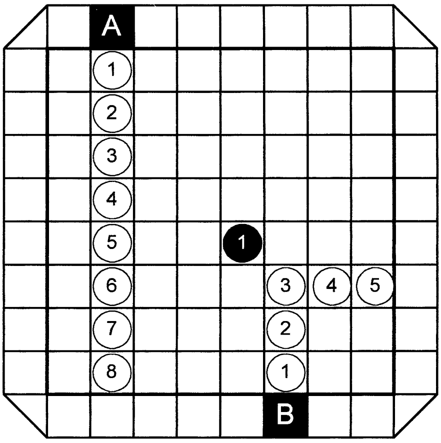


Fig. 9. Participants' explanation of Ray A ignored the previously placed atom (A might get reflected into the atom where it would be absorbed), whereas explanations of Ray B used it.

present), but the individuals did not mention them. Note that evidence integration naturally involves the current explanation: participants combine new information with an existing hypothesis so as to refine the hypothesis space.

In summary, the case study showed that participants sometimes used their current explanation and sometimes ignored it. Upon collecting new evidence, participants did not always check to see if that evidence was explained by their existing explanation. When participants generated hypotheses they appeared to systematically either use or ignore the current explanation. Specifically, Straight and L rays were explained without reference to the current

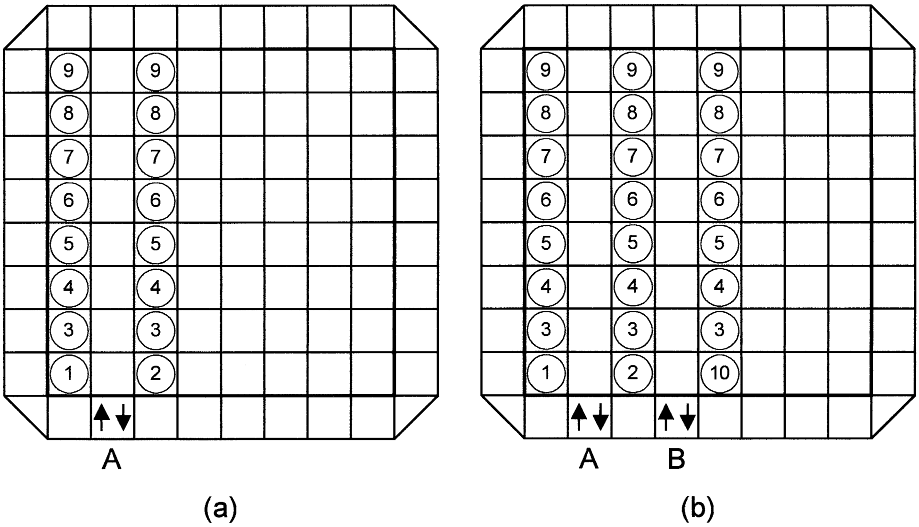


Fig. 10. Evidence integration. (a) Hypotheses produced to explain a single reflection. (b) Hypotheses produced to explain two reflections, shown prior to integration. Participants commonly select Atom 2 after integration.

explanation, whereas, other ray types made use of it. This is explored in detail below in Experiment 1. Finally, participants used their current explanation when interpreting new data that were collected to discriminate between competing hypotheses. This is explored in Experiment 2. The next section shows how the general abductive model can be applied to Black Box to account for these various uses of the current explanation.

4.3. Applying the general model to black box

As described in the previous section, our theory is formulated as a set of problem-spaces implemented in Soar. The general assumptions of the model plus domain-knowledge from Black Box were integrated to build a computational model. Implementation in Soar took place after the case study and was used to operationalize and iteratively test and revise the model, which is presented in its final form in the previous section. Running the model on Black Box cases produced specific hypotheses that we test in the experimental section.

4.3.1. Current explanation use when interpreting new data

As illustrated in the case study, participants did not always check to see whether new data were explained by their existing explanation (see Fig. 7). The general model accounts for this because it assumes that the current explanation is used when processing a new datum only when the generated hypothesis is disjunctive or abstract. In cases such as that shown in Fig. 7, this condition never arises. Specifically, each of the ray shots in Fig. 7 admit a single concrete ray path linking the input cell to the output cell.

4.3.2. Hypothesis generation and use of the current explanation in black box

In Black Box, people generate hypotheses by reasoning from a ray input-output pattern, to a path, and then to atoms. However, the case study revealed that this process is complicated by the varying effect of the current explanation on the generation process (see Fig. 7 and Fig. 9). Fig. 11 and Fig. 12 shows three more examples. Given the situation in which Atoms 1 and 2 have already been placed (Fig. 11a), participants usually explained Ray A by a straight path that ignores Atom 2 (resulting in an anomaly), however, the explanations for Rays B and C were constrained by the existing atoms, as shown in Fig. 12.

These data are easily accounted for by the hypothesis generation model, which is split into two stages: a comprehension stage followed by a refinement stage (see Fig. 4). The end result of the comprehension stage is a context-independent and possibly abstract hypothesis for the data being comprehended. If the comprehension stage results in an abstract or disjunctive hypothesis, then the refinement stage is used to refine the abstract hypothesis. This stage is a direct application of the *Refine* operator of the main abductive space. This operator refines the abstract hypothesis by taking the current explanation (previously hypothesized atoms) into account. Fig. 11b shows the results of comprehension for each of the rays shown in Fig. 11a. Fig. 12 shows the results after the hypotheses have been refined.

The path comprehension stage in Black Box generates abstract hypotheses using a problem space in which the states are the location of the ray and the operators specify movement of the ray through the box at an appropriate level of abstraction. For example, when the problem space searches for a path for a reflection, it produces a ray path that

