# Causal Reasoning through Intervention

York Hagmayer[1], Steven A. Sloman[2], David A. Lagnado[3], and Michael R. Waldmann[1]

[1]University of Göttingen, Germany, [2]Brown University, Providence, RI, USA,

[3]University College London, UK

## Introduction

Causal knowledge enables us to predict future events, to choose the right actions to achieve our goals, and to envision what would have happened if things had been different. Thus, it allows us to reason about observations, interventions and counterfactual possibilities. In the past decade philosophers and computer scientists have begun to unravel the relations amongst these three kinds of reasoning and their common basis in causality (e.g., Pearl, 2000; Spirtes, Glymour, & Scheines, 1993; Woodward, 2003).

Observations can provide some information about the statistical relations amongst events. According to the principle of common cause (Reichenbach, 1956), there are three possible causal explanations for a reliable statistical relation between two events A and B. Either A causes B, or B causes A, or both events are generated by a third event or set of events, their common cause. For example, dieting and obesity are statistically related, either because obesity causes people to go on a diet, because dieting disturbs regulatory physiological processes eventually leading to obesity (many obese people went on a diet before they became extremely overweight), or because obesity and dieting may be causal consequences of our modern eating habits. In this last case, we can say that the correlation between obesity and dieting is spurious. Regardless of the underlying causal structure, an observation of one of these events allows us to infer that other events within the underlying causal model will be present or absent as well. Thus, when we have passively observed an event, we can reason backwards diagnostically to infer the causes of this event, or we can reason forward and predict future effects. Moreover, we can infer the presence of spuriously related events.

Interventions often enable us to differentiate amongst the different causal structures that are compatible with an observation. If we manipulate an event A and nothing happens, then A cannot be the cause of event B, but if a manipulation of event B leads to a change in A, then we know that B is a cause of A, although there might be other causes of A as well. Forcing some people to go on a diet can tell us whether the diet increases or decreases the risk

of obesity. Alternatively, changing people's weight by making them exercise would show whether body mass is causally responsible for dieting.

In contrast to observations however, interventions do not provide positive or negative diagnostic evidence about the causes of the event we intervened upon. Whereas observations of events allow us to reason diagnostically about their causes, interventions make the occurrence of events independent of their typical causes. Thus, due to the statistical independence created by interventions these events will occur with their usual base rate independent of the outcome of an intervention. For example, forcing somebody to eat 50 (and only 50) grams of fat per day fixes fat intake independent of the presence or absence of other factors normally affecting choice of diet.

Counterfactual reasoning tells us what would have happened if events other than the ones we are currently observing had happened. If we are currently observing that both A and B are present, we can ask ourselves if B would still be present if we had intervened on A and caused its absence. If we know that B is the cause of A then we should infer that an absence of A makes no difference to the presence of B because effects do not affect their causes. But if our intervention had prevented B from occurring, then we should infer that A would not occur either. For example, Morgan Spurlock (director and guinea pig of the movie "Supersize Me," released in 2004) ate fast food for four weeks and gained more than 20 pounds. What would have happened if he had not eaten burgers and fries all the time? Assuming that the heavy consumption of fast food was the causally responsible factor for the increase in weight rather than the increased weight being the cause for eating, we can conclude that he would have stayed in better shape without all the carbohydrates and fats.

The example indicates that counterfactual reasoning combines observational and interventional reasoning. First we observe Morgan eating fast food and gaining weight. Second we assume that one of the events had been different. We imagine him not eating this diet, while all other observed or inferred factors (e.g., his genetic makeup, the amount of

physical exercise, etc.) are assumed to stay at the observed level. Thus, instantiating a counterfactual event is causally equivalent to an imaginary intervention on a causal model in which all variables that are not affected by the intervention are assumed to stay at the currently observed levels. Finally, the causal consequences of the intervention are inferred on the basis of the given causal model. We infer that Morgan would not have gained as much weight as he did (see next section, Pearl, 2000, and Sloman & Lagnado, 2005, for a more detailed discussion of counterfactuals).

There are important differences amongst observations, interventions, and counterfactuals. Nevertheless, they can be given a unified treatment within the causal model framework. Whereas probabilistic and associative accounts of causal knowledge fail to capture these three interrelated functions of causal knowledge, causal Bayes nets do (Glymour, 2001; Pearl, 2000; Spirtes et al., 1993). The next section will summarize these accounts. Although causal Bayes nets provide successful formal tools for expert systems, only few experiments have tested whether causal Bayes nets also capture everyday reasoning with causal models in people who are not formally trained. The remainder of the chapter will present experimental evidence from the areas of logical reasoning, learning, and decision making demonstrating the plausibility of causal Bayes nets as psychological theories.

## Modeling Observations, Interventions and Counterfactuals

We will not give a detailed description of causal Bayes nets here (see Pearl, 2000, or Spirtes et al., 1993, for detailed introductions). Research on causal Bayes nets not only focuses on causal representation and inference but also on other questions, such as learning (see Lagnado et al., this volume). We focus here on how causal Bayes nets model predictions that are based on observations, interventions, or counterfactual assumptions. Although causal Bayes nets provide tools for reasoning with complex models, experimental studies typically present problems that are within the grasp of naïve human subjects. We will therefore concentrate our brief introduction on inferences within the three basic causal models on which

most psychological research has focused: common-cause, common-effect, and causal chain models. More complex models can actually be generated by combining these three models (see Sloman & Lagnado, 2005 and Waldmann & Hagmayer, 2005, for research on more complex models).

Figure 1: Three basic causal models

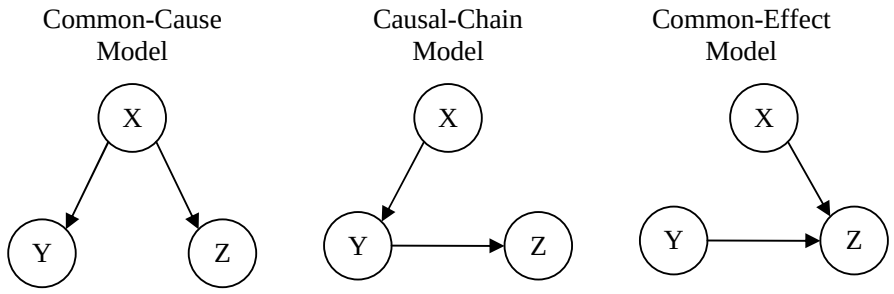



Figure 1 shows the graphs for the three models, with the nodes representing event variables and the arrows signifying direction of causal influence: (1) A common-cause model in which a single cause X influences two effects Y and Z, (2) a causal-chain model in which an initial cause X affects an intermediate event Y influencing a final effect Z, and (3) a common-effect model in which two causes X and Y independently influence a joint effect Z.

