Memory & Cognition
2009, 37 (3), 249-264
doi:10.3758/MC.37.3.249

The role of learning data in causal reasoning about observations and interventions

Björn Meder, York Hagmayer, and Michael R. Waldmann
University of Göttingen, Göttingen, Germany

Recent studies have shown that people have the capacity to derive interventional predictions for previously unseen actions from observational knowledge, a finding that challenges associative theories of causal learning and reasoning (e.g., Meder, Hagmayer, & Waldmann, 2008). Although some researchers have claimed that such inferences are based mainly on qualitative reasoning about the structure of a causal system (e.g., Sloman, 2005), we propose that people use both the causal structure and its parameters for their inferences. We here employ an observational trial-by-trial learning paradigm to test this prediction. In Experiment 1, the causal strength of the links within a given causal model was varied, whereas in Experiment 2, base rate information was manipulated while keeping the structure of the model constant. The results show that learners' causal judgments were strongly affected by the observed learning data despite being presented with identical hypotheses about causal structure. The findings show furthermore that participants correctly distinguished between observations and hypothetical interventions. However, they did not adequately differentiate between hypothetical and counterfactual interventions.

Causal knowledge is central to explaining past events, predicting future events, and the planning of actions. For example, if we experience certain symptoms, such as a cough, fever, and a headache, we can draw inferences about the underlying causes (e.g., a viral or bacterial infection). We can also capitalize on our causal beliefs to take actions, such as using painkillers to relieve the symptoms and antibiotics to treat the underlying bacterial infection. Such inferences are grounded in our capacity to recognize the asymmetry of causal relations (causes generate effects, but not vice versa) and the ability to represent the world in terms of mental models that reflect the causal texture of our physical, biological, and social environment (e.g., Glymour, 2003; Gopnik et al., 2004; Sloman, 2005; Sloman & Hagmayer, 2006; Waldmann, 1996; Waldmann, Hagmayer, & Blaisdell, 2006).

But how do people acquire, represent, and use this knowledge when making causal inferences? According to associative learning theories, causal reasoning is driven by associative relations that have been learned on the basis of observed event covariations. This view suggests that causal learning is basically a conditioning process in which organisms learn to associate particular cues (the cause events) with particular outcomes (the effect events). The general claim is that causal learning can be reduced to associative learning, that causal knowledge is basically associative knowledge, and that causal judgments are a function of associative strength (e.g., Dickinson, 2001; Shanks, 2007). This idea has been challenged by *causal model theory* approaches (Sloman, 2005; Waldmann, 1996; Waldmann & Holyoak, 1992; see also Beckers, De Houwer, & Matute, 2007; De Houwer, Beckers, & Vandorpe, 2005; Griffiths & Tenenbaum, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003), which postulate that people use the learning input to induce causal structures with properties that go beyond mere associations (for recent overviews, see Gopnik & Schulz, 2007; Waldmann et al., 2006). This argument is predicated on the idea that mental representations that merely mirror experienced patterns of covariations are insufficient to explain the competencies people have in dealing with causal situations. For example, if we learned only to associate observed events, without accessing deeper information about causality, we would be unable to understand the differences between spurious (i.e., merely correlational, noncausal) and causal relations (Meder, Hagmayer, & Waldmann, 2006; Waldmann & Hagmayer, 2001) and could not differentiate between prediction and diagnosis (Waldmann, 1996, 2000; Waldmann & Holyoak, 1992; Waldmann & Walker, 2005). For example, without the categories of cause and effect and sensitivity to causal directionality, we would be unable to understand that diseases are the causes of symptoms (and not vice versa) and that different symptoms of a disease may covary, due to a common cause, without being directly causally related.

Another challenge for associative models is the distinction between different modes of accessing causal knowledge (Meder, Hagmayer, & Waldmann, 2008; Waldmann & Hagmayer, 2005; see also Vadillo & Matute, 2007; Vadillo, Miller, & Matute, 2005). A number of researchers have emphasized that there are fundamental differences between inferences based on merely observed states of variables (*seeing*) and the very same states generated by means of external interventions (*doing*) (Dawid, 2002; Meek & Glymour, 1994; Pearl, 2000; Spirtes, Glymour, & Scheines, 1993;

B. Meder, meder@mpib-berlin.mpg.de

Woodward, 2003). For example, observing the state of a barometer allows us to make predictions about the upcoming weather (observational inference), whereas manipulating the barometer does not license such a prediction (interventional inference). Whereas observational inferences allow us to capitalize on both causal and noncausal correlations, interventional predictions are based only on predictive causal relations. Contrary to associative approaches, causal model theory (e.g., Waldmann et al., 2006) assumes that people represent causal structure in learning and reasoning. Thus, people are assumed to differentiate between causes, which may affect effect events when set by an intervention, and effects, which have no impact on their causes. Causal mental models enable people to draw inferences about both novel observations and novel interventions. Hence, these theories are able to differentiate between *seeing* and *doing* (Waldmann & Hagmayer, 2005). Different causal model accounts, however, differ in their assumptions about whether and to what extent people take into account the parameters of the causal models (e.g., the strength of causal relations, base rates of events). Qualitative accounts (e.g., Sloman, 2005) claim that people focus mainly on structure but largely ignore parameters. Power PC theory (Cheng, 1997) and causal model theory (Waldmann, 1996) assume that people infer the most likely parameters of a single hypothesized causal model from observations. Causal Bayes net theories, finally, assume that people derive probability distributions over multiple candidate causal models and their parameters from data (e.g., Griffiths & Tenenbaum, 2005).

The present article focuses on two questions: The first goal is to examine to what extent people take into account both causal structure and learning data during causal learning and subsequent reasoning. More specifically, we investigate whether people learn about the parameters of a causal model during observational trial-by-trial learning and, in turn, later use the parameterized model to make causal inferences about novel situations. The second goal is to investigate whether people differentiate between hypothetical observations, hypothetical interventions (i.e., interventions participants have not taken or seen before), and counterfactual interventions, which hypothetically change a factual state of the world. Although most theories of causal learning (e.g., associative accounts) lack the representational power to express the conceptual differences between these inferences, causal model theories capture this crucial distinction. Our goal is to examine whether causal model theory captures the intuitions of people, too.

Previous Empirical Evidence

Thus far, only a few studies have addressed the question of whether people are sensitive to the difference between observations (*seeing*) and interventions (*doing*). Sloman and Lagnado (2005) compared logical with causal reasoning and demonstrated that people correctly distinguished between observations and interventions and arrived at different conclusions, depending on whether events were merely observed or actively generated. Waldmann and Hagmayer (2005) investigated reasoning about hypothetical observations and interventions in a learning task. The participants in their experiments were first shown diagrams depicting causal models—that is, hypotheses about the structure of the causal situation (similar to Figure 1, but without the numbers). Subsequently, the participants received a list of cases that they were supposed to use to pick up the models' parameters (causal strength and base rates of causes). Reasoning with both deterministic and probabilistic causal relations was examined. The results showed that the participants understood the differences between seeing and doing and that they used the parameters, which were gleaned from the learning data, in their inferences. For example, given a simple causal model with one cause event and two effect events, the participants inferred the presence of one effect from observing the other, but they did not draw this conclusion when the presence of the first effect was not merely observed but was generated by means of external intervention. One limitation of this study as a learning experiment, however, was the presentation of the data in list format. One could argue that this type of highly aggregated data presentation turned the task more into a reasoning than a learning task. Therefore, Meder et al. (2008) moved one step further into the realm of learning by using a trial-by-trial learning paradigm. Their findings showed that learners are capable of deriving interventional predictions from passive observations of causal systems—a result inconsistent with the assumption that predictions of the consequences of instrumental actions require a prior phase of interventional learning. A second finding of the study was that the way the learning data were presented affected learners' judgments. Meder et al. (2008) manipulated temporal cues while holding the instructed causal model and learning data constant. Participants' judgments were more sensitive to confounding pathways when temporal order during learning conformed to the causal order (predictive learning from causes to effects) than when temporal order was reversed (diagnostic learning from effects to causes). In this study, sensitivity for the parameter values was not directly assessed, though.

Goal of the Experiments

Thus far, only the Meder et al. (2008) study has used a trial-by-trial learning paradigm with human participants to examine causal inferences about observations and interventions (but for related studies with rats, see Blaisdell, Sawa, Leising, & Waldmann, 2006; Leising, Wong, Waldmann, & Blaisdell, 2008). However, in this study, the parameters of the learning input were not manipulated, only the way in which the data was presented (diagnostic vs. predictive learning). In contrast, in the Waldmann and Hagmayer (2005) experiments, the statistical patterns were varied, but a trial-by-trial learning paradigm was not employed. Thus, we currently have no empirical evidence regarding the question of whether people learn about the parameters of causal models during passive trial-by-trial observational learning and later use these parameters to derive inferences about the consequences of potential actions they have never seen or taken before. Therefore, the first main goal of the present experiments was to investigate whether people are sensitive to different parameterizations of a given causal model when deriving observational and interventional predictions. To test this assumption, two experiments were conducted in which participants were given identical causal structures

