Seeing versus Doing:
Causal Bayes Nets as Psychological Models
# of Causal Reasoning

DISSERTATION

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen
vorgelegt von

BJÖRN MEDER
aus Hannover

Göttingen, im März 2006

D7

Referent:	Prof. Dr. Michael R. Waldmann
Koreferent:	Prof. Dr. Gerd Lüer

Tag der mündlichen Prüfung: 3. Mai 2006

Prediction is very difficult, especially about the future.

Niels Bohr (1885 –1962)

Abstract

This dissertation is concerned with the question of how people infer the consequences of active interventions in causal systems when only knowledge from passive observations is available. Causal Bayes nets theory (Spirtes, Glymour & Scheines, 1993; Pearl, 2000) provides a rational account of causality which explicitly distinguishes between merely observed states of variables ("seeing") and identical states due to external interventions ("doing"), and which provides mechanisms for predicting the outcomes of hypothetical and counterfactual interventions from observational knowledge. By contrast, alternative models of causal cognition (e.g., associative theories) fail to capture the crucial difference between observations and interventions and thus are likely to generate erroneous predictions when the implications of observations and interventions differ.

The basic research question of the eight experiments presented in this thesis was to investigate whether people who have observed individual trials presenting the states of a complex causal model can later predict the consequences of hypothetical and counterfactual interventions in a way predicted by causal Bayes nets. Consistent with the Bayes nets account learners were surprisingly good at inferring the consequences of interventions from observational knowledge in accordance with the structure and the parameters of the observed causal system. The experiments also show that participants were capable of taking into account the implications of confounding variables when reasoning about complex causal models. Although participants' inferences were largely consistent with the predictions of causal Bayes nets, the studies also point to some boundary conditions of the competencies of lay reasoners. For example, learners had problems distinguishing hypothetical interventions from counterfactual interventions.

In summary, the experiments strongly support causal Bayes nets as a model of causal reasoning. Alternative theories of causal cognition lack the representational power to express the crucial differences between observations and interventions and therefore fail to account for the results of the experiments.

Zusammenfassung

Diese Dissertation geht der Frage nach, wie Menschen Vorhersagen über die Folgen von aktiven Interventionen in kausalen Systemen zu treffen, wenn sie diese Systeme zuvor nur passiv beobachtet haben. Die Theorie der kausalen Bayes-Netze (Spirtes, Glymour & Scheines, 1993; Pearl, 2000) stellt einen rationalen Ansatz zur Repräsentation von Kausalwissen dar und formalisiert den Unterschied zwischen passiv beobachteten Ereignissen („seeing“) und identischen Ereignissen, die durch Interventionen aktiv erzeugt wurde („doing“). Dadurch ermöglicht es der Formalismus, die Folgen von hypothetischen und kontrafaktischen Interventionen aus Beobachtungswissen abzuleiten. Alternative Theorien kausalen Denkens hingegen, die den Unterschied zwischen passiv beobachteten und aktiv erzeugten Ereignissen nicht berücksichtigen, generieren fehlerhafte Vorhersagen, wenn Beobachtungen und Interventionen unterschiedliche Implikationen haben.

Die grundlegende Forschungsfrage der acht Experimente dieser Arbeit ist, ob Menschen die Folgen von hypothetischen und kontrafaktischen Interventionen aus Beobachtungswissen ableiten können, das in einem passiven Trial-by-Trial Lernverfahren erworben wurde. In Übereinstimmung mit der Theorie kausaler Bayes-Netze zeigte sich, dass die Versuchsteilnehmer überraschend gut darin waren, die Folgen von Interventionen aus Beobachtungswissen abzuleiten, und dass sie dabei auch die Struktur und die Parameter des beobachteten Kausalmodells einbeziehen. Zudem zeigen die Befunde, dass konfundierende Variablen bei den jeweiligen Vorhersagen adäquat berücksichtigt werden. Obwohl die Schlussfolgerungen der Versuchsteilnehmer insgesamt den Vorhersagen der Theorie kausaler Bayes-Netze entsprachen, zeigen die Befunde auch einige Randbedingungen auf. So hatten die Probanden zum Beispiel Probleme, zwischen den Implikationen von hypothetischen und kontrafaktischen Interventionen zu differenzieren.