The graphs encode assumptions about dependence and independence, which simplify the representation of the causal domain. One important assumption underlying Bayes nets is the Markov assumption, which states (speaking informally) that each event in a causal graph is independent from all events other than its descendants -- i.e. its direct and indirect effects -- once the values of its parent nodes (i.e., its direct causes) are known.

The graph of the common-cause model expresses the spurious correlation between effects Y and Z (due to their common cause) and their independence once the state of cause X is known. This is a consequence of the Markov condition. Once we know that X is present, the probability of Y is the same regardless of whether Z is present or not. Similarly, the

causal chain implies that the initial cause X and the final effect Z are dependent but become independent when the intermediate event Y is held constant. Once we know that Y, the direct cause of Z, is present, the probability of Z stays constant regardless of whether X has occurred or not. Finally, the common-effect model implies independence of the alternative causes X and Y, and their dependence once the common effect is held fixed. This is an example of explaining away. X and Y should occur independently. But once we know that X and its effect Z are present, it is less likely that Y is also present.

Independence is advantageous in a probabilistic model not only because it simplifies the graph by allowing omission of a link between variables but also because it simplifies computation. Conceived as computational entities, Bayes nets are merely partial representations of a joint probability distribution -- *P*(X,Y,Z) in Figure 1 -- that provides a more complete model of how the world might be by specifying the probability of each possible state. Each event is represented as a variable. Causal relations have some relation to the conditional probabilities that relate events; how conditional probabilities and causal relations relate depends on one's theory of the meaning of causation. The factorizations of the three models at issue are:

(1) Common-Cause Model: $P(X,Y,Z) = P(Y|X)\,P(Z|X)\,P(X)$

(2) Causal-Chain Model: $P(X,Y,Z) = P(Z|Y)\,P(Y|X)\,P(X)$

(3) Common-Effect Model: $P(X,Y,Z) = P(Z|Y,X)\,P(Y)\,P(X)$

The equations specify the overall probability distribution of the events within the model on the basis of the strength of the causal links and the base rates of the exogenous causes that have no parents (e.g., X in the common-cause model). Implicit in the specification of the parameters of a Bayes' net are rules specifying how multiple causes of a common effect combine to produce the effect (e.g., noisy-or rule), or (in the case of continuous variables) functional relations between variables. A parameterized causal model allows it to make

specific predictions of the probabilities of individual events or patterns of events within the causal model.

**Modeling Observations**

Observations not only tell us whether a particular event is present or absent, they also inform us about other events that are directly or indirectly causally related to the observed event. Therefore, the structure of the causal model is crucial for inference. Observing an event increases the probability of its causes and of its effects. For example, if someone has a high level of cholesterol, then you can make the diagnostic inference that he or she has probably followed an unhealthy diet (cause) and you can predict that her risk of contracting heart problems is relatively high (effect). These inferences can be justified on the basis of the structure of the causal model. No specific information about the strength of the causal relations or the base rates of the events is necessary to make these qualitative predictions. More specific predictions of the probabilities of events can be made when the model is parameterized.

Formally, observations are modeled by setting the event variables to the values that have been observed. Based on the equations shown above and the probability calculus, the probabilities of other events conditional on the observed variable can be calculated. The structure of the causal model is crucial for these calculations. Imagine that an effect Y of a common cause X which also generates Z is observed. The resulting increase in probability of the cause X can be computed using Bayes rule:

$$P(X=1|Y=1) = P(Y=1|X=1)\,P(X=1) / [P(Y=1|X=1)\,P(X=1) + P(Y=1|X=0)\,P(X=0)]$$

For example, if the base rate of following an unhealthy diet is $P(X=1)= 0.5$, the probability that an unhealthy diet will cause being overweight is $P(Y=1|X=1) = 0.9$, and the probability of being overweight despite eating healthy food is $P(Y=1|X=0) = 0.1$, then being overweight indicates a probability of $P(X=1|Y=1) = 0.9$ that the diet was unhealthy. The probability of the other effect Z can be computed by using the updated probability of the common cause and

the conditional probability P(Z|X) referring to the causal relation connecting the common cause and the second effect. For example, if the probability of having high levels of cholesterol given an unhealthy diet is P=0.4, and P=0.1 otherwise, then the observation of a person's being overweight implies that the probability of having high level of cholesterol is 0.37. Note that this calculation rests on the assumptions that the events are connected by a common cause model. The very same conditional probabilities have different implications given other causal structures.

**Modeling Interventions**

There are different types of interventions (see Woodward, 2003). Interventions can interact with the other causes of events. For example, when we increase fat in our diet then the resulting cholesterol level in our blood depends on our metabolism, prior level of cholesterol, and many other factors. The causal Bayes net literature has focused on a specific type of intervention that completely determines the value of the variable the intervention targets (see Pearl, 2000; Spirtes et al., 1993; Woodward, 2003). For example, if we set the temperature of a room to 20 degrees Celsius, our intervention fixes room temperature while disconnecting all the other factors determining temperature. In this chapter we will focus on this strong type of intervention.