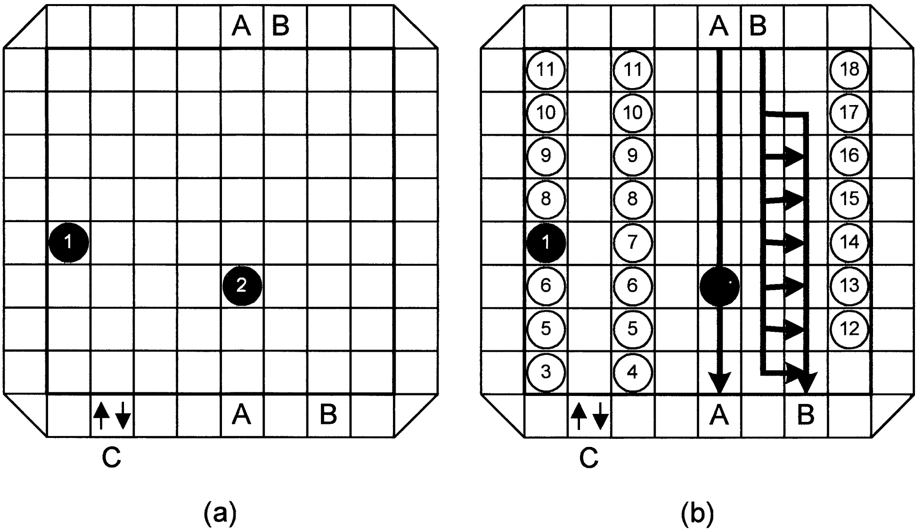


Fig. 11. Comprehension of evidence given previously hypothesized atoms. (a) Atoms 1 and 2 were placed to explain previously collected data (not shown in the figure). Rays A, B, and C were shot after the placement of Atoms 1 and 2. (b) The explanations for each ray (A, B, and C), as produced by the comprehension stage. Hypothesized atoms are white with black numbers. For clarity, not all of the hypothesized atoms are shown for Ray B. The hypothesis for Ray A is that it passed straight through the box, which implies that there are no atoms along its path or to either side of its path. The hypothesis for Ray B implies that it turned right, then turned left.

indicates that the ray went into the box, turned 180 degrees at some unspecified point, then traveled back out of the box. This corresponds to three operators: move-forward (to an unspecified point), rotate-180, move-forward to exit. When explaining an L, such as that in Fig. 8, the problem space uses three operators: move-forward to row 7, turn-right, move-

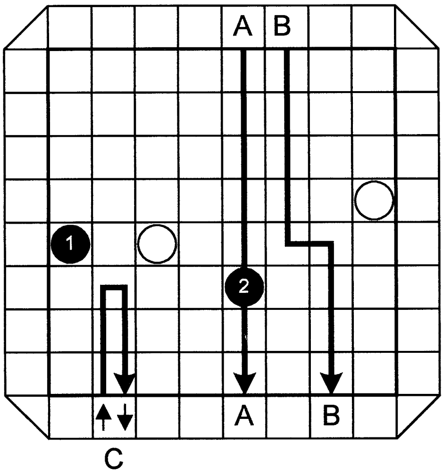


Fig. 12. Ray explanations after the refinement stage (see Fig. 11b for the input to this stage). White atoms are possible explanations that have not yet been marked on the black box.

forward to exit. A zig-zag, such as Ray B in Fig. 11, results in a more complex sequence of operators: move-forward (to an unspecified point), turn-left, move-forward to column 8, turn-right, move-forward to exit. This problem space makes use of a distance reduction metric to constrain the search for paths when given rays with specific input and output cells. It does this by preferring operators that take the most direct path from the input to the output cell.

After a path is produced to explain a ray, the path must be comprehended to determine the location of atoms. This process makes use of traditional backward reasoning over causal knowledge. The space traces the path and for each change of direction or absorption in the path, hypothesizes one or more atoms to support the change. An abstraction in the specification of a ray path (such as the unspecified locations in the paths for reflections and zig-zags), leads to abstract hypotheses for the locations of atoms to support the paths.

Therefore, the result of the second comprehension operator (the atoms to support the path), often consists of several alternative explanations, some concrete and some abstract. For instance, Fig. 11b shows three ray shots and the explanations that the comprehension stage produces for a straight (Ray A), a zig-zag (Ray B) and a reflection (Ray C). In all cases, the existing atoms are ignored. As shown, there are nine possible explanations for Ray C (labeled 3–11), and seven possible explanations for Ray B (labeled 12–18). Fig. 12 illustrates the results after the refinement stage for Rays B and C. Both explanations have been refined (i.e., specialized) using the current explanation. This two-stage model—comprehension followed by refinement—explains why participants appear to ignore their current explanation when explaining the straight ray but use it when explaining the zig-zag and reflection.

The use of abstract paths and the separation of path generation from atom generation also explain why case study participants generate both one and two-atom hypotheses for reflections (see Fig. 10a). For most other rays, the simplest hypothesis is produced and considered before considering the next more complex hypothesis (see Table 1). However, in the hypothesis generation model described above, the abstract path for a reflection specifies that the ray enters the box and then turns around and exits (i.e., it does not specify where the ray turns around). When the atoms are generated to support this path, backward reasoning reveals two ways to support the path: either an atom to one side or the other or two atoms somewhere out in the box.

Another problem is to explain participants' use of the current explanation when explaining absorptions, as discussed earlier and illustrated in Fig. 9. To review, Ray B is accounted for by hypothesizing an atom somewhere along the path shown in the figure—a hypothesis that makes use of the current explanation. We hypothesize that the current explanation is used because absorptions specify an input cell, but no output cell, which means that a simple hypothesis to account for the data cannot be established. As a result, participants apparently work forward from the input cell, using the current explanation to constrain their search for a possible path. This leads to the behavior illustrated in the figure.

4.3.3. *Evidence integration and use of the current explanation in black box*

The evidence integration problem is to explain how data, such as those shown in Fig. 10b are integrated to select from among the hypotheses generated to explain the evidence. In contrast to models that use properties like the number of hypotheses or other domain-independent mechanisms our theory assumes that people explicitly represent different hypotheses in a model. These are then used as context when explaining new data. For example, after seeing Ray A in Fig. 10a and realizing that there is insufficient evidence to select from among the alternative hypotheses for the new datum, our model-based theory assumes that the person explicitly represents all of the alternatives in their situation model. The new situation model contains all of the information in Fig. 10a. Upon seeing Ray B, shown in Fig. 10b, comprehension of the ray occurs as usual, returning nine possible hypotheses for Ray B. In the figure, these are numbered from 2 to 10 for notational convenience only–at this point the system doesn't know that the center column of atoms overlap with those hypothesized for Ray A. When the hypotheses are refined, however, the theory assumes that the person notices that Atoms 2 and 10 are not present, but that Atom 2 was hypothesized as a possible explanation of the previous ray. Since Atom 2 is the only previously generated explanation that completely overlaps with the new explanations, Atom 2 is placed.

What would happen if Atom 10 had already been placed prior to seeing the second ray? The theory assumes that existing atoms take precedence over hypothesized atoms, thus Atom 10 would be noted as the explanation for the second ray and further data would have to be collected to discriminate among the explanations for the first ray.

The model theory offers an explanation of evidence integration that does not depend on counting elements of the representation or computing the likelihood of each hypothesis based on the available evidence. Instead, it relies on local processing of data using model information derived from a sequence of similar local processing steps. If some or all of the previously derived information cannot be represented in the model (due to memory limitations, or any other reasons), the evidence integration process will be affected.