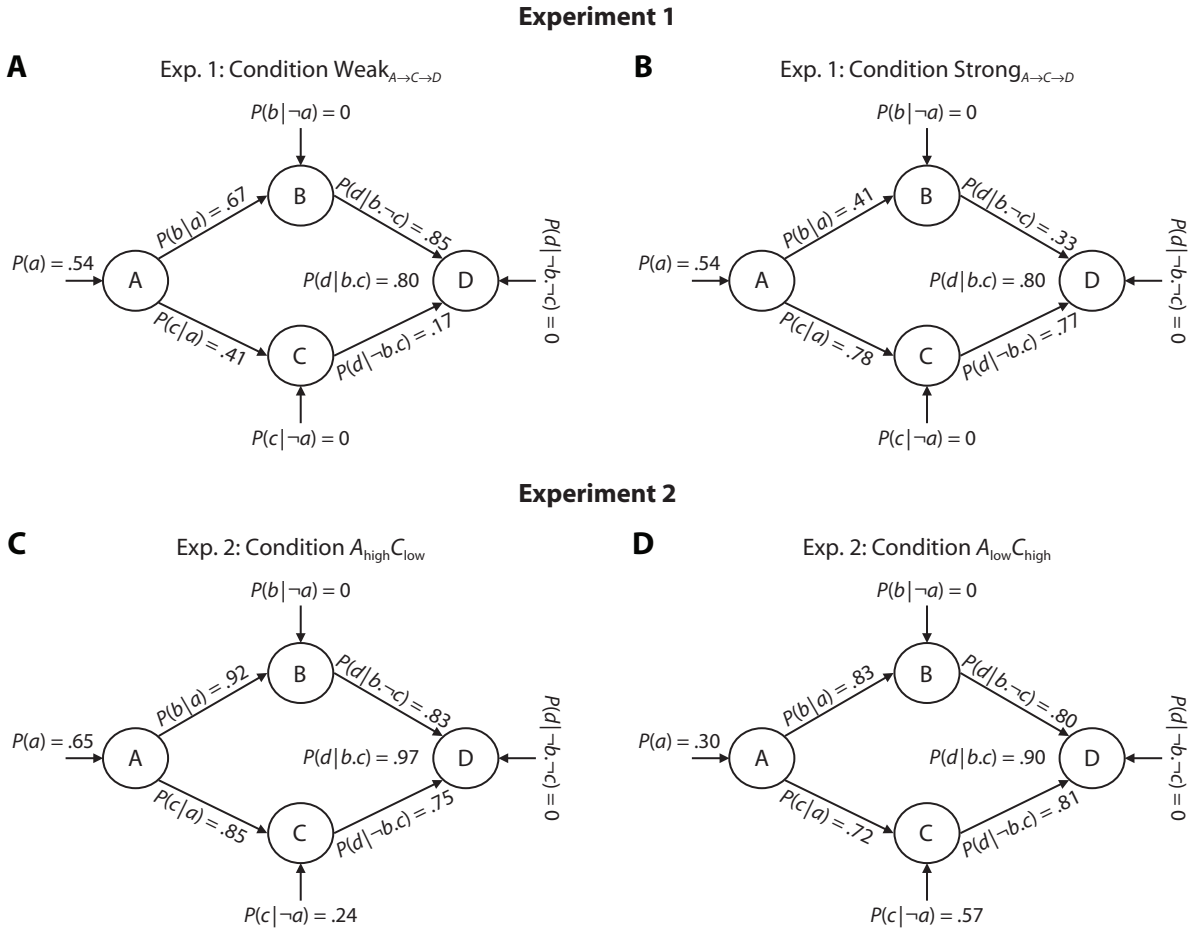

Figure 1. Parameterized causal models used in Experiments 1 and 2. Arrows indicate causal relations between variables; probabilities encode the strengths of these relations. The parameter $P(d|b.c)$ approximates a noisy-OR gate (Pearl, 1988).

but were provided with different kinds of data during observational learning. In Experiment 1, the causal strength of the links within the causal model was varied, whereas in Experiment 2, the base rates of the causes were manipulated. If learners took into account the learning data, their causal inference should vary systematically in accordance with the provided learning input. By contrast, if they consider only causal structure, no differences should result.

A second goal of the experiments was to investigate inferences about different types of interventions. All previous studies had focused on the difference between observation and intervention. Accordingly, participants' inferences based on passively observed states of events were compared with inferences based on the very same states generated by hypothetical interventions. Although, in the literature, both hypothetical and counterfactual reasoning is sometimes generally referred to as *what if* thoughts (see Dawid, 2006), these two types of inferences are clearly different from a theoretical point of view. Consider the following example. Imagine tinkering with your friend's scale. If you rig the scale *before* he uses it (i.e., a hypothetical intervention), no inferences can be drawn from the reading about his actual weight. However, if you manipulate the scale *after* reading it and observing that he has lost 10 pounds (an intervention that changes the factual reading—i.e., a counterfactual intervention), you still can be pretty sure that he will need new clothes, although the scale may now indicate the opposite. Causal models allow for differential inferences for hypothetical and counterfactual interventions (see the next section). To study whether people also differentiate between these two types of inferences, we asked participants in both experiments to make inferences about novel observations, hypothetical interventions, and counterfactual interventions.

Modeling Causal Inferences About Observations and Interventions

A formal treatment of causal models is provided by causal Bayes net theories (Pearl, 1988, 2000; Spirtes et al., 1993). The formalism uses directed acyclic graphs (DAGs) to represent causal relations between variables and parameters to express the strength of these relations (e.g., conditional probabilities) and the base rates of the involved events. An example is given in Figure 1.

This causal model consists of four (binary) variables, A, B, C, and D, in which A can cause D via either B or C. In the Bayes net framework, the joint probability distribution of this model is decomposed into a set of conditional dependence and independence relations by applying the

causal Markov condition to the causal model (for further details, see Pearl, 2000; Spirtes et al., 1993). For each variable, the Markov condition defines a local causal process in which the state of the variable is a function only of its direct causes (its Markovian parents). This causally based factorization and the implied relations of conditional dependence and independence enable and facilitate the modeling of different types of probabilistic causal inferences within complex causal networks (Pearl, 2000).

Modeling observations. On the basis of the structure of the causal model and its parameters, the probabilities implied by observed values of variables can be computed using standard probability calculus. For example, observing the presence of C implies that its cause, A, has occurred with a probability of $P(a \mid c)$, which can be computed from the parameters of the model by using Bayes rule, $P(a \mid c) = P(c \mid a) \cdot P(a) / P(c)$. A more interesting example is the prediction of Variable D from an observation of Variable C. Obviously, there is the direct causal link connecting C to D; but there is also a second causal pathway connecting C to D via A and B. Pearl (2000) vividly calls such confounding pathways *backdoors*. Observational probabilities include this alternative causal path because observed values of C provide diagnostic evidence for A and, therefore, also include the influence of D's alternative cause, B (see the Appendix for the exact derivations).

Modeling hypothetical interventions. Predictions for interventions can also be derived from the parameterized model. The literature on causal Bayes nets has focused on ideal ("atomic") interventions in which the intervention changes the value of a variable independently of the state of the variable's parents (for more precise characterizations of interventions, see Woodward, 2003). The characteristic feature of such interventions is that they create independence, since the value of the variable intervened on is no longer dependent on its typical causes. For example, if we manipulate the barometer reading, its state no longer depends on its "normal" cause, atmospheric pressure. A number of different notations have been suggested to represent interventions in causal networks. The most prominent one is Pearl's (2000) do-operator, *do* (•). Using this notation, the probability $P(a \mid c)$ refers to the probability that A will be present given that C was *observed* to be present, whereas the expression $P(a \mid \text{do } c)$ refers to the interventional probability that A will be present given that C was generated by means of intervention. A similar approach is found in the work of Spirtes et al. (1993), who used the expression *set* [e.g., $P(a \mid \text{set } c)$] to denote the difference between different merely observed and actively generated events. A general representation of outside interventions within causal models is provided by augmenting causal model representations with intervention nodes representing additional cause variables (see Dawid, 2002; Spirtes et al., 1993). Within this framework, predictions of the outcomes of different types of interventions can be modeled as probabilistic inferences (Waldmann, Cheng, Hagmayer, & Blaisdell, 2008). To simplify derivations, we will here use the *do*-notation to distinguish observations from interventions.

To illustrate the difference between observational and interventional inferences, consider the diamond-shaped causal model shown in Figure 1 (see the Appendix for the formal derivations.). According to this model, the initial event, A, can generate the final effect, D, by way of the intermediate variables, B and C. Within this model, for example, the observed states of C provide diagnostic evidence for the state of its cause, A—thus, $P(a \mid c) > P(a \mid \neg c)$ (given a generative causal relation $A \rightarrow C$). By contrast, due to the asymmetry of causal relations, manipulations of C do not change the probability of A, which therefore remains at its base rate [i.e., $P(a \mid \text{do } c) = P(a \mid \text{do } \neg c) = P(a)$]. The fact that interventions create independence also implies a difference between observations and interventions when reasoning from C to D. Obviously, there is the direct causal link connecting C to D, but there is also a second cause of D, Event B. This *backdoor* path is crucial for reasoning about observations of and interventions in C. For example, observing C to be absent indicates that A and, therefore, also B are likely to be absent. Therefore, Event D has only a low probability of being present. However, the situation is different when C is not merely observed to be absent but is actively prevented from occurring. Although this intervention ensures that Event D is not influenced by C, the model's initial Event A might still occur with its base rate probability $P(a)$ and influence D by way of B. Thus, D is less likely to be present when C is observed to be absent than when C is actively prevented by means of an intervention [i.e., $P(d \mid \neg c) < P(d \mid \text{do } \neg c)$].

Note that the do-operator allows us to model the consequences of actual interventions (i.e., what will happen if the intervention is actually executed) and to derive predictions for hypothetical interventions that may actually never be taken (i.e., what would happen if an intervention were taken). In both cases, a variable's state is set by the external intervention.

Modeling counterfactual interventions. Causal models and Bayes nets can also be used to model inferences about counterfactual interventions. Whereas hypothetical interventions assume that the manipulated variable's state prior to the intervention is not known, counterfactual interventions are defined as actions that run counter to the factual course of events. The crucial difference between modeling hypothetical and counterfactual actions is that the latter require us to take into account the diagnostic information provided by the factual observation. Thus, counterfactual interventions combine inferences about observations with inferences about interventions. For example, consider the causal model shown in Figure 1: Substance A causes Substances B and C, each of which can then independently cause Substance D. Now assume that we have observed Substance C to be present. A counterfactual inference might be the following: "C is present. What would have been the state of A if C (and only C) had been prevented from occurring by means of an external intervention?" This question refers to a counterfactual intervention, since the action is logically incompatible with the known occurrence of C in the actual world. Formally, this counterfactual inference is represented by the counterfactual probability $P(a \mid c.\ \text{do } \neg c)$. This term denotes the conditional probability of $A = a$, given that $C = c$ was observed but counterfactually removed (the period in the formula refers to "&").

The basic logic of computing such counterfactual probabilities is to combine observational and interventional inferences (see Pearl, 2000). First, the causes of the variable targeted by the counterfactual intervention are updated (i.e., conditionalized on) in accordance with the diagnostic information provided by the observation. In the example, observing C to be present raises the probability of its cause, A [i.e., $P(a \mid c) > P(a)$]. This probability update is followed by applying the do-operator to C, which renders C independent of A. The crucial point is that the do-operator is applied *after* the probability of C's cause, A, has been updated by the observed evidence. Finally, these two pieces of information are integrated to make a causal inference. For example, if we counterfactually remove the presence of C, this does not change the probability of A, which has been updated by the actually observed presence of C. Thus, $P(a \mid c.\ \text{do}\ \neg c) = P(a \mid c)$. This, in turn, has consequences for the probability of D, which depends on the factual observation of C being present *and* the counterfactual intervention of removing C [i.e., $P(d \mid c.\ \text{do}\ \neg c)$]. Whereas the counterfactually generated absence of C breaks the link $C \rightarrow D$ in the counterfactual world, the actually observed presence of C indicates that D's alternative cause, Event B, is still likely to be present. Thus, in this case, the causal analysis of the situation would indicate that even if C had been prevented from occurring, D would still have a high probability of being generated by its alternative cause, B. In fact, this situation constitutes a particular (probabilistic) version of *causal overdetermination*, a situation that has been used to examine the problems of counterfactual theories of causality (Lewis, 1973).