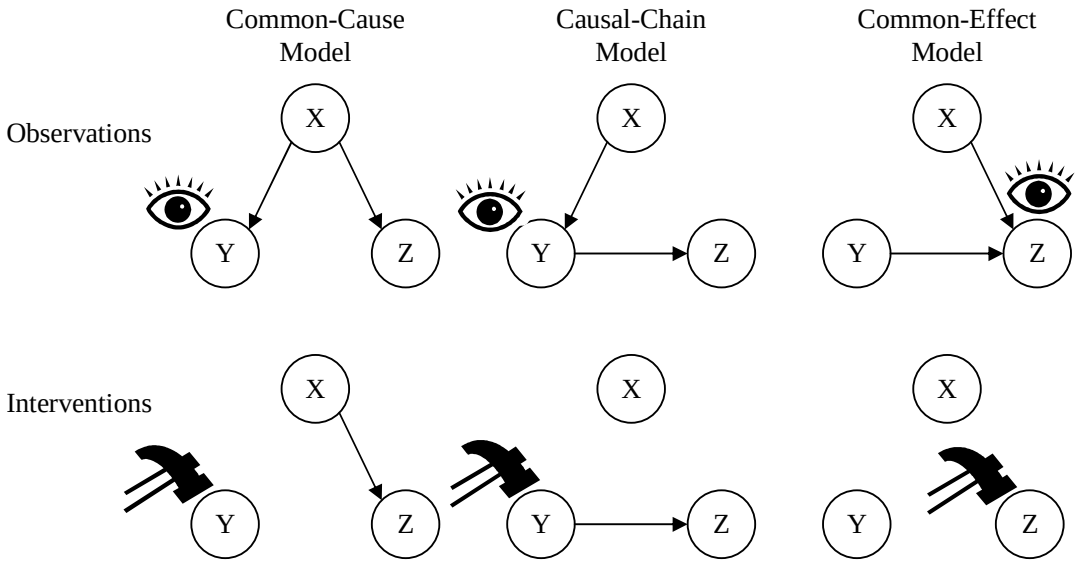
How can interventions be formally modeled? The most important assumption can be traced back to Fisher's (1951) analysis of experimental methods. Randomly assigning participants to experimental and control groups creates independence between the independent variable and possible confounds. Thus, if we, as external agents, set cholesterol levels to a specific value, then the level of cholesterol is independent of other factors normally determining its level. To qualify as an intervention of this strong kind, the manipulation has to force a value on the intervened variable (e.g., cholesterol), thus removing all other causal influences (e.g., diet). Moreover the intervention must be statistically independent of any variable that directly or indirectly causes the predicted event (e.g., all causes of cholesterol),

and it should not have any causal relation to the predicted event except through the intervened-on variable (see Pearl, 2000; Spirtes et al., 1993; Woodward, 2003).

As with observation, predictions of the outcomes of hypothetical interventions are based on specific values of event variables, but whereas observations leave the surrounding causal network intact, interventions alter the structure of the causal model by rendering the manipulated variable independent of its causes. Thus, predictions on the basis of interventions need to be based on the altered causal model, not the original model. For example, the passive observation of low cholesterol level indicates a healthy diet because of the causal link between diet and cholesterol, but medically inducing a specific cholesterol level does not provide evidence about a person's eating habits. Manipulating cholesterol independent of the prior value and other factors creates independence between cholesterol level and diet. Thus, predictions about eating habits can only be based on assumptions about base rates, not on evidence about cholesterol level.

The changes in a causal model caused by interventions (of the strong type) can be modeled by procedures that Pearl (2000) has vividly called "graph surgery." These procedures result in a "manipulated graph" (Spirtes et al., 1993). The key idea is that interventions introduce an external independent cause that fixes the value of the manipulated event. As a consequence, all other causal arrows pointing toward this event need to be removed because these causal influences are not operative during the intervention. Thus, both types of predictions are grounded in a representation of the underlying causal model. However, whereas observational predictions are based on the original causal graph, interventional predictions are based on the manipulated graph. Figure 2 illustrates for the three causal models from Figure 1 how observing differs from intervening. In general, the manipulated graphs are generated by removing the incoming causal links that point to the manipulated variable.

Figure 2: Examples of observations of (symbolized as eyes) and interventions on (symbolized as hammers) the three basic causal models




Traditional Bayes nets (e.g., Pearl, 1988) and other probabilistic theories are incapable of distinguishing between observations and interventions because they lack the expressive power to distinguish between observational and interventional conditional probabilities. Both types are subsumed under the general concept of conditional probability. To distinguish observations from interventions, Pearl (2000), following previous work by Spirtes et al. (1993), has introduced a do-operator. The do-operator represents an intervention on an event that renders the manipulated event independent of all its causes (i.e., it is the formal equivalent of graph surgery). For example, do(Y=1) represents the event that Y was fixed to the value of 1 by means of an intervention, thus removing all previous causal influences in Y. Applying the do-operator allows it to make specific interventional predictions about events within the causal model. For example, the equations for the factorization of the joint distribution of the causal-chain model (Fig. 2) in which the intermediate event is observed to be present (Y=1) or manipulated (do(Y=1)), respectively, are:

Observation of Y: $P(X,Y=1,Z) = P(Z|Y=1)\,P(Y=1|X)\,P(X)$

Intervention on Y: $P(X,do(Y=1),Z) = P(Z|Y=1)\ P(X)$

If the parameters of the causal model are known, we can calculate the probabilistic consequences of interventions. The hypothetical intervention on Y (i.e., Y is fixed to the value of 1, and therefore known to be present) in the causal chain implies that Z occurs with the observational probability conditional upon the presence of Y ($P(Z|Y=1)$, and that X occurs with a probability corresponding to its base rate ($P(X)$). Notice that the interventional probability requires fewer parameters because graph surgery involves simplification by inducing independence between a variable and its causes.

As a second example, consider the common-cause model in Figure 1. Whereas observing Y allows us to reason diagnostically back to its cause X and then reason forward predictively to its spurious correlate Z, predictions for hypothetical interventions in effect Y need to be based on the manipulated graph in Figure 2 in which the link between X and Y is removed. Formally, this can be expressed by the equation:

$P(X,do(Y=1),Z) = P(Z|X)\ P(X)$[1]

Thus, fixing Y at the value 1 removes the link to this variable from the causal model. However, predictions are still possible on the basis of the manipulated graph. The common cause X should occur with a probability corresponding to its base rate and Z is determined by the base rate of its cause X and the strength of the probabilistic relation between X and Z.

**Modeling Counterfactuals**

Counterfactuals combine observations and interventions. The current state of the world is modeled as an observation and then the counterfactual is set by an imaginary intervention altering the state of the variables assumed to be different. For example, we may currently tend to eat unhealthy fast food. For a counterfactual analysis we would first model this fact as if it were an observation by inferring the consequences for other unobserved events within the causal model. We may infer that we have an increased probability of contracting diabetes.

[1] The implications of interventions cannot always be derived from observations (see Pearl, 2000, chapter 3, for a specification of the conditions).