5. Experimental model evaluation

The major goal of the experiments was to empirically investigate some major assumptions and predictions of the model with respect to the use of the current explanation in evidence interpretation. The theory assumes that the current explanation is not used to explain new evidence if there is a simple hypothesis to account for that evidence. In contrast, it assumes that the current explanation is used when attempting to discriminate among several alternative hypotheses for new evidence. This occurs when refining abstract hypotheses to produce a specific hypothesis or when discriminating among competing hypotheses in a disjunctive hypothesis. In all other cases new hypotheses are generated without referring to the current explanation. Experiment 1 investigates the assumption that the current explanation is not used when there is a simple hypothesis to account for the new evidence. Experiment 2 investigates the assumption that the current explanation is used when refining an abstract hypothesis. Experiment 3 examines whether changes to the current explanation affects the interpretation of new data. It tests the model's prediction of current explanation use, as well as the assumption that the situation model is not reevaluated as it is updated.

5.1. Experiment 1: current explanation use in hypothesis generation

The goal of the first experiment was to investigate the assumption that the current explanation is not used when a simple (i.e., nondisjunctive, concrete) hypothesis is available to explain the new data. The experiment tests this assumption by giving participants new evidence (a single ray) that is already accounted for through a path that makes use of previously placed atoms (the current explanation) but which the participants can also account for by adding a new atom. An example of such a situation is shown in Fig. 7, wherein Ray D is already explained by the three black atoms (Fig. 7b), but could also be explained by hypothesizing the single white atom (Fig. 7a). If participants use their current explanation to interpret the new evidence, they should notice that the evidence is already explained, but if the interpretation is done without reference to the current explanation, they will place a new atom.

5.1.1. Method

The experimental hypothesis, that the current explanation is not used if a simple hypothesis can be established, was examined in a repeated-measures design where participants solved a number of Black Box tasks that supported two alternative hypotheses for the final ray: one that used the current explanation and one that ignored it.

5.1.1.1. Participants. 15 participants between 20 and 25 years old ($M = 22.3$ years, $SD = 1.99$) participated in the experiment. There were 11 women and 4 men. The participants responded to announcements posted at the campus of the University of Regensburg and received course credit for participating. None had significant prior experience with the task (Black Box).

5.1.1.2. Apparatus. Personal Computers were used for stimulus presentation and response collection. The task was presented on a 17 in. color monitor. Participants were instructed to use the mouse to either place or remove an atom or to request more data. The program logged and time-stamped all requests for data and atom placements and removals.

5.1.1.3. Stimuli. The experiment consisted of four phases: 1) A training phase in which the rules of ray travel were explained to participants and their knowledge of these rules was tested. Ten configurations were constructed for use in this phase of the experiment. 2) An acquisition phase in which participants solved a series of Black Box trials. Six three- to five-ball configurations were randomly generated for use during this phase of the study. 3) A test phase for testing the predictions. For this phase 6 test tasks and 12 distracter tasks were constructed. Every test task supported both a context-dependent and context-independent explanation. A context-dependent explanation is one in which previously placed atoms could be used to explain the new datum (see Fig. 7b). The six test tasks were constructed by varying two dimensions: rotation (0^o 90o, 180o), and the distance between the previously placed atoms and the target datum (near or far). Distance refers to whether the existing atoms were close to the target datum or far away. For example, the three black atoms in Fig. 7a are close to Ray D. These three atoms could be moved further from Ray D by shifting the upper two atoms all the way to the top of edge of the box and shifting the rightmost two atoms all the way to the right edge of the box. This more "distant" configuration still produces the same input-output pattern for Ray D using a path similar to, but longer than that shown in Fig. 7b.

Based on the model predictions, the six tasks used in this experiment are isomorphic, because distance and rotation should not affect the answer. Each of the test tasks consisted of eight rays explainable by three or four atoms. Stimuli included a preset sequence of four ray shots before the critical decision had to be made after the fifth ray shot. The first four ray-patterns were constructed so as to lead the participants to construct an experimenter-designed explanation. Participants could easily explain the four ray patterns by placing two to three atoms. These atoms served as context for the interpretation of the critical fifth ray shot. The remaining three rays were consistent with both the context-dependent and context-independent solutions. The 12 distracter tasks were regular Black Box tasks without the patterns used for the test cases. Two sequences of 18 tasks each were built with the constraint that two distracter tasks were located between test cases. Between the two sequences, the rotation and distance factors were counterbalanced. 4) The final phase was a post-test in which participants selected the most plausible explanation from two competing explanations: one context-independent and one context-dependent.

5.1.1.4. Procedure. Participants were tested individually in an experimental session that lasted from 38 to 123 min (M = 85.3 min, including training, SD = 14.3 min). The experiment was split into training, acquisition, test, and post-test phases. Since our goal is to model abductive skill for tasks in which people are familiar with the function and structure of a device, it was important to ensure that the individuals completely understood the Black Box rules of ray travel before beginning to play the game.

(1) Training phase. The novices first read instructions describing how the atoms affect the light rays. Then they were given a series of three-ball to five-ball configurations, where the atoms were visible and the task was to predict the path of a ray, given an input cell. To proceed to the acquisition phase a participant had to correctly predict the paths for all of the rays for two consecutive configurations. This ensured that every participant understood the causal mechanisms underlying Black Box.

(2) Acquisition phase. At the beginning of the acquisition phase, the experimenter explained the Black Box abduction task—to locate atoms by shooting rays into the box. The participants were told that they should apply the previously acquired causal knowledge in order to locate atoms. They also were told that they would be doing this task for the remaining sessions, and that the purpose of the experiment was to see how their performance improved. Each participant then solved six practice tasks. During the acquisition phase the experimenter was permitted to answer questions concerning the rules of the task.

(3) Test phase. Next the participants solved the 6 test and 12 distracter tasks in a fixed order. The participants' task in the acquisition and test phases was to develop an explanation of rays by placing atoms. During a trial, they could place an atom marker, remove it, or ask for new data by clicking on a button labeled "More Data." Clicking on this button highlighted one of the perimeter cells of the Black Box matrix. This showed the participant where the ray would be shot into the box. The participant then clicked on the highlighted cell to fire the ray, which then revealed the result. Thus, the data were presented sequentially as it would be if

Table 2
Proportion of participants who used the current explanation

Distance	Isomorph		
	0	90	180
Near	4/11, p $<$ .001*	4/10, p $<$ .001*	5/10, p $<$ .01*
Far	10/3, p $<$ .13	14/1, p $<$ .55	9/5, p $<$ .009*

the participants were actually shooting the rays themselves. This procedure allowed the experimenters to control the information the participants saw and therefore the knowledge people could acquire by solving the tasks.