EXPERIMENT 1

The aim of Experiment 1 was to investigate the role of the learning input by varying the strength of the causal links connecting the observed events of the diamond-shaped causal model (see Figure 1). The experiment's rationale was that potential differences between observations, hypothetical interventions, and counterfactual interventions do not depend only on the structure, but also on the particular parameters of the model, which can be learned from observations. For example, if C is prevented by an inhibitory action, the probability of the final effect, D, crucially depends on the causal strength of the links constituting the alternative causal chain, $A \rightarrow B \rightarrow D$. When the causal path consists of strong causal relations, there is a high probability that D will be generated via this causal path. By contrast, if this alternative causal chain consists of rather weak causal relations, the influence of this backdoor path is negligible. In this case, only minor differences between observational and interventional probabilities will result. If learners' inferences reflect both the model and its parameters, their estimates of the observational, interventional, and counterfactual probabilities should be affected by manipulations of causal strength. By contrast, if their causal inferences are based mainly on the suggested causal structure, no differences are to be expected.

Method

Participants and Design

Thirty-six undergraduate students from the University of Göttingen participated. The *learning data* factor was varied between groups; the *type of inference* and *presence versus absence of C* factors were varied within subjects. The participants received course credit for participation. All the participants were randomly assigned to either of the two conditions.

Procedure and Materials

Causal model instruction. The diamond-shaped causal structure shown in Figure 1 was used. The four variables of the causal model were introduced as fictitious chemical substances causally interacting in wine casks (see Meder et al., 2008). The participants were shown a graph of the causal model similar to that in Figure 1 (but without any numbers) and were told that Substance A causes Substances B and C, each of which can then independently cause Substance D. It was also pointed out that the causal relations are probabilistic. There was no time limit for the inspection of the graphical representation of the causal structure. The participants were instructed to learn more about these causal relations by observing the states of the substances in a number of wine casks. The kinds of questions the learners would have to answer after the learning phase were not mentioned until the test phase. The learners did not see the figure showing the model during either the learning or the test phase.

Observational learning phase. Whereas the same causal structure was suggested to all the participants, the learning data were varied between conditions. The learning phase consisted of 50 trials (see Table 1) in a randomized order that implemented the parameters of the $\text{Weak}_{A \rightarrow C \rightarrow D}$ and $\text{Strong}_{A \rightarrow C \rightarrow D}$ conditions, respectively (see Figures 1A and 1B). In the $\text{Weak}_{A \rightarrow C \rightarrow D}$ condition, the causal path $A \rightarrow C \rightarrow D$ consisted of weak probabilistic relations, whereas the alternative causal chain $A \rightarrow B \rightarrow D$ contained strong probabilistic relations. In the alternative condition, $\text{Strong}_{A \rightarrow C \rightarrow D}$, this pattern was reversed. In this condition, the causal path $A \rightarrow C \rightarrow D$ involved strong probabilistic relations, but the backdoor path $A \rightarrow B \rightarrow D$ comprised only weak probabilistic relations. The trials presented information on a computer screen about the states of the four variables, with each trial referring to a different wine cask. Each chemical substance was represented by a circle with a fictitious label (see Figure 2 for a trial example). At the beginning of each trial, all four circles were labeled with question marks, indicating that the variables' states in this wine cask were not yet known. Then temporally ordered infor-

Table 1
Learning Data of Experiments 1 and 2

Data Pattern	Experiment 1 (Causal Strength)		Experiment 2 (Base Rates)	
	$\text{Weak}_{A \rightarrow C \rightarrow D}$	$\text{Strong}_{A \rightarrow C \rightarrow D}$	$A_{\text{high}}C_{\text{low}}$	$A_{\text{low}}C_{\text{high}}$
a. b. c. d	4	7	29	9
a. b. c.¬d	1	1	1	1
a.¬b. c. d	1	10	2	2
a.¬b. c.¬d	5	3	1	1
a. b.¬c. d	11	1	5	4
a. b.¬c.¬d	2	2	1	1
a.¬b.¬c. d	0	0	0	0
a.¬b.¬c.¬d	3	3	0	0
¬a. b. c. d	0	0	0	0
¬a. b. c.¬d	0	0	0	0
¬a.¬b. c. d	0	0	4	20
¬a.¬b. c.¬d	0	0	1	4
¬a. b.¬c. d	0	0	0	0
¬a. b.¬c.¬d	0	0	0	0
¬a.¬b.¬c. d	0	0	0	0
¬a.¬b.¬c.¬d	23	23	16	18
Σ	50	50	60	60

Note—Numbers indicate observed frequencies.

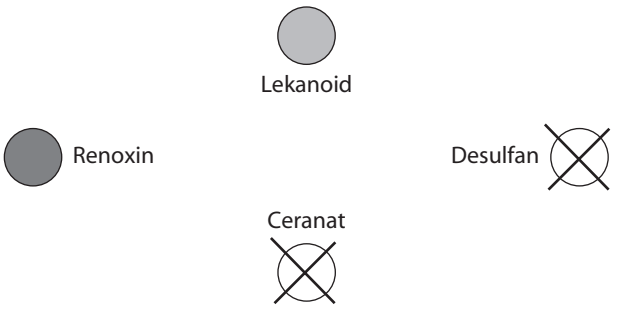

Figure 2. Example of a learning trial in Experiment 1. Events *A* ("Renoxin") and *B* ("Lekanoid") are present, and Events *C* ("Ceranat") and *D* ("Desulfan") are absent (crossed-out circles).

mation about the state of the four variables was given. The presence of a chemical substance was depicted by a colored circle, its absence by a crossed-out circle. Information about the initial Event *A* was given first, and then, simultaneously, information about the presence or absence of *B* and *C* was given. Finally, information about the state of *D* was provided. The interstimulus interval was 1 sec. After the sequence was finished, the information remained for another 2 sec on the screen before the next trial began automatically. The participants passively observed the 50 trials without making overt predictions.

Test phase. Each participant was requested to answer six test questions about observations, hypothetical interventions, and counterfactual interventions. Observational, interventional, and counterfactual questions were grouped into three blocks with two questions each; the order of the blocks was counterbalanced across participants. For the observational questions, the participants were instructed to imagine observing the presence (absence) of Substance *C* in a previously unseen wine cask and then to estimate the probability that Substance *D* is present, too [i.e., they estimated $P(d|c)$ and $P(d|\neg c)$]. For the interventional questions, the learners were asked to imagine that Substance *C* was added to a new wine cask [i.e., $P(d|\text{do } c)$] or that *C* was inhibited from developing by adding a substance called "Anti-*C*" to a cask [i.e., $P(d|\text{do } \neg c)$]. For the counterfactual generative intervention question, the learners were asked to imagine a previously unseen cask in which *C* was observed to be absent, but to imagine that Substance *C* had been added to this very cask [i.e., the learners had to estimate $P(d|\neg c.\ \text{do } c)$]. Conversely, to judge the probability of *D* given a counterfactual prevention of *C* [i.e., to estimate $P(d|c.\ \text{do } \neg c)$], the participants were requested to imagine a wine cask in which *C* was observed to be present, but to imagine that, in this cask, the development of Substance *C* had been prevented by adding "Anti-*C*." All the estimates for the observational and interventional questions were given on a rating scale ranging from 0 = *Substance D is definitely not present* to 100 = *Substance D is definitely present*. For the counterfactual questions, the same scale was used, but labeled with 0 = *Substance D definitely would not have been present* and 100 = *Substance D definitely would have been present*.

Results and Discussion

Table 2 shows the results for the observational, interventional, and counterfactual inference questions, along with the probabilities derived from a causal model analysis (see the Appendix for details). This analysis, which is based on causal models and Bayes net theory, provides the normative solution for the requested causal queries. The analysis derives parameter estimates from the observed data and takes into account the conceptual differences between observations, hypothetical interventions, and counterfactual interventions. To examine people's causal judgments, we used two approaches. First, we compared their estimates with predictions from three different models that implement different assumptions as to how people respond to the different causal queries. Second, we conducted a number of focused statistical tests to further examine how people's judgments vary within and between conditions.

The first model we compared with the human data is based on a *full-fledged causal model analysis*. This model assumes that the parameters of the causal model are estimated from the learning data, assuming that these data accurately reflect the actual parameters. The model also assumes that people differentiate between observations, hypothetical interventions, and counterfactual interventions and use their causal model representations to respond appropriately to these queries. For example, depending on the causal model's parameters, this model may entail differences between the merely observed presence of *C*, the hypothetical generation of *C*, and the counterfactual generation of *C*, which requires a belief update prior to estimating the effects of the intervention. Hence, $P(d|c) \neq P(d|\text{do } c) \neq P(d|\neg c.\ \text{do } c)$. These probabilities are derived in accordance with the computations outlined in the Appendix (see Pearl, 2000).

The second model, the *observation-versus-intervention model*, again implements the idea that people estimate the

Table 2
Mean Probability Judgments for the Causal Inference Questions in Experiment 1 ($N = 36$)

	Observations		Hypothetical Interventions		Counterfactual Interventions	
Causal Strength	$P(d\|c)$	$P(d\|\neg c)$	$P(d\|\text{do } c)$	$P(d\|\text{do } \neg c)$	$P(d\|\neg c.\ \text{do } c)$	$P(d\|c.\ \text{do } \neg c)$
$\text{Weak}_{A\rightarrow C\rightarrow D}$						
Causal model	59	23	39	30	34	56
Experiment						
M	50.56	35.56	41.11	35.56	45.00	40.00
SD	25.55	14.23	24.47	14.23	24.07	22.49
$\text{Strong}_{A\rightarrow C\rightarrow D}$						
Causal model	81	3	79	7	78	14
Experiment						
M	63.33	18.47	55.97	18.47	58.89	19.03
SD	22.49	14.98	25.97	15.75	22.98	19.31

Note—See the Appendix for the derivation of the causal model predictions. All judgments were made on a 0–100 scale; the probabilities derived from the causal model analyses were mapped to this scale by multiplying them by 100.

parameters of the causal model from the data. In contrast to the previous model, it assumes that people treat all interventions equally—that is, as interventions that change the status of a certain variable having implications only for events that are affected by this variable. Research on counterfactual reasoning points in this direction, since people sometimes mentally negate a candidate cause in order to assess whether this particular event was causally responsible for the occurrence of a target outcome (see Spellman & Mandel, 1999, for an overview). Whereas, in this case, peoples' counterfactual thinking focuses on the predictive link from cause to effect, in our experiments particularly the diagnostic evidence provided by the factual state of the event was crucial for correctly assessing the implications of a counterfactual intervention. Thus, this second model recognizes that active interventions may have different implications than passive observations, but it treats counterfactual interventions in the same manner as hypothetical interventions. In other words, regarding counterfactual interventions, this model ignores the available evidence (i.e., the factually observed state of a variable) and focuses only on the counterfactual state of the variable. This account, for example, would predict that $P(d \mid c) \neq P(d \mid \text{do } c)$ but that $P(d \mid \text{do } c) = P(d \mid \neg c.\ \text{do } c)$.