Next we want to know what would happen if we had eaten healthy food instead. We model this counterfactual by means of a hypothetical intervention that fixes the value of the diet variable. Note that counterfactuals differ from interventions, because counterfactual interventions alter causal models, which have been updated before on the basis of the given facts.

As in the case of observations and interventions, graphical causal models are sufficient to draw qualitative inferences from counterfactuals. For example, consider a causal-chain model connecting diet, weight and diabetes. To model the statement "If she were not obese, she would not have developed diabetes," we first assume that we observe diabetes and obesity in a woman. Based on these observations we can infer that the woman probably tends to eat an unhealthy diet. Next, we hypothetically eliminate obesity by means of an intervention that influences this variable by means of a factor external to the chain model (e.g., by assuming that the woman exercises a lot). This hypothetical intervention would cut the causal link between diet and weight but the link between weight and diabetes would stay intact. Therefore, the counterfactual implies that the person in this alternative world would be spared diabetes, while her eating habits would stay the same.

Formal modeling of counterfactuals requires an updating of the model twice. First the probabilities of all events are calculated conditional upon the facts stated in the counterfactual treating facts as observations. Second the counterfactual event is set by the do-operator which entails a re-analysis of the probabilities of the events in the manipulated graph. Thus, assuming the validity of the causal model and the attached parameters, causal Bayes nets allow us to generate precise predictions for counterfactuals.

**Summary**

Causal Bayes nets capture the structure of causal models. They allow us to generate qualitative predictions for observations, interventions, and counterfactuals. Moreover, parameterized causal models enable us to make precise predictions about the probabilities of

events within the causal model. Whereas observational predictions are within the grasp of traditional associative or probabilistic (including Bayesian) theories, modeling interventions and counterfactuals transcends the conceptual power of these models. To correctly model hypothetical interventions and counterfactuals, a preliminary stage has to be assumed in which the structure of the causal model generating the predictions is modified. Based on this modified causal model precise predictions can be made for situations that may have never been observed before.

The distinction between observation and intervention is crucial for the theory of causal Bayes nets. While observations allow us to draw inferences about causes and effects of the observed event, interventions cut the event off from its causes by deleting the causal links pointing towards the event. Sloman and Lagnado (2005) coined the term "undoing" for this process. If causal Bayes nets are veridical models of intuitive human causal reasoning, participants have to be sensitive to undoing. Thus a key issue will be whether human participants are capable of predicting outcomes of hypothetical interventions and of reasoning about causal counterfactuals. This competency would imply that people have access to reasoning processes that modify causal representations prior to deriving predictions. The next three sections will report evidence concerning this question.

## Causal reasoning versus logical reasoning

Causal Bayes nets can be used to represent and model qualitative logical inferences in causal domains. One implication of this account is that causal inference differs from inference in a context in which the standard rules of propositional logic also apply. While standard logic does not distinguish between the observation of an event and the generation of the same event by an intervention, the distinction is central to causal Bayes nets. Causal models have the ability to represent both action (intervention in the world) and imagination (intervention in the mind). If participants are sensitive to the difference between observation and intervention, they should infer that the observation of an event is diagnostic for the

presence of its causes, but when the same event is physically or mentally manipulated, it no longer is.

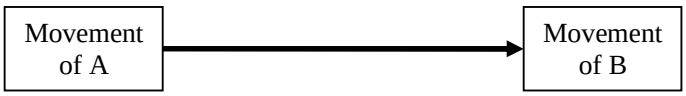
**Observation versus intervention in counterfactual scenarios**

To verify that people are sensitive to the difference between observation and intervention, Sloman and Lagnado (2005) gave a group of students the following scenario

> All rocketships have two components, A and B. Movement of component A causes component B to move. In other words, if A, then B. Both are moving.

Notice that this scenario describes the simplest possible causal model involving only a single link (see Figure 3). Furthermore, the current values of the variables A and B are stated.

Figure 3: Simplest possible causal model




After reading the scenario, half the group was then asked the observational counterfactual question concerning what they would expect if they had *observed* components not moving:

> (i) Suppose component B were observed to not be moving, would component A still be moving?

The other half was asked the interventional counterfactual question concerning what they would expect if components had been intervened on and thereby prevented from moving:

> (ii) Suppose component B were prevented from moving, would component A still be moving?

The difference between observation and intervention should show up in the comparison of (i) and (ii). Observing that the effect B is not moving should be diagnostic of A, suggesting that A too is not moving. In contrast, the logic of intervention says that we should represent an intervention on B as P(A moves | *do*(B does not move)), which reduces to P(A moves)

because B is disconnected from its normal cause A under the do operation. As participants were told before that A is moving, they should stick to that belief and answer "yes." This is just what happened: 85% of participants answered "yes" to (ii) but only 22% answered "yes" to (i). B's movement was only treated as diagnostic of A's movement when B was observed not to move, not when its movement was prevented. This shows that people are sensitive to the logic of a counterfactual intervention in a situation with a transparent causal structure.

**Causal reasoning versus propositional logic**

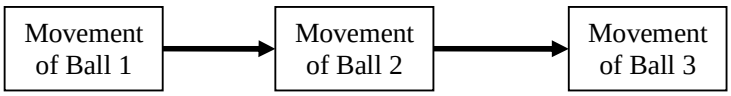
The causal-model framework predicts that people are sensitive to the logic of intervention when reasoning causally, not necessarily when reasoning in other ways. Sloman and Lagnado (2005) compared reasoning in a situation with causal relations to one with parallel relations that were not causal.

Consider the following causal problem described in terms of conditional ("if...then...") statements:

> Causal conditional There are three billiard balls on a table that act in the following way: If Ball 1 moves, then Ball 2 moves. If Ball 2 moves, then Ball 3 moves.
>
> Imagine that Ball 2 could not move, would Ball 1 still move?

The fact that we're talking about billiard balls – prototypical causal elements – strongly suggests that the conditional statements should be interpreted as describing causal relations. The causal model underlying this scenario is depicted in Figure 4.