(4) Post-test. After participants had solved the last trial in the series, six pairs of the six test tasks were sequentially shown to them as figures on paper (like Fig. 10) and they had to decide which alternative they considered more plausible.

5.1.2. Results

Data were the placements of atoms by the participants during a task. There were 90 lists of atom placements (15 participants on six test-tasks). The average time to solve the test tasks was 82.3 sec (SD = 50.2). ANOVA showed that there was an improvement between tasks in terms of speed, $F(5,24) = 2.82$, $p < .02$. Student-Newman-Keuls posthoc tests revealed a significant difference between the first and the last trial. In a two-factorial ANOVA with distance and rotation as factors no significant main effects, $F_{distance}$ (1, 84) = 0.52, $p <$ .47; $F_{rotation}$ (2, 84) = 0.48, $p <$ .67, nor a significant interaction, F (2, 84) = 0.11, $p <$ .89, could be found in processing time. The same holds for the time participants spent for the "critical" datum where they had a choice: neither the main, F (1, 84) = 0.16, $p <$ .69, F (2, 84) = 1.65, $p <$ .2, nor the interaction term was significant, F (2, 84) = 0.11, $p <$ .9. The critical decision between a context-dependent or context-independent hypothesis required that participants establish the right context–in terms of placed atoms–in advance (while trying to explain the first four shots). This was the case in 95.6% (86 out of 90) of all solutions. The low error rate also means that the participants understood the rules of the game. The four incorrect solutions were not used in further statistical analysis.

Frequency tables were built by counting the number of participants who used the current explanation to account for new observations (see Table 2). Our first experimental hypothesis states that the existing explanation is never used for explaining the new observation if a weak-method can be used. We found that the current explanation was ignored in only 53.5% of all trials. In all the other trials people did not place an additional atom. For statistical analysis it was assumed that according to our theory participants should not use the current explanation in more than 90% of all tasks. A binomial test was calculated with H_o: $p \geq .9$. Since the direction of the difference was predicted in advance, a one-tailed rejection region is given. With $u = 11.29$, $p < .0001$ the first experimental hypotheses has to be rejected. Contrary to the model assumptions people often use the current explanation in generating explanations even when a simple hypothesis is available.

In order to investigate the association between distance and rotation, a χ^2 test was calculated. No significant interaction between the two factors could be found, χ^2 (2, N =

15) = 2.2, $p < .34$). In a separate analysis we investigated whether current explanation use depends on distance or rotation. The factor distance reached statistical significance, χ^2 (1, N = 15) = 20.8, $p < .0001$, but not rotation, χ^2 (2, N = 15) = 1.3, $p < .52$. This means that participants were more likely to place an atom in the far distance tasks.

In an analysis of the cells of Table 2 six binomial tests were calculated with H_o: $p \geq .9$. The total number of observations per cell differs because of missing values in the data. Table 2 shows that under two far-distance conditions (where rotation was 0- and 90-degrees) participants placed additional atoms, that is they explained new observations independent of the current explanation. In all other conditions, the data significantly deviate from the model predictions. Hypothesis 1 therefore can only be kept with regard to a very specific situation.

In their final decision during the post-task test only 9 out of 15 participants preferred a context-independent solution for tasks with a far distance (meaning that they selected the explanation containing the additional, redundant atom). For tasks with a near distance, 8 out of 15 participants preferred a context-independent solution. Again binomial tests showed significant effects, based on H_o: $p = .9$, $p < .002$ and $p < .001$. Thus in a situation where both solutions are presented to the participants, they do not select a simple, context-independent solution.

5.1.3. Discussion

Processing time did not differ significantly between the six test tasks. This means that the results cannot be explained by assuming different speed-accuracy trade-offs among the individuals. Nor is it likely that effects were due to differences in the difficulty of the tasks. The postexperimental questioning of the participants made it clear that people had enough domain-specific knowledge to be aware of both alternatives. Contrary to our first experimental hypothesis participants did not consistently explain new evidence independent of the existing explanation. They sometimes used the existing causal explanation—the pattern of atoms—when they interpreted new evidence. Only when the existing atoms lay far from the ray path did they place another atom. One explanation for this deviation of participant behavior from our model predictions might start with the visual search individuals must do to process the display. In Black Box a hypothesis is built using a search process that generates a path starting at the input-cell and ending at the output-cell. Since this search process depends on the visual display and its representation, this path is identified by scanning the display from the input- to the output-cell. Therefore, context is not just defined by task features, like previously placed atoms, but also depends on the focus of attention and the part of the display a person is looking at. A low context situation is when the existing atoms are far from the visual search path. On the other hand, if the previously placed atoms are close to the visual search path, a strong context situation is given.

This experiment suggests that people sometimes check to see if new data are already explained by their existing hypothesis. Only when the focus of attention is not large enough to "see" the relevant context does a person immediately explain the new data, independent of the current explanation. If previously placed atoms interfere with their display-based reasoning, individuals use the current explanation to constrain their search. Thus, local simplicity at the path level can be over-ruled if visual attention indicates that new data can be explained by the existing hypothesis.

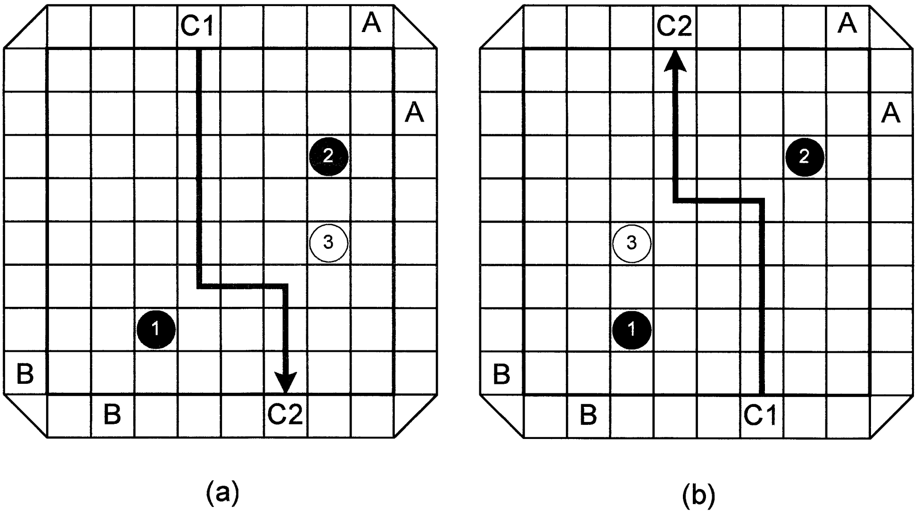


Fig. 13. The effect of directionality and context on hypothesis generation. (a) When Ray C is shot in at C1 and exits at C2, it is commonly explained by placing Atom 3 as shown. (b) When Ray C is reversed (shot in the opposite direction), it results in a completely different explanation.