Insgesamt stützen die Ergebnisse klar die Theorie der kausalen Bayes-Netze als psychologisches Modell kausalen Denkens. Alternative Theorien kausaler Kognitionen, die die Unterschiede zwischen beobachteten und durch Interventionen erzeugten Ereignissen nicht repräsentieren, können die Ergebnisse der Experimente nicht erklären.

Contents

1 Introduction

The ability to acquire and use causal knowledge is a central competency necessary for explaining past events and predicting future events. What are the causes of cancer? How does inflation affect economic growth? How will greenhouse gases influence our climate? Causality is “the cement of the universe”, as the philosopher Mackie (1974) once put it, and both in science and everyday reasoning we aim to reveal the causal texture of the world we live in. However, the question of how we acquire knowledge of causal relations has puzzled both philosophers and psychologists for centuries. It was the philosopher David Hume (1711–1776) who, with his striking analysis of causality, posed the fundamental challenge all theories of causal induction have ever since had to address: how do we learn about causal relations even though our sensory input contains no direct causal knowledge? The solution offered by Hume was that we induce causal relations from spatio-temporal contiguity and covariational information: if two events are repeatedly observed to vary together in space and time we will infer that they are causally related. Causal knowledge derived from such observations enables us to predict one event from the other: from observing the cause event we can infer the presence of the effect event, and from observing the effect we can infer the presence of the cause event.

Seeing versus Doing: Causal Inferences with Observations and Interventions

Causal knowledge acquired from passive observations can be contrasted with causal knowledge concerning the consequences of our actions. Would I develop a rash if I ate this fruit? What would happen if I pressed this red button that says “Do not push”? One way to directly acquire this kind of interventional knowledge is by trial and error. If people have tried out the interventions on previous occasions they know the potential outcomes of their actions. Similarly, in scientific studies the candidate cause is manipulated to learn about its effects. Learning from interventions directly provides us with causal knowledge about the consequences of interventions. However, learning through intervention is not always possible. In some sciences (e.g., astronomy) and also in many everyday contexts, often only observational knowledge is available. The question is then how we can infer the consequences of our actions from observational knowledge. A tempting solution would be to equate observational knowledge with instrumental knowledge and proceed from there. Unfortunately, this strategy will often lead to ineffective actions. For example, the status of a barometer is statistically related to the approaching weather due to their common cause, atmospheric pressure (cf. Figure 1). Even though this correlation does not indicate a genuine causal relation, observational predictions can capitalize on such spurious statistical relations. In contrast, interventional predictions cannot, since manipulating the barometer obviously does not affect the weather. Effects do not change their causes; thus, manipulating the barometer does not affect its cause, atmospheric pressure, and therefore has no causal influence on the weather. While observational inferences are often warranted by correlational data alone, interventional predictions require us to represent the causal structure underlying our observations.

The difference between observing ("seeing") and intervening ("doing") is compelling in the barometer example, and, at first glance, the example may look rather trivial: is it not obvious that manipulations of the barometer will not influence the weather? However, this simple example elucidates a general problem of formal models such as standard probability calculus: these accounts provide no formal means to express the difference between merely observed states of variables and the very same states generated by external interventions. This is also mirrored in most traditional theories of causal cognition which, in one way or the other, model the way covariational information is processed to derive causal judgments. As a consequence, these models collapse observational and interventional knowledge and are likely to generate erroneous predictions when the implications of observations and interventions differ. For example, associative theories of causal induction distinguish between observational learning (classical conditioning) and interventional learning (instrumental conditioning), but, as the barometer example shows, they fail when predictions for instrumental actions have to be derived from observational learning.