Finally, we also implemented a model that does not distinguish between observations and interventions at all. This *observation-only model* proposes that people merely encode the observed probabilistic relations among the events. These are the probabilities that associative theories would also be sensitive to. Traditional Bayes net theories that are sensitive only to probabilistic, but not causal, relations also do not differentiate between observations and interventions (Pearl, 1988). Queries about probabilities are assumed to be answered on the basis of observed conditional probabilities. Thus, this model entails no differences in probability estimates for different causal queries—that is, $P(d \mid c) = P(d \mid \text{do } c) = P(d \mid \neg c.\ \text{do } c)$.[1]

To clarify the differences between these implementations, consider the causal model predictions for the $\text{Weak}_{A \rightarrow C \rightarrow D}$ condition in Table 2. These six values are derived from the full causal model analysis (see the Appendix). As can be seen, certain differences are predicted between observations and hypothetical interventions [e.g., $P(d \mid c) = 59 > P(d \mid \text{do } c) = 39$], as well as between hypothetical and counterfactual interventions [e.g., $P(d \mid \text{do } c) = 39 > P(d \mid \neg c.\ \text{do } c) = 34$]. Whereas the full causal model analysis entails different predictions for the hypothetical and counterfactual generation of C, the observation-versus-intervention model treats counterfactual interventions just as it treats hypothetical interventions. Thus, this model entails that $P(d \mid \text{do } c) = P(d \mid \neg c.\ \text{do } c) = 39$. Finally, the observation-only model does not distinguish between observations and interventions and, therefore, predicts that $P(d \mid c) = P(d \mid \text{do } c) = P(d \mid \neg c.\ \text{do } c) = 59$.

To evaluate people's probability estimates, the models' predictions are compared with participants' responses to the corresponding test questions. In addition to computing the overall correlation between model predictions and data, we also report root-mean squared errors (RMSEs). This measure provides additional information on how strongly the models' predictions deviate from the human data (see Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). Finally, to examine the relative fits of the models, we also report likelihood ratios (see Glover & Dixon, 2004). These analyses are based on comparing the models' residual variations (i.e., sum-squared errors [SSEs]).

Note that the three models do not have any free parameters that are fitted to the data. We computed model fits separately for each condition, as well as the overall correlation across conditions (Table 3). Generally, the separate analyses are more informative, because there are a number of conditions in which the models make very similar predictions. Likelihood ratios are computed across conditions.

The $\text{Weak}_{A \rightarrow C \rightarrow D}$ condition is the critical condition in Experiment 1 since, here, the normative probabilities entail differences between observations and hypothetical and counterfactual interventions. A first inspection of the descriptive data indicates that the participants tended to respond differently to observational and hypothetical intervention questions, but similar estimates were obtained for the hypothetical and counterfactual intervention questions. This impression is corroborated by the model fits (see Table 3). The results show that the observation-versus-intervention model had the highest fit ($r = .94$) and the smallest RMSE. The observation-only model ($r = .81$) fits slightly better than the predictions derived from the full causal model analysis ($r = .73$). However, the latter had a smaller RMSE, which indicates that its predictions deviated not as strongly as the predictions entailed by the observation-only model. The $\text{Strong}_{A \rightarrow C \rightarrow D}$ condition served as a control condition in which no differences were expected between observational and interventional judgments. In accordance with this prediction, all three models fit equally well for this condition (see Table 3). However, both the full causal model analysis and the observation-versus-intervention model had a smaller RMSE than did the observation-only model. We also computed the overall model fits across conditions, although the fact that the models make very similar predictions for the control group (the $\text{Strong}_{A \rightarrow C \rightarrow D}$ condition) renders this analysis less informative. The overall model shows that the best account for the participants' estimates is provided by the observation-versus-intervention model, which had the highest fit and the smallest RMSE. The full causal model and the observation-only model had identical fits, but the observation-only model had a higher error.

To further evaluate the three models, we also computed likelihood ratios, which indicate the relative fit of the models. Briefly, the value of the likelihood ratio gives a measure of how much more likely the data is under one model than under the other (for details, see Glover & Dixon, 2004). Table 4 shows the likelihood ratios for the pairwise comparisons of the three models, based on sum-squared errors over all 12 judgments obtained in Experiment 1 (Table 1). As can be seen from the table, both the full-fledged causal model analysis and the observation-versus-intervention model are clearly superior to the observation-only model ($\lambda = 11.82$ and $\lambda = 14.89$, respectively). However, the likelihood ratio for the comparison of these two models indicates that the observation-versus-intervention model can explain only slightly more variance than can the full-fledged causal model ($\lambda = 1.26$).

Table 3
Model Fits for Experiments 1 and 2: Correlations Between Predicted and Observed Means (With RMSEs)

		Full Causal Model Analysis		Observation-Versus-Intervention Model		Observation-Only Model	
	Condition	*r*	RMSE	*r*	RMSE	*r*	RMSE
Experiment 1							
Predictive judgments	$\text{Weak}_{A\to C\to D}$	.73	9.79	.94	7.46	.81	13.70
	$\text{Strong}_{A\to C\to D}$	.99	16.35	1.00	17.14	.99	18.99
	Overall	.96	13.48	.98	13.22	.96	16.55
Experiment 2							
Diagnostic judgments	$A_{\text{high}}C_{\text{low}}$	.75	27.15	.53	25.37	−.24	35.11
	$A_{\text{low}}C_{\text{high}}$	.30	6.02	.11	5.08	−.39	8.67
	Overall	.84	19.66	.79	18.30	.28	25.57
Predictive judgments	$A_{\text{high}}C_{\text{low}}$	.70	17.35	.82	15.59	.72	25.98
	$A_{\text{low}}C_{\text{high}}$	.99	9.50	.99	10.05	.99	11.12
	Overall	.90	13.99	.93	13.12	.85	19.98

Note—The values for the causal model analysis (left columns) were computed as outlined in the Appendix. The two other models can be considered as simplified cases of the full-fledged causal model analysis. The *observation-versus-intervention model* (middle columns) entails different predictions for observations and interventions, but no difference is made between hypothetical and counterfactual interventions. For the *observation-only model* (right columns), it is assumed that learners do not differentiate between observations, hypothetical interventions, and counterfactual interventions. See the text for details.

Statistical analyses. Finally, we conducted a number of within- and between-subjects comparisons. The within-subjects comparisons focus on the distinction between hypothetical observations and hypothetical interventions and the differences between hypothetical and counterfactual interventions. In line with the high fit of the observation-versus-intervention model, the within-subjects analyses of the $\text{Weak}_{A\to C\to D}$ condition revealed no reliable differences between the hypothetical and the counterfactual intervention questions. An ANOVA with state of *C* (present vs. absent) and type of question (hypothetical vs. counterfactual intervention) as within-subjects factors did not yield the interaction effect entailed by the probabilities derived from the causal model analysis ($F < 1$). However, the predicted interaction between observations and hypothetical interventions in the $\text{Weak}_{A\to C\to D}$ condition was also surprisingly small and only approached significance [$F(1,17) = 2.46$, $p = .14$]. We therefore conducted an additional analysis based on the pairwise comparisons of the observation and hypothetical intervention judgments for the presence and absence of *C* [i.e., $P(d\,|\,c)$ vs. $P(d\,|\,\neg c)$ and $P(d\,|\,\text{do}\ c)$ vs. $P(d\,|\,\text{do}\ \neg c)$] to explore the trends in the data. Consistent with the prediction that observations and interventions should be treated differently, in the $\text{Weak}_{A\to C\to D}$ condition the observational judgments differed significantly from each other [$t(17) = 2.37$, $p = .03$], whereas no reliable difference was obtained for the intervention questions [$t(17) = 0.94$, $p = .36$]. By contrast, in the $\text{Strong}_{A\to C\to D}$ condition, not only did the observational judgments differ from each other [i.e., $P(d\,|\,c)$ vs. $P(d\,|\,\neg c)$; $t(17) = 8.55$,

Table 4
Comparisons of the Model Fits in Experiments 1 and 2: Likelihood Ratio λ Indexing the Relative Fits of Two Models, on the Basis of Sum-Squared Errors (*SSE*s)

	Experiment 1 (Predictive Judgments)		Experiment 2 (Predictive Judgments)		Experiment 2 (Diagnostic Judgments)	
	SSE	λ	*SSE*	λ	*SSE*	λ
Observation-only model	3,289		7,848		4,790	
vs.	vs.	11.82	vs.	23.42	vs.	72.29
Full-fledged causal model	2,179		4,640		2,347	
Observation-only	3,289		7,848		4,790	
vs.	vs.	14.89	vs.	55.65	vs.	156.22
Observation-versus-intervention model	2,097		4,017		2,064	
Full-fledged causal model	2,179		4,640		2,374	
vs.	vs.	1.26	vs.	2.38	vs.	2.16
Observation-versus-intervention model	2,097		4,017		2,064	

Note—A likelihood ratio of $\lambda > 1$ indicates a better fit for the model named last in the left column. See the text for details. The likelihood ratio λ is computed according to the following formula: $\lambda = (SSE_1 / SSE_2)^{n/2}$, where SSE_1 and SSE_2 are the sum-squared deviations from the means predicted by Model 1 and Model 2, respectively, and *n* is the number of observations (see Glover & Dixon, 2004, for details).

$p < .001$], but also the interventional judgments [i.e., $P(d|c)$ vs. $P(d|\text{do } \neg c)$; $t(17) = 5.64$, $p < .001$].