5.2. Experiment 2: use of the current explanation in refining abstract hypotheses

The goal of the second experiment was to investigate the assumption that people will use the current explanation if they need to refine an abstract or disjunctive hypothesis. For example, in Black Box an abstract hypothesis is often given for a zig-zag pattern (see Fig. 11). From the input- and the output-cells the path can only be inferred on an abstract level (as described in Fig. 11b). Our theory assumes that the current explanation is used to prune the space of alternative hypotheses for the new evidence. As explained earlier, the first step in hypothesis generation is independent of the current explanation, but if the generated hypothesis is abstract or disjunctive, it is refined by referring to the existing explanation. The model does this refinement by sequentially processing the ray path from start to finish, comparing the atoms needed to support abstract parts of the path with any previously placed or hypothesized atoms. This implies that the previously placed atoms used to refine the new explanation should depend on directionality of processing—the direction in which the ray path is examined.

From a formal point of view, directionality should not matter because the same evidence is used along with identical causes. Therefore, the probabilities for different hypotheses do not depend on the direction in which the path is processed. However, our model predicts that, depending on the direction, different atoms will be used to constrain the search.

If directionality and the current explanation matters then people will develop different causal explanations when the existing explanation is kept constant, but the direction in which the ray is processed is reversed. Fig. 13 illustrates how directionality can result in different explanations. In both scenarios, Rays A and B are shot first followed by Ray C. If Ray C is shot from C1 and emerges at C2 as shown in Fig. 13a, participants should tend to place the third atom below Atom 2. However, if Ray C is reversed, as shown in Fig. 13b, participants should tend to place the third atom above Atom 1, according to the predictions of our model.

5.2.1. Method

5.2.1.1. Participants. The same 15 students who took part in the first experiment also participated in the second experiment. The experiment was conducted one week after the first study.

5.2.1.2. Apparatus. The same equipment as in experiment one and two was used.

5.2.1.3. Stimuli. Five configurations were constructed for the evaluation of the experimental predictions. Four configurations contained a zig-zag pattern with two atoms (such as Ray C in Fig. 13a and b), one configuration had a more complex ray path that used four atoms. The four two-atom tasks were generated by rotating the zig-zag pattern shown in Fig. 13 according to the four sides of the Black Box. The zig-zag pattern was given after the participant had placed atoms to explain at least four other observations. This explanation could then be used as context for explaining the new observation. A second set of five tasks was constructed by exchanging the input and output cells of the critical zig-zag pattern. According to the model, swapping the input and output cells (i.e., reversing the direction of ray travel) should lead the participants to use different subparts of the existing explanation to constrain hypothesis generation for the new observation. So for every task, two isomorphic versions existed that differed only in that the input and output cell of the zig-zag pattern were swapped. Again two sequences of ten tasks each were constructed. Between the two sequences the factors rotation and complexity (number of hidden atoms) were counterbalanced.

5.2.1.4. Procedure. Since the participants already knew the rules of the Black Box, no training phase was needed. The session started with three practice trials to make sure that participants still remembered and understood the rules. Other than this, the same procedure was used as in the first experiment. The experimental sessions lasted between 21 and 35 min ($M = 29.3$).

5.2.2. Results

Again statistical analysis used only those solutions in which the individuals placed the necessary atoms prior to the critical new observation and in which the explanation for the new observation used an atom in one of the two critical positions. From 150 solutions (10 tasks x 15 participants) 140 could be used (93.3%). The trials lasted from 17.4 sec to 251 sec ($M = 56.4$, $SD = 34.4$). With regard to processing time a single factor ANOVA showed no effect of directionality ($M_1 = 56.47$; $M_2 = 54.32$), F (1, 146) $= 0.52$, $p < .48$, but an effect of complexity ($M_{\text{simple}} = 50.32$; $M_{\text{complex}} = 80.68$), F (1, 146) $= 21.7$, $p < .001$. This means that results with respect to the critical factor "directionality" are not confounded with task difficulty or participant motivation.

In building a path, participants could use either of two previously placed atoms as context

Table 3
Proportion of participants who placed an atom on location 1 or 2 as a function of ray direction

Game	Direction			
	1		2	
	Atom location		Atom location	
	1	2	1	2
1	10/11	0/13	0/14	14/14
2	13/15	0/15	0/15	15/15
3	14/14	0/15	0/15	15/15
4	14/14	0/14	3/15	12/15
5	13/13	0/13	6/14	8/14

and then add one more atom to explain the ray. The placement of the additional atom depended on which previously placed atom the subjects used as context. Table 3 shows the number of individuals who placed an atom on one of these two positions as a function of the game and the direction of the path. We code the two positions as Positions 1 and 2 and the two ray directions as Directions 1 and 2. The model predicted that participants would place the atom at Location 1 when the ray was shot in Direction 1 and at Location 2 when the ray was shot in Direction 2. Even without statistical analysis it is obvious that participants traced the path from the input to the output-cell using the atom that was first hit as context to constrain the search for an explanation. With isomorphic tasks that differed only by reversing ray direction, two different multicausal explanations were given.

5.2.3. Discussion

In accordance with our experimental hypothesis the study clearly showed the effect of directionality. The current explanation is definitely used when refining an abstract hypothesis. This leads to different explanations, based on the location of the input cell, independent of the normative structure of the task. This demonstrates how explanatory context, combined with processing heuristics, can lead to potential errors. The error observed here is a kind of order effect on the direction of processing the evidence. The next experiment explores order effects due to different sequences of the same data.

5.3. Experiment 3: the effect of changes to the current explanation on data interpretation

The first experiment showed that the current explanation is still sometimes used if a concrete hypothesis is available. However, results from Experiment 2 clearly point out that the current explanation is used to constrain the search space when an abstract hypothesis is generated. The effect in Experiment 2 hinged on the directionality of processing an abstract ray path, but kept the current explanation constant. Another test of the model is to show that changes in the current explanation affect the interpretation of new data. The third experiment does this by varying the current explanation while keeping constant the datum to be explained. This is done using cases that are designed to elicit order effects, such that evidence presented in one sequence will elicit a different explanation when presented in a different sequence (Hogarth & Einhorn, 1992). The order effect arises because a critical ray appears in the middle of a sequence of ray shots, such that reversing the sequence causes the critical ray shot to be interpreted in the context of a different current explanation. Order effects of this type are considered a bias of human reasoning, because there is no normative reason to prefer one explanation over another simply because of a change in the sequence of evidence presentation. Order effects were shown in work on impression formation (Asch, 1946), deductive reasoning (Johnson-Laird & Steedman, 1978) and causal reasoning (Hogarth & Einhorn, 1992).