Figure 4: Causal Chain Model used by Sloman and Lagnado (2005)

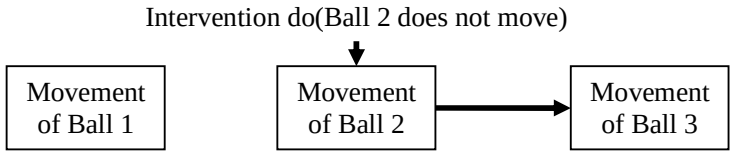


The causal modeling framework represents the two questions using the do operator because an outside agent is preventing the ball from moving, represented as *do*(Ball 2 does not move):

P(Ball 1 moves | *do*(Ball 2 does not move)).

To evaluate this, we must assume that Ball 2 does not move. We must also simplify the causal model by removing any links into Ball 2 as depicted in Figure 5.

Figure 5: Causal chain model altered by an intervention on the intermediate event (adapted from Sloman & Lagnado, 2005)




It is immediately apparent, parallel to the last example, that the value of Ball 1 is no longer affected by Ball 2 and therefore the causal Bayes model framework predicts that Ball 2's lack of movement is not diagnostic of its normal cause, Ball 1. 90% of participants agreed, affirming that Ball 1 could move if Ball 2 could not.

Standard propositional logical systems have no way to represent this argument. Not only do they not have a representation of cause, they have no way of representing an intervention. A conventional logical analysis of this problem might go as follows: The problem tells us that if Ball 1 moves, then Ball 2 moves. We know that Ball 2 does not move. Therefore, Ball 1 does not move by modus tollens. This particular argument does not explain people's judgments which are that Ball 1 can move even if Ball 2 cannot.

In the noncausal realm, modus tollens can be a perfectly valid form of argument for deriving definite conclusions. For example, modus tollens would be an appropriate inference schema to use on a problem similar to the causal one just shown but based on logical if-then relations rather than causal ones. Maybe people would make inferences conforming to modus

tollens with such an argument. To find out, Sloman and Lagnado (2005) gave a group of people the following scenario:

> Logical conditional. Someone is showing off her logical abilities. She is moving balls without breaking the following rules: If Ball 1 moves, then Ball 2 moves. If Ball 2 moves, then Ball 3 moves.

and then asked them the same question as for the causal case:

> Imagine that Ball 2 could not move, would Ball 1 still move?

In this case, only 45% of participants said "yes." The majority gave the inference consistent with modus tollens, "no." Clearly there is less consistency than in the causal case probably because participants are more confused in a logical than in a causal context. Their answers are more wide ranging and they tend to express less confidence. People's discomfort with logical problems relative to causal ones arises either because there are different forms of logic and they are not sure which one to pick or because no form of deductive logic comes naturally.

The experiments by Sloman and Lagnado (2005) show that causal reasoning is neither adequately modeled by standard propositional logic formalisms, nor by traditional probabilistic theories that do not distinguish intervention from observation. Causal Bayes nets are the best currently available account that models this competency.

## Reasoning with parameterized causal models

The previous section has shown that people can qualitatively reason with causal models, and that they do distinguish between observation and intervention. Waldmann and Hagmayer (2005) have addressed similar questions in the realm of learning. Following the framework of causal-model theory (Waldmann, 1996; Waldmann & Martignon, 1998; see also Lagnado et al., this volume), participants were instructed about a causal model underlying the learning domain prior to receiving learning data. The learning data consisted of individual cases that allowed participants to estimate the parameters of the assumed causal model (e.g., causal

strength, base rates). The main question was whether learners were capable of deriving precise predictions on the basis of the parameterized models, and whether their predictions differ depending on whether the predictions are based on hypothetical observations or hypothetical interventions. Again causal Bayes nets provided the formal tools to analyze this competency.

Associative theories are the dominant approach in the realm of learning. They can differentiate between observing and intervening by postulating separate learning modes: Whereas *classical conditioning* might be viewed as underlying observational predictions, interventional predictions might be driven by *instrumental conditioning* (Dickinson, 2001; see Domjan, 2003, for an overview). Thus, we might learn in an observational learning paradigm (classical conditioning) that the barometer reading predicts the weather, and in an interventional learning paradigm (instrumental learning), we might additionally learn that fiddling with the barometer does not change the weather. However, although this approach approximates causal knowledge in many contexts, it fails to capture the relations between observation and intervention. The separation between classical and instrumental conditioning predicts that without a prior instrumental learning phase we should be incapable of correctly predicting what would happen in case of an intervention in situations in which our knowledge is based on observational learning. Waldmann and Hagmayer's (2005) experiments show that this is wrong. People not only were capable of deriving predictions for hypothetical interventions after a purely observational learning phase, their predictions were also sensitive to the structure of the underlying causal model and the size of the parameters.

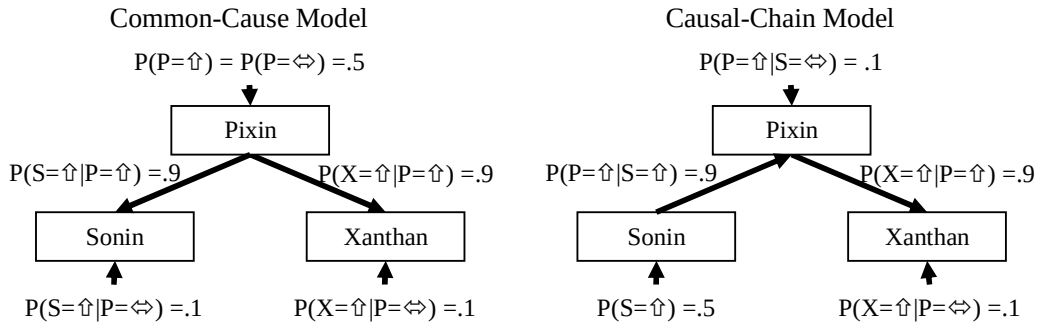
**Predicting the outcomes of hypothetical interventions from observations**

Experiment 2 of Waldmann and Hagmayer (2005) provides an example of the learning task. In this experiment participants were taught either a common-cause or a causal-chain model. In a fictitious medical scenario that involved hormone levels of chimpanzees, they were told either that an increased level of the hormone pixin causes an increase in the level of

sonin and of xanthan (common-cause model), or that an increase in the level of sonin causes the level of pixin to rise which in turn increases the amount of xanthan (causal-chain model) (see Fig. 6). Waldmann and Hagmayer compared these two models because the common-cause model normatively implies a dissociation between observational and interventional predictions whereas the chain model implies identical predictions for both types, allowing it to test whether people correctly differentiate between different causal models.