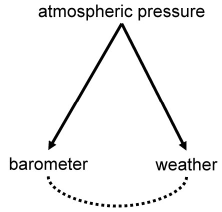


Figure 1. Simple causal model with three variables. Arrows indicate causal relations; dashed line indicates a spurious correlation.

Causal Bayes Nets as Formal Account of Causal Cognition

Recently, causal Bayes nets theory (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993) has been developed as a normative formal account of causal representation, causal learning, and causal reasoning. Originally developed in the context of computer science, philosophy, and statistics, it has been argued that the formalism also captures important aspects of human causal cognition (e.g., Glymour, 2001; Gopnik et al., 2004; Meder, Hagmayer, & Waldmann, 2005; Sloman & Lagnado, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Waldmann & Hagmayer, 2005). Causal Bayes nets combine graphical causal models with probability calculus to represent causal knowledge and formalize causal learning and causal reasoning. In contrast to most other theories of causal cognition, which consider causal induction as a purely data-driven process, causal Bayes nets theory assumes that top-down and bottom-up processes interact in both the acquisition and use of causal knowledge. A hallmark of causal Bayes nets theory is that the formalism explicitly distinguishes merely observed states of variables from identical states due to external interventions. By providing a formal account of interventions in causal systems the theory allows for the derivation of precise predictions for the consequences of hypothetical and counterfactual interventions from observational knowledge and graphical representations of causal systems.

Structure and Aims of this Dissertation

The goal of this dissertation is to test some of the fundamental predictions of causal Bayes nets theory. My intention is neither to test all aspects of causal Bayes nets theory nor to claim that the Bayes nets formalism provides a universal model of causal cognition. Clearly, the scope of Bayes nets is beyond that of psychological models. For example, Bayes nets can easily handle complex causal systems with hundreds of variables and employ sophisticated algorithms which can analyze large amounts of statistical data. Consequently, in many situations the account will make psychologically implausible assumptions about the necessary information processing capacities. Nevertheless, I will argue that causal Bayes nets provide a useful tool to model important aspects of human causal cognition which conventional models of causal cognition fail to capture. A central emphasis is placed on the assumption that causal induction is not a purely data-driven process but that bottom-up and top-down processes interact in causal learning and causal reasoning.

The aim of the experiments presented here is to investigate three key issues. First, the basic research question is whether learners distinguish between observations and interventions and have the capacity to derive interventional predictions from observational knowledge. A demonstration of learners' capacity to derive adequate interventional predictions from observations would not only support causal Bayes nets theory but also question the traditional separation of representations acquired from observational and interventional learning, as, for example, proposed by associative learning theories. To compare the findings with the predictions of associative theories of causal learning, the experiments employ passive trial-by-trial learning which is assumed to provide optimal conditions for the operation of associative learning mechanisms (cf. Shanks, 1991). Second, the experiments present learners with causal models that contain confounding causal relations. Whereas randomized experiments ensure the independence of the candidate cause from all other potentially confounding variables, merely observed statistical relations may include the influence of confounding variables which are both related to the potential cause and the presumed effect. In such situations reasoners have to disentangle the direct causal influence from a concurrent spurious relation to derive adequate interventional predictions. The goal of the experiments is to tap into participants' understanding of the causal logic of confounds. Finally, the experiments presented here also allow for an investigation of the boundary conditions of learners' capacity to reason in accordance with the normative framework of causal Bayes nets. The causal inferences participants are requested to draw therefore differ with respect to the kind of intervention, the number of variables and causal relations that have to be taken into account, and the way the learning data is presented.