5.3.1. Method

5.3.1.1. Participants. 19 students between 23 and 27 years old ($M = 25.4$ years, $SD = 0.7$) participated in the experiment. There were 5 women and 14 men. The participants responded to announcements posted at the campus of the University of Regensburg. Participants received course credit for participating. None had significant prior experience with Black Box.

5.3.1.2. Apparatus. The same equipment as in experiment one and two was used.

5.3.1.3. Stimuli. 28 four-atom configurations were constructed. They consisted of 14 test tasks that were designed to provoke order effects. The other 14 tasks were control tasks, in which no order effects were predicted, because every datum could be explained independently of others. For the control tasks no order effect was expected. Every 4-ball configuration consisted of six rays. The 14 test tasks as well as the control tasks consisted of seven paired tasks (A B) that only differed in the sequence of evidence. So in sequence B the last evidence of version A was presented first, the next to the last was presented second, and so on (see Fig. 14). The ray shots in the test tasks were constructed so that the rays could be explained by placing either the second or third atom at either of two different locations. The dependent variable was whether this atom placement depended on the order of evidence.

5.3.1.4. Procedure. Participants were tested individually in an experimental session that lasted 92 min on average (from 62 to 148 min, $SD = 12.1$). In the training phase the individuals first had to learn the rules and then had to apply the rules in the prediction mode. Training was completed as soon as the participants correctly predicted all rays for three trials in a row. In the test phase the 28 tasks were given in random order with the constraint that the two versions of every test game never occurred together.

5.3.2. Results

Data were again placements of the atoms and total processing time. The average processing time for all tasks was 84.2 s. In an ANOVA with sequence (sequence A: $M = 86.1$ s; sequence B: $M = 82.3$ s) and test versus control (test: $M = 80.3$ s, control: $M = 86.7$ s.) as factors no significant main or interaction effects, $F(1, 542) = 1.38$, $p < .24$, for test versus

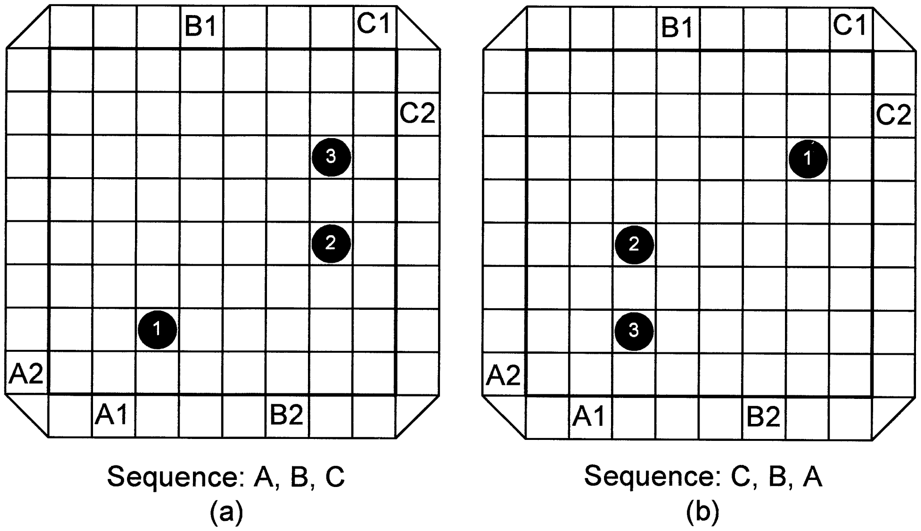


Fig. 14. The effects of order of evidence on hypothesis generation. (a) The explanation commonly given when rays are presented in the order A, B, C. (b) The explanation given when the rays are presented in reverse order (C, B, A). The numbers next to each letter specify the input cell (1) and the output cell (2).

control, F (1, 542) $=$ 0.49, $p < .48$, for sequence and F (1, 542) $=$ 1.06, $p < .3$, for the interaction were found. This means that the tasks did not differ in difficulty.

It was investigated whether test- and control-tasks differed in general between sequence A and B, that is, if the sequence of data presentation affected explanation generation. This was done by determining how many placements in version A were identical to version B. In the control-tasks, the locations of atoms (hypotheses) were identical in 94.2% of all tasks; in the test-tasks, however, only 78.9% of the placements of version B were identical with placements in version A. In 420 out of 532 test tasks, participants placed atoms between the versions in identical places, compared to 501 out of 532 in the control tasks. By means of a binomial test (asymptotic) a significant difference between condition A and B was found, $u = 14.5$, $p < .0001$. This means that the test-tasks differed in versions A and B much more than the control tasks. On a more global level we analyzed how many of the participant's solutions differed between the two sequences of ray shots. Fig. 15 shows the frequency distribution built by counting how many participants produced 0 to 7 identical solutions within the 14 control or test tasks. The frequency distribution shows that participants in control tasks more often chose the same locations between the two sequences.

A more detailed analysis looked at "critical" atoms. The tasks were constructed in a way that different sequences should affect only specific atoms. Depending on the sequence, the location pattern of these critical atoms should differ from the placement of other atoms. Therefore, we counted the number of cases in which the placement of critical atoms differed between the sequences. Noncritical atoms were identical in 97.6% of all cases. This ratio did not differ between control and test-tasks so the major difference between placements is clearly due to those atoms that we expected to be affected by sequence.

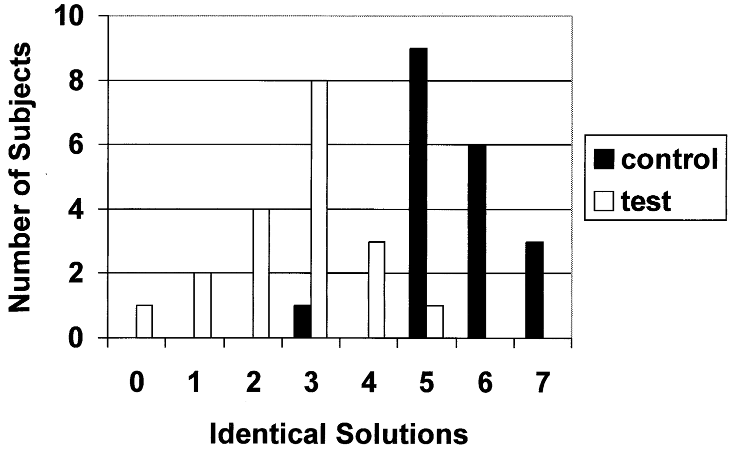


Fig. 15. Number of participants who placed in 0 to 7 cases critical atoms on the same position between the two sequences.