After the initial instructions participants received descriptions of the hormone levels of a set of twenty individual chimpanzees as observational data. The causal relations were probabilistic (see Fig. 6). Using the data, learners could estimate the parameters of the causal models. Causal-chain and common-cause models have the same structural implications (they are Markov equivalent); therefore only one set of data was presented that was coherent with both causal models. The models and the implied parameters are shown in Figure 6.

Figure 6: Conditions and data of Experiment 2 by Waldmann and Hagmayer (2005). Upward arrows symbolize increased hormone levels, sideways arrows normal levels. The parameters represent causal strength (conditional probabilities) and base rates (unconditional probabilities).



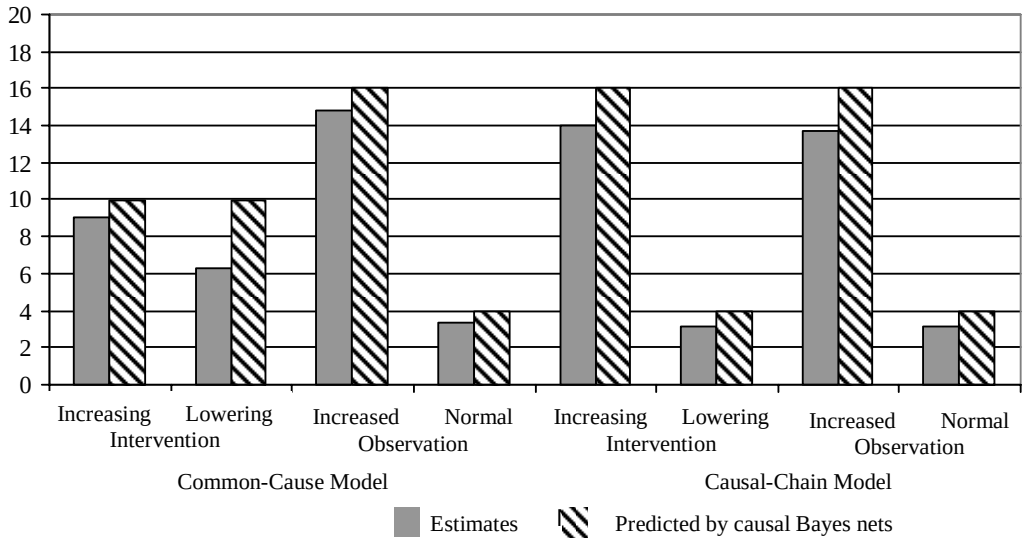

A Bayesian analysis of these parameterized models implies for both models that the probability of increased levels of xanthan conditional upon sonin being observed to be at an elevated level is P(X=⇧|S=⇧) =.82, whereas the corresponding conditional probability is

$P(X=\Uparrow|S=\Leftrightarrow) = .18$ when the sonin level is normal. The base rate of the exogenous causes in both models (i.e., sonin in the common-cause model, pixin in the chain model) was set to 0.5.

For the causal chain model, the interventional probabilities are identical to the observational probabilities. For example, regardless of whether sonin is observed to be increased or whether an increased level was caused by means of an inoculation, the other two hormones should be equally affected. However, an intervention in sonin in the common-cause model entails the removal of the causal arrow connecting pixin and sonin. Therefore the probability of xanthan depends only on the base rate of its cause pixin and the causal impact of this hormone on xanthan. Thus, the interventional probability of xanthan is $P(X=\Uparrow|do[S=\Uparrow]) = P(X=\Uparrow|do[S=\Leftrightarrow]) = .5$, regardless of whether sonin is increased or normal.

To test whether participants' judgments follow these predictions, they were asked to make predictions about hypothetical observations and hypothetical interventions after having studied the learning data. All participants were requested to estimate for a set of twenty new, previously unseen chimpanzees the number of animals showing elevated levels of xanthan based on the hypothetical observations that sonin was observed to be at either an increased or normal level in these animals. The corresponding questions about hypothetical interventions asked participants to imagine inoculations that increased or lowered the level of sonin in the twenty animals. The order of the test questions was counterbalanced. The mean response to the test questions and the answers predicted by the causal model framework are shown in Figure 7.

Figure 7: Results of Experiment 2 of Waldmann and Hagmayer (2005). Mean responses and predicted frequencies to observation and intervention question.




The pattern of results shows that participants correctly differentiated between observational and interventional predictions and that they were sensitive to the different implications of the contrasted causal models. Whereas for the causal chain model learners correctly predicted similar levels of xanthan independent of whether sonin levels were observed or generated, a clear dissociation was observed for the common-cause model. The majority of participants concluded that the probability of xanthan is independent of the type of intervention upon sonin. A second interesting finding was that on average estimates were as predicted although in some cases there was a slight tendency to underestimate. The largest deviation between the estimates and the normative values was found for the intervention lowering the level of sonin (second pair of columns in Figure 7), which is probably due to the fact that participants had no data about what would happen if the level of one hormone would fall below a normal level.

These results are beyond the grasp of associationist theories. This is most obvious in the common-cause model in which the predictions of the outcomes of the hypothetical interventions turned out close to the predicted value of 50 percent, despite the fact that

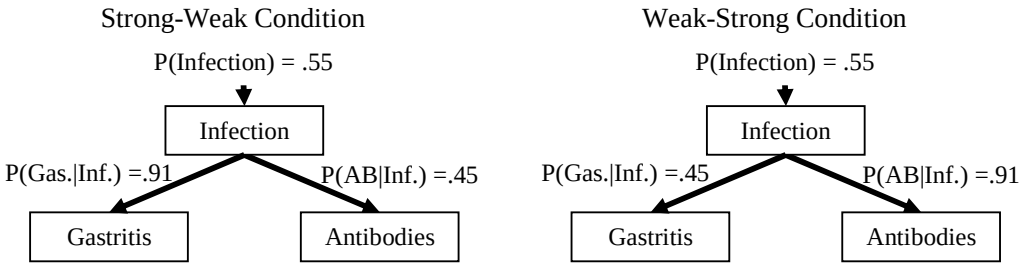
learners had never observed this value in the learning phase. These predictions clearly support causal models as descriptions of human reasoning. Apparently reasoners rely not only on the observed associations but also on the underlying causal model to generate predictions.