The structure of this work is as follows. I will first give an introduction to Hume's analysis of causality and critically review his account of causal induction. I will continue with an overview of psychological theories of causal learning and conclude the theoretical section by introducing causal Bayes nets theory as a formal framework of causal inference. In the empirical section I will present a series of experiments which aim at investigating the adequacy of Bayes nets formalism as a psychological model of causal reasoning. Throughout the experiments learners are asked to draw observational, interventional, and counterfactual inferences from causal models and observational knowledge. Thus, a main goal is to investigate participants understanding of these three types of causal inference. In Experiments 1 and 2 learners are provided with identical learning input but are suggested different causal structures. Conversely, in Experiments 3 and 4 learners are provided with identical causal models but the learning data is manipulated. The goal of these studies is to highlight the interaction between top-down and bottom-up processes in causal reasoning. In Experiments 5 and 6 the robustness of learners' competency to derive interventional predictions from observational learning is tested by manipulating the way the learning data is presented. To achieve this goal, temporal order during learning is pitted against causal order, which provides a potentially misleading cue to causality. Finally, Experiments 7 and 8 further investigate learners' understanding of confounds. In these studies, learners are presented with alternative causal models which generate very similar observational data but strongly differ with respect to the consequences of interventions. Thus, the observational data learners are provided with must be simultaneously used to choose between competing models and to estimate the chosen model's parameters. The dissertation concludes with a discussion of the empirical findings and their implications for psychological models of causal cognition.

2 Philosophical Background

The question of how we learn about causal relations is not only a concern of psychological theories but has also been an important issue throughout the history of philosophy. The following section is intended to give a brief introduction to the philosophical debate on causation and causal learning. Contrary to psychological theories which mainly address the epistemological question of how we acquire and make use of causal knowledge, many philosophers have rather focused on the ontological aspects of causality. However, even though philosophical and psychological theories often address different questions, many current theories of causal cognition have been inspired by philosophical accounts of causality. Especially the fundamental analysis of causality given by the Scottish philosopher David Hume has been of great importance to the development of theories of causal learning. His suggestion that the acquisition of causal knowledge is a purely data driven process in which causal knowledge is derived from covariational data is still alive in most psychological theories of causal induction.

2.1 Hume's Riddle of Causal Induction

Hume's epistemological orientation was that of a radical empiricism according to which all knowledge is derived from experience. Traditionally, this account has been contrasted with rationalist approaches which emphasize the role of reason (e.g., deductive inferences) and deny the claim that all human knowledge originates in experience. Rationalist thinkers such as Descartes (1596-1650) and Leibniz (1646-1716) proposed that there is *a priori* knowledge, that is, knowledge independent of experience from which we can derive new knowledge. For example, Descartes' famous "*cogito, ergo sum*" was claimed to be an *a priori* truth since it is gained through reason alone and not from experience. Some philosophers in the rationalist tradition, such as Leibniz, also allowed for the possibility of innate ideas.

Contrary to the rationalist position, the British empiricists John Locke (1632-1704), George Berkeley (1684-1753), and David Hume (1711-1776) claimed that the ultimate source of human knowledge is sense experience, not reason. The empiricists also denied the existence of any innate knowledge (“innate ideas”), a position vividly expressed in Locke’s notion of the *tabula rasa*. In its most radical version, the empiricists’ position was that all our knowledge is *a posteriori*, that is, directly derived from experience. Since all knowledge depends upon sense experience, the empiricists’ position implies that our causal knowledge must originate in experience, too. The question is then which of our experiences can give rise to knowledge of causal relations, one of the main issues Hume addressed in his writings.

Hume divides the mental realm into thoughts (“ideas”) and perceptions (“impressions”) which provide our mind with experience. According to Hume, even the most elaborated and abstract concepts (“complex ideas”) stem from, and are reducible to, atomic pieces of knowledge (“simple ideas”). These simple ideas, in turn, originate in the content of our experience (Hume’s so-called “copy thesis”). In his *Treatise of Human Nature* (1739/2000) Hume argues that “(…) all our simple ideas in the first appearance are deriv’d from simple impressions, which are correspondent to them, and which they exactly represent” (p. 9).