The model's predictions provide an "expected" location for the critical atoms. Within the control tasks this location should not differ between the two sequences. However, for test tasks, depending on the sequence, there are two different locations for the critical atom. If the current explanation is used, and if therefore an order-effect is produced, then the distribution of atom placements on "expected" positions should not differ between sequences. This prediction was supported by a χ^2 -test. No significant differences between the frequency distributions of the two sequences of test tasks, χ^2 (1, $N = 19$) = 0.79, $p > .1$, and control tasks, χ^2 (1, $N = 19$) = 7.5, $p > .05$, could be found.

5.3.3. Discussion

No differences between the test and control trials were found with regard to processing time. This means that the tasks were comparable in terms of difficulty. On average individuals spent similar time solving the tasks, independent of the explanation generated.

The explanations developed by the individuals clearly depended on the sequence of data presentation. So an order-effect in the sense of Hogarth and Einhorn (1992) was verified. This result is consistent with the experimental hypothesis that the current explanation is used to refine abstract or disjunctive hypotheses. Order effects occur because the simplest consistent hypothesis for a new datum is immediately added to the situation model and then used to constrain the interpretation of succeeding data. By changing the order of data, the context in which new evidence is evaluated is changed, leading to the selection of different hypotheses.

6. General discussion

Although the current explanation can narrow the search for hypotheses, it can also blind the problem solver to alternative, possibly better, hypotheses. This trade-off provided the motivation for studying how and when people use the current explanation as context. Based on a task analysis and general cognitive constraints, we proposed a model in which the current explanation is not used if a simple hypothesis is available to explain new data, but is always used when discriminating among alternative hypotheses for new data. To test the model, we conducted three experiments. Our results are consistent with the assumption that the current explanation is used when discriminating among alternative hypotheses. However, the assumption that the current explanation is not used when a simple hypothesis is available, received only limited support. Participants in four out of six task types used the current explanation to constrain their interpretation of new data. This suggests that context-independent strategies compete with context-dependent ones—an interpretation that is consistent with recent work on strategy selection during problem solving. Lovett and Anderson (1996) showed that the probability of selecting a problem solving strategy was independently affected by problem context and the strategy's past history of success.

It is worth comparing the model proposed in this paper to alternative models of abduction. Several Artificial Intelligence (AI) models have been proposed (de Kleer & Williams, 1987; Pearl, 1988; Peng & Reggia, 1990; Reiter, 1987, see Josephson & Josephson, 1994, for a summary), but most of these are formal prescriptive theories of abduction, hence they do not make very good cognitive models of human behavior. For example many of the AI models consider vast numbers of possible explanations and then generate all possible explanations that fit certain formal criteria for a best explanation. In contrast, even in complex tasks such as diagnosis, expert's consider only a small number of alternatives and usually work to select a single best explanation (Arocha & Patel, 1991; Feltovich et al., 1984).

Thagard's (Thagard, 1989) Theory of Explanatory Coherence (TEC) is a model of hypothesis selection (or evaluation), not hypothesis generation. TEC assumes that people select the explanation that best coheres with their beliefs. TEC prescribes a set of principles that define coherence (and incoherence). These are implemented in a connectionist model, called ECHO, which combines evidence in parallel to determine the most coherent explanation. Some researchers have argued that ECHO is not a good cognitive model because it assumes that humans have the ability to consider all of the interdependencies between an explanation and the available data (see open peer commentary in Thagard, 1989). ECHO assumes that the relationship between all observations and all possible explanations can be quickly updated in parallel. However, this parallel process seems appropriate for the kinds of automatic unconscious processes involved in evaluating explanations (see Thagard & Kunda, 1998, for a discussion of the relation between conscious deliberative and unconscious parallel aspects of abduction). In addition, several researchers have modified ECHO to produce some of the same biases exhibited by humans reasoning under uncertainty, such as order effects (Hoadley et al., 1994; Wang et al., 1998; Wang et al., 2000). The model described in this paper makes no attempt to model automatic processes; however, Johnson et al. (1997) have argued that a complete model of abduction requires a hybrid model composed of the kind of sequential, deliberate model described in this paper, and a subsymbolic, highly parallel model, such as ECHO.

Although Bayesian networks have proven quite useful for capturing and deploying expert knowledge of uncertainty and causality, they seem ill-suited as models of human cognition, because they cannot capture the kinds of biases that we found in our experiments. Bayesian networks determine the normative probability distribution of every unknown observation and hypothesis in the network, hence they cannot capture biases such as the order effect. In addition, Bayesian networks do not provide a model of hypothesis generation, since the network is assumed to contain all possible observations and hypotheses.

The generation of causal explanations was also investigated in work on scientific discovery. Klahr and Dunbar's (1988) model of Scientific Discovery as Dual Search (SDDS) is based on the coordination of search in a hypothesis space and an experiment space. This theory provides a general explanation for how search in the hypothesis and experiment spaces interacts. It also clearly defines three roles for experiments (exploring, hypothesis testing, and hypothesis refinement) and indicates how these roles affect the developing hypothesis. Klahr and his colleagues have been quite successful at using the model to explore differences among good and bad reasoners and to study developmental differences (Dunbar & Klahr, 1989). 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. For example, *Evaluate Evidence* is listed as a process in SDDS that "decides whether the cumulative evidence—as well as other considerations—warrants acceptance, rejection, or continued consideration of the current hypothesis."(Dunbar & Klahr, 1989, pg. 33) Thus, while SDDS models 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 the subtasks of abduction. In contrast to SDDS, our model proposes detailed theories for many of the important subtasks of abduction, including evidence integration and hypothesis generation.

Future research must evaluate the model we have outlined in more detail and improve the model's behavior based on comparisons with human data. The mixed results regarding the use of the current explanation during evidence interpretation deserve further attention. Many practical applications of this research, such as understanding and improving medical decision making, will require a more complete understanding of the situations in which current explanations are used or ignored. Practical applications must also bear in mind that use of the current explanation is a double-edged sword—it assists the problem solver by limiting the space of possibilities, but can also blind the problem solver to potential solutions.