The between-subjects comparisons directly focused on the question as to how variations in the learning data affected the participants' causal judgments. We conducted separate analyses for observational, interventional, and counterfactual questions. An ANOVA for the observational questions with state of C (present vs. absent) as the within-subjects factor and parameterization ($\text{Weak}_{A\rightarrow C\rightarrow D}$ vs. $\text{Strong}_{A\rightarrow C\rightarrow D}$) as the between-subjects factor showed the expected interaction effect [$F(1,34) = 13.19$, $p < .001$]. Reliable interaction effects were also obtained for the hypothetical intervention questions [$F(1,34) = 12.93$, $p = .001$] and the counterfactual intervention questions [$F(1,34) = 10.86$, $p < .01$]. These analyses show that the learners' responses to the observation questions, as well as to the hypothetical and counterfactual intervention questions, were strongly influenced by the learning data. These findings corroborate the hypothesis that people use the available data to estimate the parameters of the causal model, which then are used to answer causal queries. As a consequence, different parameterizations of the same causal structure led to different response patterns. However, the model fits as well as the results of the within-subjects comparisons also suggest that the participants failed to adequately differentiate between counterfactual and hypothetical interventions. The observation-versus-intervention model, which entails identical judgments for hypothetical and counterfactual intervention questions, best accounted for the human data. The rank order of the participants' intervention judgments [$P(d|\text{do } c) > P(d|\text{do } \neg c)$ and $P(d|\neg c.\ \text{do } c) > P(d|c.\ \text{do } \neg c)$] indicates that the participants tended to neglect the implications of the factual observation for the instantiation of the backdoor path and, instead, treated the counterfactual queries like hypothetical interventions.

EXPERIMENT 2

The results of Experiment 1 showed that manipulations of the model's causal strength parameters strongly affected people's judgments. To further investigate the role of the learning data and of alternative parameterizations of a causal model, in Experiment 2, we manipulated base rate information while keeping causal strength constant. Thus, the rationale of the experimental setup (i.e., identical causal structures, different learning data) follows that of Experiment 1. Whereas the previous experiment showed that learners took into account variations of causal strength, the goal of Experiment 2 was to investigate whether learners are sensitive to base rate information when making different types of causal judgments. Base rates not only are relevant for observational inferences modeled by standard probability calculus (e.g., Bayes's theorem), but also need to be considered when deriving interventional probabilities. For example, when the causal path $A\rightarrow C\rightarrow D$ is broken by preventing C from occurring, the instantiation of the alternative causal chain $A\rightarrow B\rightarrow D$ depends not only on the strength of the involved causal links, but also on the probability of the initial Event A. If A is frequent (i.e., has a high base rate), it is more likely that D will be generated via the alternative causal chain than when A is rare (i.e., has a low base rate). To examine these predictions in more detail, we here investigated both diagnostic causal reasoning from C to A and predictive causal reasoning from C to D. Causal model theory predicts that both types of inferences should be affected by variations in A's base rate.

Method

Participants and Design

Forty-eight undergraduate students from the University of Göttingen participated for course credit; none of them had taken part in Experiment 1. The *learning data* factor was varied between conditions; the *type of inference* and *presence versus absence of C* factors were varied within subjects. All the participants were randomly assigned to one of the two conditions.

Procedure and Materials

Causal model instruction. For this study, the causal structure shown in Figure 1 was instantiated as a medieval communication system transmitting signal fires between four watch towers (Figure 3). Furthermore, the participants were told that Towers A and C were close to a particular border they watch. If either of these two towers spots enemy troops, a fire is lit, and the signal is transmitted to Tower D (i.e., there are two possible hidden causes that can initiate the signal transmission). This allows for the manipulation of the probability with which Events A and C occur (i.e., their base rates). As in Experiment 1, the participants were shown a graph of the causal model to illustrate the direction of the signal transmission and to exclude any bidirectional links. The participants were then instructed to learn "how well the communication system works" from observing the watch tower system on several days. No information about the model's parameters was given. It was pointed out, however, that the relations might be probabilistic—for example, because bad weather might prevent a tower's guards from detecting a signal. The kind of questions the participants would have to answer after the learning phase was not mentioned until the test phase.

Observational learning phase. The learning phase consisted of 60 trials presenting information about the states of the four variables on a computer screen, with each trial referring to a new day on which the communication system was observed. Table 1 shows the learning input that implements the parameters of the causal models shown in Figures 1C and 1D. In the $A_{\text{high}}C_{\text{low}}$ condition, the initial Event A has a high base rate [$P(a) = .65$], but the probability of C's occurring in

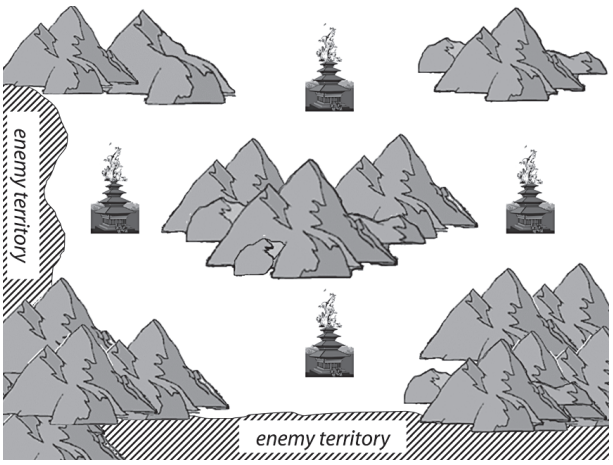

Figure 3. Example of a learning trial in Experiment 2. In this trial, all events are present (i.e., there is a signal fire on each tower). "Enemy territory" represents hidden causes that might initiate the fire signal transmission among the towers.

the absence of A is low [$P(c|\neg a) = .24$]. This pattern is reversed in the $A_{low}C_{high}$ condition. In this condition, the initial event's base rate is rather low [$P(a) = .3$], but C often occurs when A is absent [i.e., $P(c|\neg a) = .57$]. Raising and lowering the parameter $P(c|\neg a)$ inversely proportional to the base rate of the initial Event A allows for keeping the frequency of D's being present approximately equal across the two conditions [$P(d) = .67$ and $P(d) = .58$ in the $A_{high}C_{low}$ and $A_{low}C_{high}$ conditions, respectively]. In this experiment, the states of all four events were displayed simultaneously. Learners could continue at their own pace, but they were not allowed to revisit a trial.

Test phase. Subsequent to the observational learning phase, the participants were again asked three types of causal inference questions: observational, interventional, and counterfactual questions. For the observational questions, the learners were requested to imagine a new day on which a signal fire on Tower C was observed and then to estimate the probability of a fire on Tower A. Subsequently, they were asked to estimate the probability of a fire on Tower D. The same two questions were asked for the case in which no fire on Tower C was observed. Thus, these questions required an estimation of the conditional probabilities $P(a|c)$, $P(a|\neg c)$, $P(d|c)$, and $P(d|\neg c)$. The generative interventional question stated that lightning had struck the tower and lit the signal fire. The inhibitory interventional question stated that the tower's guards had forgotten to collect new firewood, and therefore, no fire could be lit that day. Thus, the participants had to estimate $P(a|\text{do } c)$, $P(a|\text{do } \neg c)$, $P(d|\text{do } c)$, and $P(d|\text{do } \neg c)$. The questions referring to the counterfactual generation of C first requested that the learners should assume that no fire was observed this day on Tower C (factual observation of C's being absent), but to imagine that, on this very day, lightning had caused a signal fire (counterfactual generation). Thus, these questions referred to the counterfactual probabilities $P(a|\neg c.\ \text{do } c)$ and $P(d|\neg c.\ \text{do } c)$, respectively. Conversely, the counterfactual inhibitory questions stated that a signal fire was observed to be present on Tower C. The learners were then asked to imagine that the guards had forgotten to collect new firewood that very day and to estimate the probability of a fire on Towers A and D, respectively [i.e., the participants estimated $P(a|c.\ \text{do } \neg c)$ and $P(d|c.\ \text{do } \neg c)$].

The estimates for the observational and interventional questions were given on a rating scale ranging from 0 = *There definitely is no signal fire on Tower A [D]* to 100 = *There definitely is a signal fire on Tower A [D]*. For the counterfactual questions, the same scale was used, but labeled with 0 = *There definitely would not have been a signal fire on Tower A [D]* and 100 = *There definitely would have been a signal fire on Tower A [D]*. Interventional, observational, and counterfactual questions were blocked; the order of blocks was counterbalanced.

Results and Discussion

As in Experiment 1 we analyzed the results by comparing the three models' predictions with the human data. We also conducted a number of planned within- and between-subjects comparisons that focused on the predicted differences. For the sake of clarity, we will report the results for the diagnostic and predictive inference questions separately.

Diagnostic Inferences

The results for the diagnostic inference questions are shown in Table 5; the model fits are depicted in the middle part of Table 3. Regarding the $A_{high}C_{low}$ condition, the predictions derived from the full causal model analysis accounted best for the data ($r = .75$), followed by the observation-versus-intervention model ($r = .53$); but the latter had a slightly smaller RMSE. The observation-only model performed worst; this model's predictions correlated negatively with the participants' responses ($r = -.24$). A similar pattern was obtained for the $A_{low}C_{high}$ condition, although in this condition, all the models had a rather poor fit ($r = .30$, .11, and $-.39$, respectively). A closer inspection of the data revealed the reasons for these poor fits. To our surprise, in both conditions, the participants gave higher estimates for an inhibitory intervention (i.e., do $\neg c$) than for a generative intervention (i.e., do c). It turns out that these effects are due mainly to a small number of participants who strongly underestimated the probability $P(a|\text{do } c)$, although it is not clear to us why this was the case. Finally, we computed model fits across conditions. The overall fit showed that the full causal model analysis fitted best ($r = .84$), closely followed by the observation-versus-intervention model ($r = .79$). The observation-only model yielded the poorest overall fit ($r = .28$) and also had a much higher RMSE than did the other two models.