**Sensitivity to Parameters**

To examine whether learners used the learned parameters for their predictions, Waldmann and Hagmayer (2005) ran additional studies manipulating parameter values across conditions. Experiment 4 provides an example of this manipulation. In this experiment, participants were instructed that a fictitious bacterial infection in dogs has two causal effects, gastric problems and increased antibodies (i.e., common-cause model). In two conditions, two different data sets were shown to participants in a list format. The two data sets varied the strength of the two causal relations. In one condition ("strong-weak") the bacterial infection had a strong influence upon gastric problems ($\Delta P=.91$) and only a medium influence on the presence of antibodies ($\Delta P=.45$). ($\Delta P$ is a measure of contingency that reflects the numeric difference between the probability of the effect, gastric problems, conditional upon the presence and absence of the cause (e.g., bacterial infection).) In the other condition, the assigned causal strength was reversed ("weak-strong") (see Fig. 8). The base rate was the same (0.55) in both conditions.

Participants were requested to estimate the frequency of antibodies in a new set of 20 dogs assuming either that gastritis was observed to be present or absent or that the presence or absence of gastritis was caused by means of an external intervention (inoculation).

Figure 8: Conditions and data of Experiment 4 of Waldmann and Hagmayer (2005)



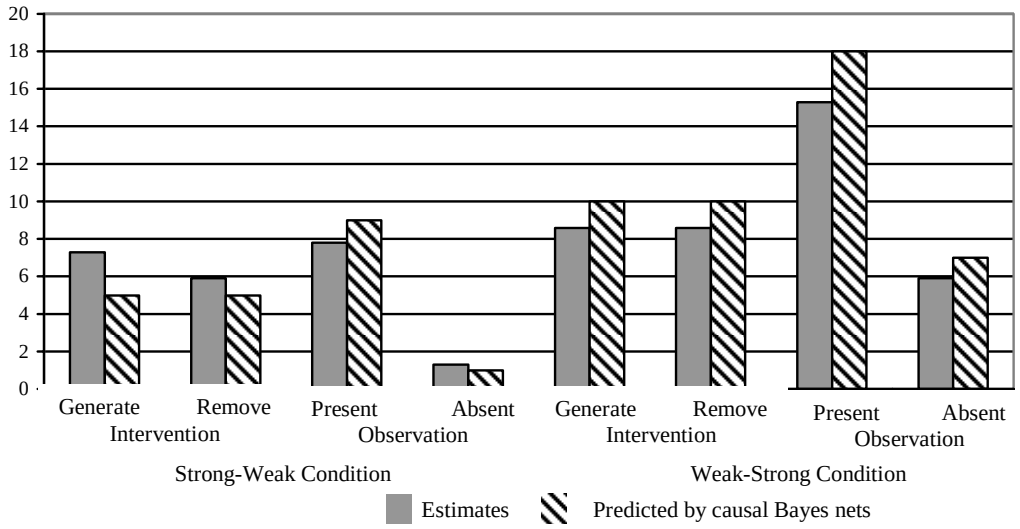

Although the structure of the causal model is identical in both conditions, the parameters implied by the two data sets have distinctive implications for the different types of predictions (see Fig. 8). Because of the underlying common-cause model, an external intervention in gastric problems has no causal influence on the infection rate and the presence of antibodies. This is due to graph surgery that requires a removal of the causal arrow between the common cause infection and gastritis. The probability of antibodies is solely determined by the base rate of the bacterial infection and its causal impact on the antibodies. Therefore antibodies are more likely in the condition in which bacterial infection has a strong influence (i.e., "weak strong") than when it has only a weak impact (i.e., "strong-weak").

The different parameters in the two conditions not only imply different predictions for the intervention questions but also for the observation questions. In general, the implied probabilities are higher if gastritis is observed to be present than if it is absent. In addition, the probability of antibodies is higher in the "weak-strong" condition than in the "strong-weak" condition.

In Figure 9 the mean responses are compared with the values predicted by the causal model. The results show that participants again differentiated between predictions for hypothetical observations and hypothetical interventions. Moreover, the estimates also demonstrate that participants were sensitive to the parameters of the causal model. On average, participants' estimates were quite accurate although there are again small deviations

that could be due to regression effects. This competency is rather surprising considering the complexity of the task.

Figure 9: Results of Experiment 4 of Waldmann and Hagmayer (2005). Mean responses and predicted frequencies for the observation and intervention questions.




Sensitivity to the size of parameters was not only shown for the causal strength parameters but also for the base rate parameters. In another experiment (Waldmann & Hagmayer, 2005, Experiment 3), the base rate of the common cause was manipulated while holding causal strength constant. This should particularly affect the interventional predictions (based on interventions on the first effect) as the probability of the predicted second effect in this case varied in proportion to the base rate of its cause (see Fig. 2). The results showed that participants incorporated the base rate information in their predictions in a way that was surprisingly close to the normative predictions of causal Bayes nets.

## Causal Decision Making

The distinction between observation and intervention also has practical implications for decision making. For example, if we observe low values on a barometer we will probably take our umbrella because the probability of rain is high. But we also know that setting the

barometer by means of an intervention will not affect the weather. The evidential relation between the barometer reading and the weather is spurious and mediated by atmospheric pressure which acts as a common cause that independently affects the barometer and the weather. Thus, observing a low reading of the barometer due to tampering should not influence our decision to take an umbrella. This example shows that causal models and the distinction between observation and intervention are highly relevant to decision making. Specifically, choice is a form of intervention and should be modeled as such, by breaking the edge between the variable whose value is being chosen and its normal causes. However, most theories of decision making, certainly most normative theories, analyze decision-making on the basis of evidential relations between variables (e.g., subjective expected utility theory).

In contrast, in line with the analyses of causal Bayes nets and previous work on causal expected utilities (Nozick, 1969, 1995), Hagmayer and Sloman (in prep.) propose that choice is equivalent to an intervention in a causal network. They claim that in decision making people first consider a causal model of the decision context and then explore the causal consequences of their possible interventions.

**Simple Choices**

Hagmayer and Sloman presented participants with simple decision problems, such as the following:

> Recent research has shown that of 100 men who help with the chores, 82 are in good health whereas only 32 of 100 men who do not help with the chores are. Imagine a friend of yours is married and is concerned about his health. He read about the research and asks for your advice on whether he should start to do chores or not to improve his health. What is your recommendation? Should he start to do the chores or not?