Notes

1. This distinction between abduction and induction differs from Peirce's later view in which abduction is defined as the generation of an explanatory hypothesis (which might be a rule) and induction as the process that justifies the explanation. See Flach (1996) for a discussion of Peirce's definitions of abduction.
2. Black Box was designed by Eric Solomon and has been marketed under various names.
3. This explanation actually contains one error and completely ignores another possibility. Atom 4 (see Fig. 9) would actually cause Ray B to reflect to the left, not to be absorbed. In addition, an atom located immediately to the right of the existing atom would absorb Ray B. We believe that these mistakes are due to a heuristic often

mentioned by the participants. Namely, that an absorption is caused by an atom that lies somewhere along the trajectory of the ray as defined by the existing atoms.

Acknowledgments

This research was supported in part by a seed grant from The Ohio State University, a German American Collaborative Research Grant from the American Council of Learned Societies and the German Academic Exchange Program (DAAD), and a grant from Vielberth-Stiftung, University of Regensburg, Germany. Hans Bogenberger programmed the Windows version of Black Box used in the three experiments. Much of this research was conducted while Todd R. Johnson was at The Ohio State University, Department of Pathology and Josef Krems was at the University of Regensburg, Department of Psychology. We would like to thank Kurt VanLehn, Paul Thagard, Todd Griffith, Patricia Schank, and Hongbin Wang for their insightful comments concerning this manuscript.

References

Arocha, J. F., & Patel, V. L. (1991). Hypothesis generation and the coordination of theory and evidence in medical diagnostic reasoning. In K. J. Hammond & D. Gentner (Eds.), *Proceedings of the thirteenth annual conference of the cognitive science society* (pp. 623–628). Hillsdale, NJ: Lawrence Erlbaum Associates.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258–290.

Bayazitoglu, A., Smith, J. W., & Johnson, T. R. (1992). A diagnostic system that learns from experience. In E. Bolger (Ed.), *Proceedings of the sixteenth annual symposium on computer applications in medical care* (pp. 685–689). New York: McGraw-Hill, Inc.

Bylander, T., Allemang, D., Tanner, M. C., & Josephson, J. R. (1991). The computational complexity of abduction. *Artificial Intelligence, 49*(1–3), 25–60.

de Kleer, J., & Williams, B. (1987). Diagnosing multiple faults. *Artificial Intelligence, 32,* 97–130.

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, New Jersey: Lawrence Erlbaum Associates.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving—An analysis of clinical reasoning.* Harvard: Harvard University Press.

Feltovich, P. J., Johnson, P. E., Moller, J. H., & Swanson, D. B. (1984). LCS: The role and development of medical knowledge in diagnostic expertise. In W. J. Clancey & E. H. Shortliffe (Eds.), *Readings in medical artificial intelligence.* Reading: Addison-Wesley.

Flach, P. (2001). (1996). Abduction and induction: Syllogistic and inferential perspectives. In P. A. Flach & A. Kakas (Eds.), *Proceedings of the ECAI '96 workshop on abductive and inductive reasoning* (pp. 31–35). Budapest: 12th European Conference on Artificial Intelligence. Available at: http://citeseer.nj.nec.com/129082.html. Accessed July 5.

Gardenfors, P. (1992). Belief revision: An introduction. In P. Gardenfors (Ed.), *Belief revision.* New York: NY: Cambridge University Press.

Groen, G. J., & Patel, V. L. (1998). The relationship between comprehension and reasoning in medical expertise. In M. Chi & R. Glaser (Eds.), *The nature of expertise* (pp. 287–310). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoadley, C. M., Ranney, M., & Schank, P. (1994). WanderECHO: A connectionist simulation of limited coherence. In A. Ram & K. Eiselt (Eds.), *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 421–426). 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, K. A., Johnson, T. R., Smith, J. W., Jr., DeJongh, M., Fischer, O., Amra, N. K., & Bayazitoglu, A. (1991). RedSoar—A system for red blood cell antibody identification. In P. D. Clayton (Ed.), *Proceedings of the fifteenth annual symposium on computer applications in medical care* (pp. 664–668). New York, NY: McGraw-Hill.

Johnson, T. R., & Smith, J. W. (1991). A framework for opportunistic abductive strategies. In K. J. Hammond & D. Gentner (Eds.), *Proceedings of the thirteenth annual conference of the cognitive science society* (pp. 760–764). Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, T. R., Krems, J., & Amra, N. K. (1994). A computational model of human abductive skill and its acquisition. In A. Ram & K. Eiselt (Eds.), *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 463–468). Lawrence Erlbaum Associates.

Johnson, T. R., Zhang, J., & Wang, H. (1997). A hybrid learning model of abductive reasoning. In R. Sun & F. Alexandre (Eds.), *Connectionist symbolic integration* (pp. 91–112). Mahweh, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology, 10,* 64–99.

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

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163–182.

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

Krems, J. (1994). *Wissensbasierte urteilsbildung* [Knowledge-based diagnostic reasoning]. Bern: Huber.

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

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

Lovett, M. C., & Anderson, J. R. (1996). History of success and current context in problem solving: Combined influences on operator selection. *Cognitive Psychology, 31*(2), 168–217.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs. N.J.: Prentice-Hall.

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

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

Peirce, C. S. (1839–1914). *Collected papers of Charles Sanders Peirce.* Edited by C. Hartshorne, P. Weiss, & A. Burks. Cambridge, MA: Harvard University Press, (1931–1958).

Peng, Y., & Reggia, J. A. (1990). *Abductive inference models for diagnostic problem-solving.* New York: Springer.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 521–533.

Reiter, R. A. (1987). A theory of diagnosis from first principles. *Artificial Intelligence, 32,* 57–95.

Smith, J. W., Jr., Bayazitoglu, A., Johnson, T. R., Johnson, K. A., & Amra, N. K. (1995). One framework, two systems: flexible abductive methods in the problem-space paradigm applied to antibody identification and biopsy interpretation. *Artificial Intelligence in Medicine, 7,* 201–225.

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.

Thagard, P., & Kunda, Z. (1998). Making sense of people: coherence mechanisms. In S. J. Read & L. C. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 3–26). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, H., Johnson, T. R., & Zhang, J. (1998). UECHO: A model of uncertainty management in human abductive reasoning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the twentieth annual meeting of the cognitive science society* (pp. 1113–1118). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, H., Zhang, J., & Johnson, T. R. (2000). Human belief revision and the order effect. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the cognitive science society* (pp. 547–552). Mahweh, NJ: Lawrence Erlbaum Associates.