Finally, we again computed likelihood ratios to determine the relative model fit (Table 4). The ratio was derived from the models' sum-squared errors computed over the 12 diagnostic judgments. As in Experiment 1, both the full-fledged causal model analysis and the observation-versus-intervention model fit the data much better than did the observation-only model ($\lambda = 23.42$ and 55.65, respectively). The comparison of the observation-versus-intervention model with the full-fledged analysis shows that the data are more likely given the observation-versus-intervention model than given the full-fledged causal

Table 5
Mean Probability Judgments for Diagnostic Inference Questions in Experiment 2 ($N = 48$)

	Observations		Hypothetical Interventions		Counterfactual Interventions	
Base Rates	$P(a\|c)$	$P(a\|\neg c)$	$P(a\|\text{do } c)$	$P(a\|\text{do } \neg c)$	$P(a\|\neg c.\ \text{do } c)$	$P(a\|c.\ \text{do } \neg c)$
$A_{high}C_{low}$						
Causal model	87	27	65	65	27	87
Experiment						
M	50.83	38.75	36.67	47.08	35.42	45.00
SD	20.41	18.25	23.90	16.81	25.36	24.67
$A_{low}C_{high}$						
Causal model	35	22	30	30	22	35
Experiment						
M	35.42	33.33	28.33	34.58	29.17	31.25
SD	18.41	14.65	16.59	16.15	23.20	15.97

Note—See the Appendix for the derivation of the causal model predictions. All judgments were made on a 0–100 scale; the probabilities derived from the causal model analyses were mapped to this scale by multiplying them by 100.

model analysis ($\lambda = 2.38$). However, the magnitude of the likelihood ratio provides only moderate evidence in favor of the observation-versus-intervention model.

Statistical analyses. The within-subjects comparisons of the observational and interventional probabilities indicated that the participants clearly distinguished between passive observations and active interventions. In the $A_{high}C_{low}$ condition, the causal model's parameters entailed quite large differences between the observational and interventional probabilities. Accordingly, an ANOVA with state of C (present vs. absent) and type of question (observation vs. hypothetical intervention) as within-subjects factors yielded a significant interaction effect [$F(1,23) = 14.37$, $p < .001$]. An interaction effect was also obtained for the $A_{low}C_{high}$ condition [$F(1,23) = 5.87$, $p = .02$]. As in Experiment 1, we also examined how people responded to the counterfactual intervention questions and contrasted them with the estimates obtained for the hypothetical intervention questions. At variance with the predictions derived from the full causal model analysis, but consistent with Experiment 1, in none of the two conditions was a reliable difference obtained between the hypothetical and counterfactual intervention questions. Thus, as in the previous experiment, the participants did not differentiate between these two types of interventional questions.

We also examined the participants' sensitivity to the learning input by contrasting their causal judgments across the two parameterizations. An ANOVA for the observation judgments, with state of C (present vs. absent) as the within-subjects factor and parameterization ($A_{high}C_{low}$ vs. $A_{low}C_{high}$) as the between-subjects factor, revealed a main effect of parameterization [$F(1,46) = 5.97$, $p = .01$], but there was only a statistical tendency for the interaction between parameterization and state of C [$F(1,46) = 2.79$, $p = .10$]. We therefore conducted an additional test focusing on the participants' judgments regarding $P(a \mid c)$, for which the causal model analysis entails a large difference [only a negligible difference is entailed for $P(a \mid \neg c)$]. Consistent with this prediction, a significant difference was obtained [$t(46) = 2.75$, $p < .01$]. For the hypothetical intervention questions, the normative probabilities entailed only a main effect of condition, but no interaction. Consistent with these predictions, the analysis revealed a main effect of condition [$F(1,46) = 5.03$, $p = .03$] but no interaction ($F < 1$). Finally, the analysis of the counterfactual judgments revealed only a weak effect of the between-subjects variable [$F(1,46) = 2.94$, $p = .09$]; the expected interaction did not prove significant [$F(1,46) = 1.63$, $p = .21$]. However, the participants' judgments for the counterfactual inhibition of C [i.e., $P(a \mid c.\ \text{do}\ \neg c)$] were affected by the learning data [$t(46) = 2.29$, $p < .05$].

Predictive Inferences

Table 6 shows the results for the predictive inference questions, along with the probabilities derived from the causal model analysis; the lower part of Table 3 shows the model fits. The model fits of the $A_{high}C_{low}$ condition show that the observation-versus-intervention model had the highest fit ($r = .82$) and smallest RMSEs. The full causal model and the observation-only model had approximately equal fits ($r = .70$ and .72, respectively), but the very high RMSE for the observation-only model refutes this account as a descriptive model of people's causal judgments. Since the alternative $A_{low}C_{high}$ condition does not entail many differences between observations and interventions, all three models here make very similar predictions, resulting in equally high model fits ($r = .99$ for all the models). The final analysis concerns the overall model fits across conditions. These analyses support again the observation-versus-intervention model, which had the highest fit ($r = .93$) and the smallest RMSE, closely followed by the full causal model analysis ($r = .90$). The observation-only model had the lowest fit ($r = .85$) and a substantially higher RMSE than did the two other models.

The likelihood ratios computed to assess the relative fit of the models corroborated the previous analyses (Table 4). As before, the observation-only model has a substantially worse fit than did the other two models ($\lambda = 72.29$ and 156.22, respectively). The observation-versus-intervention model matched the observed data slightly better than did the full-fledged analysis ($\lambda = 2.16$).

Statistical analyses. As before, we additionally conducted focused within- and between-subjects comparisons. The comparison of the observation and hypothetical inter-

Table 6
Mean Probability Judgments for Predictive Inference Questions in Experiment 2 ($N = 48$)

	Observations		Hypothetical Interventions		Counterfactual Interventions	
Base Rates	$P(d \mid c)$	$P(d \mid \neg c)$	$P(d \mid \text{do}\ c)$	$P(d \mid \text{do}\ \neg c)$	$P(d \mid \neg c.\ \text{do}\ c)$	$P(d \mid c.\ \text{do}\ \neg c)$
$A_{high}C_{low}$						
Causal model	92	23	88	50	80	67
Experiment						
M	80.00	38.75	76.67	47.08	70.83	42.08
SD	15.88	23.46	20.78	19.89	25.18	24.84
$A_{low}C_{high}$						
Causal model	84	17	84	20	83	23
Experiment						
M	80.42	25.83	72.50	31.67	70.42	27.92
SD	14.89	18.40	15.95	20.36	18.99	17.44

Note—See the Appendix for the derivation of the causal model predictions. All judgments were made on a 0–100 scale; the probabilities derived from the causal model analyses were mapped to this scale by multiplying them by 100.

vention questions revealed a marginally significant interaction term, both in the $A_{low}C_{high}$ condition [$F(1,23) = 3.99$, $p = .06$] and in the $A_{high}C_{low}$ condition [$F(1,23) = 3.89$, $p = .06$]. In line with the obtained model fits, these findings suggest that the participants responded differently to the observation and hypothetical intervention questions. As in the previous analyses, the comparisons between the hypothetical and counterfactual intervention questions did not show the expected interaction effects. This pattern of findings indicates that the participants again did not adequately differentiate between these two types of interventional inferences.

Next, we contrasted the participants' responses across conditions (i.e., $A_{low}C_{high}$ vs. $A_{high}C_{low}$). For the observation questions, the probabilities derived from the causal model analysis do not entail many differences for the two parameterizations. Accordingly, there was only a statistical tendency for the between-subjects factor [$F(1,46) = 2.8$, $p = .10$], as well as for the interaction contrast (i.e., parameterization $\times$ state of C) [$F(1,23) = 3.06$, $p = .09$]. In contrast to the observation questions, the causal model analysis predicts a dissociation between the two parameterizations for the hypothetical intervention questions, which was actually found. The results of the ANOVA for these questions yielded a significant main effect of condition [$F(1,46) = 5.58$, $p = .02$], but only a weak interaction between parameterization and state of C was found [$F(1,46) = 2.26$, $p = .14$]. However, due to the strong probabilistic link between Events C and D the probabilities derived from the causal model analysis actually entailed only a substantial difference for the hypothetical prevention of C [$P(d \mid \text{do } \neg c)$], which we obtained [$t(46) = 2.65$, $p = .01$]. The analysis of the counterfactual intervention questions revealed a similar, although somewhat weaker, pattern. There was only a statistical tendency for the between-subjects factor [$F(1,46) = 2.81$, $p = .10$], as well as for the interaction contrast [$F(1,46) = 2.25$, $p = .14$]. As for the hypothetical intervention questions, participants' judgments for the preventive intervention in C [$P(d \mid c.\ \text{do } \neg c)$] varied between conditions [$t(46) = 2.29$, $p = .03$].

Taken together, these analyses corroborate the results of the model fits. Although some participants seemed to have had problems with correctly assessing base rate information, the general response pattern supported the hypothesis that the learners distinguished between hypothetical observations and hypothetical interventions and took into account the parameters of the causal model. However, no reliable differences were obtained between hypothetical and counterfactual intervention questions.

GENERAL DISCUSSION

The capacity to derive interventional predictions from observational knowledge is a touchstone of true causal reasoning, because it requires going beyond the mere representation of observed patterns of covariations. In line with the previous findings of Meder et al. (2008), the results of the present set of experiments show that the capacity to predict the consequences of novel interventions from observational knowledge is not limited to tasks in which learners are provided with lists of aggregated data (Waldmann & Hagmayer, 2005) or mere descriptions of causal structures (Sloman & Lagnado, 2005). In particular, the present findings demonstrate that both the structure of the causal model *and* its parameters are taken into account. Experiment 1 showed how the strength of the causal relations within a given causal model affected learners' judgments, whereas Experiment 2 demonstrated that participants were also sensitive to base rate information when deriving interventional predictions. Thus, both experiments support our hypothesis that people learn about the parameters of causal models during observational trial-by-trial learning and, in turn, later use these parameters to derive inferences about the consequences of situations that they have never experienced before. These findings refute the idea that such causal inferences are driven only by causal structure. Rather, the results show that causal judgments vary systematically in accordance with a causal model's parameters, which are gleaned from trial-by-trial observations.

Although the results provide strong evidence that learners were sensitive to the parameters of the initially suggested causal models, the quantitative estimates sometimes deviated from the theoretically derived predictions. For example, the participants often tended to underestimate high probabilities and to overestimate low probabilities. Moreover, they seemed to differentiate only between probabilities that differed by at least .2. There are several possible reasons for this finding. One reason might be the limited number of learning trials (50 and 60 trials for Experiments 1 and 2, respectively). The participants may have been influenced by the uncertainties of the parameter values and, therefore, exhibited regression tendencies in their judgments. Whereas we used point estimates of probabilities to derive the causal model predictions, an alternative way would be to use parameter *distributions*, as suggested by a fully Bayesian account (for more details on Bayesian inferences in causal induction, see Griffiths & Tenenbaum, 2005; Lu et al., 2008). If one assumes a flat prior probability distribution (i.e., people have no specific prior assumption about the parameters), the resulting posterior probability distribution should be fairly flat (i.e., have a high variance). As a consequence, learners' ratings should regress toward the mean, which resembles the obtained pattern of results. However, we consider the observed regression tendency to be not too problematic, since none of the presently defended theories of causal learning (including associative theories) predict that people are capable of exactly estimating statistics. Therefore, following previous research in the area of causal and associative learning, we focused on ordinal predictions.