Hagmayer and Sloman also provided participants in different conditions with one of two causal models that might underlie the correlation between chores and health. In one condition

the relation was due to a common cause, the degree of concern, that independently influences the likelihood of doing the chores and of entertaining health-related activities, or in the alternative direct-link model it was pointed out that chores are an additional exercise directly improving health.

Participants received several different decision problems involving a range of issues from the relation between high risk sports and drug abuse to the relation between chess and academic achievement. If participants equate choices with interventions they should often recommend not acting in the common-cause condition because intervening on an effect of a common cause does not alter the spuriously correlated collateral effect. Such an intervention would simply render the action independent of the rest of the model, including the desired outcome. In contrast, in the second condition, participants should recommend doing the chores because this variable is directly causally related to health. Participants' judgments turned out to be in accordance with the causal-model theory of choice. Despite learning about an identical evidential relation, only 23 percent of the participants in the common-cause condition advised their hypothetical friend to act, in contrast to 69 percent of the participants in the direct-link condition.

**Complex choices and Newcomb's paradox**

The difference between observational and interventional probabilistic relations is crucial in more complex cases as well. Newcomb's paradox is an interesting test case because it involves a conflict between two principles of good decision making: (i) maximizing expected utility and (ii) dominance (i.e., choosing the option that always leads to the better outcome) (see Nozick, 1969, 1995). Classical decision theory cannot handle this paradox, as it has no principled way to choose between these alternative criteria; however, a causal analysis in some cases can. Figure 10 illustrates a variant of Newcomb's paradox which Hagmayer and Sloman used in an experiment. In this experiment, students were asked to imagine being the marketing executive of a car manufacturer and having to choose between two advertising

campaigns. They could either promote their sedan or their minivan. However, according to the instructions the expected sales not only depend on their decision but also on the marketing decision of their main competitor (see Fig. 10).

Figure 10: Payoff matrix (in dollars) used by Hagmayer and Sloman (in prep.)

| Additional Sales | Competitor promotes sedan | Competitor promotes minivan |
|---|---|---|
| You promote sedan | 30,000 | 15,000 |
| You promote minivan | 40,000 | 20,000 |

As the payoff matrix shows, higher sales are expected for the minivan regardless of the competitor's campaign. Therefore the principle of dominance prescribes promoting the minivan. However, participants were also informed that in the past the two car companies ended up promoting the same type of car in 95 percent of the cases with either car being promoted equally often. If this additional information is taken into account the expected value of promoting the sedan turns out to be higher than that of the minivan (29.250 vs. 21.000). Thus the principle of maximizing expected value implies the opposite of the principle of dominance.

To investigate the influence of the assumed causal model, participants were additionally informed about the causal relations underlying the observed evidential relations. In one condition, participants were told that the other competitor tends to match the participant's strategy (direct-cause model), in the other condition that both car companies make their decisions independently based on the market (common-cause model). After considering the information participants were requested to choose one of the available options.

Under the direct-cause model the evidential probabilities between the choices of the two competitors indicate a stable causal relation. Therefore the causal expected utility equals the evidential expected utility, and hence the sedan should be promoted. In contrast, under a

common-cause model the choice should be viewed as an intervention that is independent of the competitor's choice, who is supposed to choose on the basis of the state of the market. Because a free choice destroys the evidential relation between the choices of the participant and the hypothetical competitor, the assumption that both choices will coincide is no longer valid. Thus, the dominant option is the best choice under a common-cause model.

The results supported the predictions derived from a causal-model theory of choice. In two different scenarios that presented variants of Newcomb's paradox, 37 percent of the participants who were given a direct-cause model preferred the dominant option in contrast to 78 percent of the participants who were given a common-cause model.

These results show that decision makers were sensitive to the structure of the underlying causal model, and that they tended to treat choices as interventions. Whereas traditional theories of decision making fail, causal Bayes nets provide a coherent account to model decision making in causal domains.

## Final Remarks

The research on the psychological validity of causal Bayes net theories has only begun (see also Gopnik et al., 2004; Lagnado & Sloman, 2004; Steyvers et al, 2003; Tenenbaum & Griffith, 2003, Waldmann & Martignon, 1998, for more findings). In this chapter, we reported some very recent evidence on the distinction between observation and intervention. Traditional probabilistic and associationist theories are incapable of distinguishing between the different predictions entailed by hypothetical observations and interventions. The results of the experiments show that people are remarkably good at distinguishing between predictions based on observed events and predictions based on hypothetical interventions. The empirical evidence supports the theory that people reason prior to making predictions. While observational predictions are based on the structure of the causal model that is being used, interventions require mentally modifying the model prior to deriving predictions to achieve

what Sloman and Lagnado (2005) call “undoing” effects in reasoning and choice (see also Waldmann & Hagmayer, 2005).

People are not only capable of deriving qualitative predictions that are implied by the structure of the causal models, they also proved capable of incorporating the learned parameters in their predictions (Waldmann & Hagmayer, 2005). The ability to distinguish between interventions and observations was found in a variety of task domains, including logical reasoning, learning, and decision making. These results clearly support the representation used by the theory of causal Bayes nets to distinguish observation from intervention.

**References**

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

Domjan, M. (2003). *The principles of learning and behavior* (5th. ed.). Belmont: Thomson/Wadsworth

Fisher, R. (1951). *The design of experiments*. Edinburgh: Oliver and Boyd.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Science, 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.

Hagmayer, Y., & Sloman, S. A. (in prep.). A causal-model theory of choice.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 856-876.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (ed.), *Essays in honor of Carl G. Hempel* (pp. 107-133). Reidel.

Nozick, R. (1995). *The nature of rationality*. Princeton. Princeton University Press.

Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.

Reichenbach, H. (1956). *The direction of time*. Berkeley and Los Angeles: University of California Press.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*.

Spirtes, P., Glymour, C., & Scheines, R. (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.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In *Advances in neural information processing systems, 15*, 35-42. Cambridge: MIT Press.

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, CA: Academic Press.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry, *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1102-1107). Mahwah, NJ: Erlbaum.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Motivation, and Cognition.*

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