The second goal of this study was to investigate whether people differentiate between hypothetical interventions (i.e., interventions they have not taken or seen before) and counterfactual interventions, which hypothetically change a state of the world known to be present. In previous research on causal reasoning, counterfactual thinking has often been discussed as a test of causality in which the candidate cause is mentally negated and the probability of the effect given this "undoing" is assessed (see Spellman & Mandel, 1999, for an overview). The focus of the present experiments, by contrast, was not on counterfactual thinking as a cue to cau-

sation but on the difference between reasoning about the outcomes of hypothetical and counterfactual interventions. Formally, the difference between the two types of interventional inferences is that predictions about counterfactual interventions first require an update of some of the causal model's probabilities in accordance with the factually observed state, which might entail diverging predictions for the outcomes of hypothetical and counterfactual interventions (for instance, because the probability with which the backdoor path is instantiated differs between the two situations). Although both experiments were designed to yield different implications between these two types of queries, the participants responded to both types of questions in a quite similar fashion. Instead of combining observational and interventional inferences to derive the counterfactual predictions, they treated the counterfactual queries rather as if they referred to hypothetical actions. This was especially salient in Experiment 1, in which the implied interventional probabilities had a rank order different from that for the implied counterfactual probabilities. The response patterns obtained suggest that the learners neglected the implications of the factually observed state of Event *C* for the instantiation of the alternative causal pathway. The model fits also indicate that the participants tended to respond to the different types of interventions in a similar fashion.

Nevertheless, we advise caution in interpreting our results. The difference between the interventional and counterfactual probabilities may have simply been too small to be detected by the participants, given such a restricted set of data. Another reason may have been the complexity of the causal model (probabilistic relations and alternative causal pathways). Finally, although the model fits generally favored the observation-versus-intervention model over the full-fledged model, the likelihood ratios also suggest a cautious interpretation of the results. Although the observation-versus-intervention model was found to be superior to the full-fledged causal model analysis, the magnitude of the likelihood ratios does not provide clear-cut evidence in favor of one model over the other. Future research needs to investigate whether people are generally insensitive to the distinction between hypothetical and counterfactual interventions (at least as modeled in the causal Bayes net framework) in more simple tasks as well.

Theoretical Implications

The results of the experiments bear on several theoretical models of learning and reasoning. We will focus briefly on three prominent theories of causal learning and reasoning: associative accounts, causal Bayes nets, and causal model theory. Associative theories of causal cognition are weakened by our findings. These theories do not have the representational power to express the differences between observations and interventions. Therefore, they cannot differentiate between the relations observed during learning and the causal consequences that would result from an intervention on the system. In learning tasks with only observational, but not instrumental, learning input, either the observed relations have to be used by associative theories to make a prediction for interventions, or no inferences pertaining to novel interventions can be made at all (for detailed discussions, see Waldmann et al., 2008; Waldmann & Hagmayer, 2005). The finding that participants differentiate between observational and interventional predictions after purely observational learning is at variance with both possibilities. The results are also inconsistent with the claim that trial-by-trial learning might operate through different learning mechanisms than does causal reasoning with aggregated data (e.g., Price & Yates, 1995; Shanks, 1991). Our results conform closely to the ones by Waldmann and Hagmayer (2005), who used aggregated data. In our view, the present findings strongly suggest that both aggregated and trial-by-trial learning input may lead to similar causal model representations, which can be flexibly accessed to answer different types of causal queries (see Meder et al., 2008; Waldmann & Hagmayer, 2005; see also Vadillo & Matute, 2007; Vadillo et al., 2005).

Causal Bayes net theories assume that participants use the instruction and the learning data to update probability distributions of possible causal models and of parameters. On the basis of these distributions, inferences are made about novel observations, hypothetical interventions, and counterfactual interventions. The finding that the participants were sensitive to the structure and the parameters of the causal model is clearly consistent with this account. Also, the participants were capable of differentiating between observations and interventions, which is a hallmark prediction of causal Bayes net theories. As we have outlined above, using a "fully" Bayesian account based on priors, likelihoods, and posterior probability distributions might additionally explain why participants' probability estimates deviated from the derived probabilities (which we derived using point estimates). More problematic for Bayes net theories may be the participants' failure to differentiate between hypothetical and counterfactual interventions, a finding that is at odds with causal Bayes net theories. As we have explained in the section on modeling, according to these theories, participants should first update their probability distributions in the light of the factually observed state and then continue to infer the outcomes of an intervention by taking into account the updated model. Future research needs to investigate whether people are generally insensitive to this distinction or whether the deviations are due, rather, to the complexity of the causal model. In our opinion, however, this finding does not invalidate the other predictions regarding the representation of causal relations and the inferences people can make on the basis of parameterized models. We think that it is a valid strategy to decompose complex theories, which initially have been proposed in the context of machine learning, and test localized predictions to decide which parts of the theory are psychologically valid and which parts are not (see Waldmann et al., 2008).

Causal model theory (e.g., Waldmann, 1996) assumes that people first infer the structure of a causal model, using available cues (e.g., the causal model instructed), and then learn about the parameters of the causal model during observational or interventional learning (see Lagnado, Waldmann, Hagmayer, & Sloman, 2007). Thus, our findings that people are sensitive to causal structure and parameters are consistent with this theory as well. In

the present set of experiments, we did not test whether people can also learn complex causal structures from covariational data alone, as postulated by causal Bayes net theories, because we doubt that learners are capable of doing so. In contrast to recent versions of Bayesian theories (Griffiths & Tenenbaum, 2005; Lu et al., 2008), no account has been proposed within causal model theory as to why participants' estimates deviate from the predictions in the way they do. One possibility, which is in accordance with many findings (including the ones reported here) is that the parameters of causal models are classified into broader categories (e.g., no impact, little impact, etc.) with rather fuzzy boundaries. This may also explain why participants are, in general, not sensitive to probability differences below .2, although they clearly take observed probabilities and their implications into account. In addition, like the causal Bayes net approach, causal model theory currently cannot explain why and under which conditions people may fail to differentiate between counterfactual and interventional probabilities. This part of causal Bayes net theories has not been independently developed within causal model theory and, therefore, was simply adopted.

Concluding Remarks

The results of the experiments reported in this article show that people are capable of learning about the parameters of a given causal model underlying an observed pattern of data. They also show that people often use these parameterized models to make inferences about the possible consequences of actions they have not taken before. This capacity is clearly of utmost importance in everyday life because it allows us to integrate knowledge about causal structures, which is often transferred socially (see Einhorn & Hogarth, 1986; Lagnado et al., 2007), and learning from observation, which does not incur the costs and risks of trial-and-error learning. Thus, this capacity not only is a touchstone of true causal reasoning, but also may be one of the few cornerstones of everyday reasoning.

AUTHOR NOTE

The experiments formed part of the doctoral dissertation of the first author (B.M.), which was supervised by the last author (M.R.W.). The research was supported by DFG Grant WA 621/20-1,2; portions of this research were presented at the 2007 Experimental Psychology conference (TeaP) in Trier, Germany. We thank Julia Iwen for collecting the data. We also thank Klaus Oberauer, Tom Beckers, and two anonymous reviewers for insightful comments on previous versions of the manuscript. Address correspondence to B. Meder, who is now at the Max Planck Institute for Human Development, Königin-Luise-Strasse 5, 14195 Berlin, Germany (e-mail: meder@mpib-berlin.mpg.de).

REFERENCES

BECKERS, T., DE HOUWER, J., & MATUTE, H. (EDS.) (2007). *Human contingency learning: Recent trends in research and theory.* Hove, U.K.: Psychology Press.

BLAISDELL, A. P., SAWA, K., LEISING, K. J., & WALDMANN, M. R. (2006). Causal reasoning in rats. *Science*, **311**, 1020-1022.

CHENG, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, **104**, 367-405.

DAWID, A. P. (2002). Influence diagrams for causal modeling and inference. *International Statistical Review*, **70**, 161-189.

DAWID, A. P. (2006). *Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality* (Research Rep. 269). London: University College London, Department of Statistical Sciences.

DE HOUWER, J., BECKERS, T., & VANDORPE, S. (2005). Evidence for the role of higher order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior*, **33**, 239-249.

DICKINSON, A. (2001). Causal learning: An associative analysis. *Quarterly Journal of Experimental Psychology*, **54B**, 3-25.

EINHORN, H. J., & HOGARTH, R. M. (1986). Judging probable cause. *Psychological Bulletin*, **99**, 3-19.

GLOVER, S., & DIXON, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, **11**, 791-806.

GLYMOUR, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences*, **7**, 43-48.

GOPNIK, A., GLYMOUR, C., SOBEL, D. M., SCHULZ, L. E., KUSHNIR, T., & DANKS, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, **111**, 3-32.

GOPNIK, A., & SCHULZ, L. (EDS.) (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.

GRIFFITHS, T. L., & TENENBAUM, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, **51**, 354-384.

LAGNADO, D. A., WALDMANN, M. R., HAGMAYER, Y., & SLOMAN, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154-172). Oxford: Oxford University Press.

LEISING, K. J., WONG, J., WALDMANN, M. R., & BLAISDELL, A. P. (2008). The special status of actions in causal reasoning in rats. *Journal of Experimental Psychology: General*, **137**, 514-527.

LEWIS, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

LU, H., YUILLE, A., LILJEHOLM, M., CHENG, P. W., & HOLYOAK, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, **115**, 955-984.

MEDER, B., HAGMAYER, Y., & WALDMANN, M. R. (2006). Understanding the causal logic of confounds. In R. Sun (Ed.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 579-584). Mahwah, NJ: Erlbaum.

MEDER, B., HAGMAYER, Y., & WALDMANN, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review*, **15**, 75-80.

MEEK, C., & GLYMOUR, C. (1994). Conditioning and intervening. *British Journal for the Philosophy of Science*, **45**, 1001-1021.

PEARL, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.

PEARL, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

PRICE, P. C., & YATES, J. F. (1995). Associative and rule-based accounts of cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1639-1655.

SHANKS, D. R. (1991). On similarities between causal judgments in experienced and described situations. *Psychological Science*, **5**, 341-350.

SHANKS, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology*, **60**, 291-309.

SLOMAN, S. A. (2005). *Causal models. How people think about the world and its alternatives.* New York: Oxford University Press.

SLOMAN, S. A., & HAGMAYER, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences*, **10**, 407-412.

SLOMAN, S. A., & LAGNADO, D. A. (2005). Do we "do"? *Cognitive Science*, **29**, 5-39.

SPELLMAN, B. A., & MANDEL, D. R. (1999). When possibility informs reality: Counterfactual thinking as a cue to causality. *Current Directions in Psychological Science*, **8**, 120-123.

SPIRTES, P., GLYMOUR, C., & SCHEINES, P. (1993). *Causation, prediction, and search.* New York: Springer.

STEYVERS, M., TENENBAUM, J. B., WAGENMAKERS, E.-J., & BLUM, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, **27**, 453-489.

VADILLO, M. A., & MATUTE, H. (2007). Predictions and causal estima-

tions are not supported by the same associative structure. *Quarterly Journal of Experimental Psychology*, **60**, 433-447.

VADILLO, M. A., MILLER, R. R., & MATUTE, H. (2005). Causal and predictive-value judgments, but not predictions, are based on cue–outcome contingency. *Learning & Behavior*, **33**, 172-183.

WALDMANN, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning* (pp. 47-88). San Diego: Academic Press.

WALDMANN, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 53-76.

WALDMANN, M. R., CHENG, P. W., HAGMAYER, Y., & BLAISDELL, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian cognitive science* (pp. 453-484). Oxford: Oxford University Press.

WALDMANN, M. R., & HAGMAYER, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, **82**, 27-58.

WALDMANN, M. R., & HAGMAYER, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 216-227.

WALDMANN, M. R., HAGMAYER, Y., & BLAISDELL, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, **15**, 307-311.

WALDMANN, M. R., & HOLYOAK, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, **121**, 222-236.

WALDMANN, M. R., & WALKER, J. M. (2005). Competence and performance in causal learning. *Learning & Behavior*, **33**, 211-229.

WOODWARD, J. (2003). *Making things happen. A theory of causal explanation*. Oxford: Oxford University Press.

NOTE

1. We also implemented an alternative version of this model, which entails that estimates for the counterfactual queries are based on the factually observed state of the target variable, not on the counterfactually generated state [i.e., $P(d|c) = P(d|\text{do } c) = P(d|c.\ \text{do } \neg c)$]. However, since this model generally provided poorer fits than did the three mentioned models, we here will omit the results.

APPENDIX

The following computations refer to the diamond-shaped causal model shown in Figure 1; capital letters denote variables, lowercase letters denote values of the variables.

Applying the causal Markov condition to the causal model in Figure 1 factorizes the associated joint probability distribution into

$$P(A.B.C.D) = P(A) \cdot P(B|A) \cdot P(C|A) \cdot P(D|B.\ C). \tag{A1}$$

This factorization provides the basis for formalizing observational, interventional, and counterfactual causal inferences. Of particular relevance to the present study is that all the computations involve only parameters that can be derived from observational data.

Modeling Observations

On the basis of structure of the causal model, its parameters, and the decomposed probability distribution, the probabilities implied by the observational data can be computed using standard probability calculus. For example, when C is observed to be present, the probability of A's being present can be computed using Bayes's rule:

$$P(a \mid c) = \frac{P(c \mid a) \cdot P(a)}{P(c \mid a) \cdot P(a) + P(c \mid \neg a) \cdot P(\neg a)} = \frac{P(c \mid a) \cdot P(a)}{P(c)}. \tag{A2}$$

Conversely, if C is observed to be absent, the probability of A's being present is given by

$$P(a \mid \neg c) = \frac{P(\neg c \mid a) \cdot P(a)}{P(\neg c \mid a) \cdot P(a) + P(\neg c \mid \neg a) \cdot P(\neg a)} = \frac{P(\neg c \mid a) \cdot P(a)}{P(\neg c)}. \tag{A3}$$

A more complex example is the prediction of Variable D from observations of Event C:

$$\begin{aligned} P(d \mid c) &= \sum_i P(A_i \mid c) \cdot P(B_i \mid A_i) \cdot P(d \mid B_i.\ c) \\ &= P(a \mid \neg c) \cdot P(b \mid a) \cdot P(d \mid b.\neg c) + P(a \mid \neg c) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.\neg \boldsymbol{c}) + \\ &\quad P(\neg a \mid \neg c) \cdot P(b \mid \neg a) \cdot P(d \mid b.\neg c) + P(\neg a \mid \neg c) \cdot P(\neg \boldsymbol{b} \mid \neg a) \cdot P(d \mid \neg b.\neg c), \end{aligned} \tag{A4}$$

and

$$\begin{aligned} P(d \mid \neg c) &= \sum_i P(A_i \mid \neg c) \cdot P(B_i \mid A_i) \cdot P(d \mid B_i.\neg c) \\ &= P(a \mid \neg c) \cdot P(b \mid a) \cdot P(d \mid b.\neg c) + P(a \mid \neg c) \cdot P(\neg b \mid a) \cdot P(d \mid \neg \boldsymbol{b}.\neg c) + \\ &\quad P(\neg a \mid \neg c) \cdot P(b \mid \neg a) \cdot P(d \mid b.\neg c) + P(\neg a \mid \neg c) \cdot \boldsymbol{P}(\neg b \mid \neg a) \cdot P(d \mid \neg b.\neg c). \end{aligned} \tag{A5}$$

By conditionalizing A on C, these computations take into account that observed states of C are diagnostic for the state of A. The probability of the final effect, D, reflects the influence of both B and C. All probabilities involved in these computations can be derived from the available learning data.

APPENDIX (Continued)

Modeling Hypothetical Interventions

Pearl's (2000) *do-operator* provides a formal means by which to represent the notion of an intervention that fixes the value of the target variable (for alternative notations see Dawid, 2002; Spirtes et al., 1993). For example, an intervention in C renders the event independent of its Markovian parent, Variable A; therefore,

$$P(a \mid \text{do } c) = P(a \mid \text{do } \neg c) = P(a). \tag{A6}$$

The probability of $D = d$, given that C is generated (do c) or inhibited (do $\neg c$) by means of intervention, is given by

$$\begin{aligned} P(d \mid \text{do } c) &= \sum_i P(A_i) \cdot P(B_i \mid A_i) \cdot P(d \mid B_i.\, c) \\ &= P(a) \cdot P(b \mid a) \cdot P(d \mid b.\, c) + P(a) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.\, c) + \\ &\quad P(\neg a) \cdot P(b \mid \neg a) \cdot P(d \mid b.\, c) + P(\neg a) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.\, c) \end{aligned} \tag{A7}$$

and

$$\begin{aligned} P(d \mid \text{do } \neg c) &= \sum_i P(A_i) \cdot P(B_i \mid A_i) \cdot P(d \mid B_i.\, \neg c) \\ &= P(a) \cdot P(b \mid a) \cdot P(d \mid b.\, \neg c) + P(a) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.\, \neg c) + \\ &\quad P(\neg a) \cdot P(b \mid \neg a) \cdot P(d \mid b.\, \neg c) + P(\neg a) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.\, \neg c). \end{aligned} \tag{A8}$$

In contrast to the computations modeling the observational inferences, Variable A is no longer conditionalized on C in these formulas but is replaced by the base rate $P(A_i)$ (cf. Equations A4 and A5). Note that on the right-hand side of the equations, only parameters are involved that can be derived from observational data.

Modeling Counterfactual Interventions

In the Bayes net framework, a counterfactual intervention is defined as an action that alters a factually observed state of the world. Depending on the variable asked for (e.g., cause or effect of the observed event) and the structure of the causal model, the observed state, the counterfactually generated state, or a combination of both determines the counterfactual probability.

For example, the counterfactual inhibition of C entails that C has been observed to be present. Since intervening in an effect variable does not influence its causes, the probability of A's being present, given that C has been observed to be present but is counterfactually removed, is determined by the factual state of C alone (since the intervention would affect only C's effects):

$$P(a \mid c.\ \text{do } \neg c) = P(a \mid c), \tag{A9}$$

and in the case of a counterfactual inhibition of C, which logically entails that C has been observed to be present,

$$P(a \mid \neg c.\ \text{do } c) = P(a \mid \neg c). \tag{A10}$$

Note the difference between modeling hypothetical and counterfactual interventions: Whereas for hypothetical interventions, the probability of C's cause A was given by A's base rate, in the case of a counterfactual intervention in C, the probability of Event A is updated in accordance with the observed state of C.

Updating the probability of C's causes also provides the basis for computing the counterfactual probabilities of the model's other variables. The computation used to derive the counterfactual probability of D, given that C is observed to be present but counterfactually removed [i.e., $P(d \mid c.\ \text{do } \neg c)$], combines observations and interventions. The observed presence of C is used to update the probability of its cause, Event A, but the probability of D (C's effect) then implies the counterfactual absence of C:

$$\begin{aligned} P(d \mid c.\ \text{do } \neg c) &= \sum_i P(A_i \mid c) \cdot P(B_i \mid A_i) \cdot P(d \mid B_i.\, \neg c) \\ &= P(a \mid c) \cdot P(b \mid a) \cdot P(d \mid b.\, \neg c) + P(a \mid c) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.\, \neg c) + \\ &\quad P(\neg a \mid c) \cdot P(b \mid \neg a) \cdot P(d \mid b.\, \neg c) + P(\neg a \mid c) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.\, \neg c). \end{aligned} \tag{A11}$$

Conversely, the probability of D, given that C is observed to be absent but is counterfactually generated, is given by

$$\begin{aligned} P(d \mid \neg c.\ \text{do } c) &= \sum_i P(A_i \mid \neg c) \cdot P(B_i \mid A_i) \cdot P(d \mid B_i.\, c) \\ &= P(a \mid \neg c) \cdot P(b \mid a) \cdot P(d \mid b.\, c) + P(a \mid \neg c) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.\, c) + \\ &\quad P(\neg a \mid \neg c) \cdot P(b \mid \neg a) \cdot P(d \mid b.\, c) + P(\neg a \mid \neg c) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.\, c). \end{aligned} \tag{A12}$$

The probability update of C's cause A makes the instantiation of the backdoor path more or less likely (as compared with the computations for hypothetical interventions, which involve A's base rate). Again, only parameters are involved that can be estimated from the data.

(Manuscript received May 27, 2008;
revision accepted for publication December 3, 2008.)