Embedded in Hume’s epistemological atomism is his analysis of causality. According to Hume, causal knowledge is inductive, not deductive. For example, when we encounter a new object of which we have no knowledge, we cannot discover its causal history or its causal powers deductively. Thus, we are not capable of determining an object’s causes or effects by reason alone. Therefore, Hume concluded, knowledge of causal relations must be derived from experience. The problem he then faced was that the sensory input, our ultimate source of knowledge, does not contain any direct causal knowledge. Even though every event is a cause or an effect (or both), there is no feature (“quality”) common to all events which are kinds of cause and effect: “And indeed there is nothing existent, either externally or internally, which is not to be consider’d either as cause or an effect; tho’ ‘tis plain there is no one quality, which universally belongs to all beings, and gives them a title to that denomination. The idea, then, of causation must be deriv’d from some *relation* among objects; and that relation we must now endeavour to discover.” (Hume, 1739/2000, p. 53, his italics).

Hume’s task then was to determine which experiences give rise to the idea of causation. He proposed that causal relations are characterized by three features which are contained in our perceptions and can serve as sensory input to the process of causal induction. First, events of cause and effect are contiguous in space and time. This relation is that of *spatio-temporal contiguity*. Second, events of cause and effect are temporally ordered since causes always precede their effects. This relation is that of *temporal succession.* However, two events might be contiguous and temporally ordered without being causally connected; therefore, contiguity and temporal priority are not sufficient to give rise to the idea of a causal relation. Hume argued that there is a third relation connecting causes and events and it is this relation that is essential to the idea of causation. Consistent with many other philosophers, Hume saw the impression of a "*necessary connexion*" to be the fundamental feature of cause-effect relations. A necessary connection between two events implies that the cause necessitates the effect. Since the cause is necessarily followed by the effect, observing the cause allows us to predict the presence of the effect event. It is in virtue of the acquaintance of this relation that we are able to transcend our past experience and make predictions about events not observed or not happened yet. However, contrary to many other philosophers before and after him, to Hume causal necessity is merely a construction of the human mind and must not be expected to exist outside our experience.

The problem is then to explain what gives rise to the impression of a necessary connection between two events. In analogy to the argument that we cannot determine an object's causal powers by reason alone, Hume was convinced that we cannot logically prove the existence of a necessary connection between a cause and an effect. He proposed that it is the relation of *constant conjunction* from which we derive the idea of a necessary connection: "The idea of cause and effect is deriv'd from experience, which informs us, that such particular objects, in all past instances, have been constantly conjoin'd with each other." (Hume, 1739/2000, p. 63). If we were only confronted with single episodes in which events occur together we would never induce a causal relation. It is the repeated observation that events vary together which gives rise to the idea of a necessary connection and, eventually, generates the impression that these events are causally related. Hume did not claim that we can discover the exact nature of this connection from our sense experience but merely that we have the idea that there is such a connection. For example, we might infer from our experience that the moon is causally related to the tides even though we do not have specific knowledge of the exact nature of the underlying connection.

According to Hume, the impression of a causal relation implies that the idea of the cause event conveys the idea of the effect event. From the experienced constant conjunction of cause and effect we infer that upon the appearance of the cause the effect will follow, just as it did in the past. The crucial difference to the rationalist account is Hume's claim that such causal inferences are not a based on reasoning but on "the union of ideas": "When the mind, therefore, passes from the idea or impression of one object to the idea or belief of another, it is not determin'd by reason, but by certain principles, which associate together the ideas of these objects, and unite them in imagination. (…) The inference, therefore, depends solely on the union of ideas" (Hume, 1739/2000, p.64). Thus, Hume denied that the presence of the effect is derived deductively from the existence of a necessary connection that binds together cause and effect. Instead, causal relations are inferred inductively according to associative learning principles; the inferences are merely "habit", as he later stated in his *An Enquiry Concerning Human Understanding* (Hume, 1748/1993, p. 50). Since all inductive knowledge is fallible, he concluded, definite knowledge of causal relations lies beyond our reach.

To sum up, according to Hume's empiricist approach the acquisition of causal knowledge is determined by spatio-temporal contiguity, temporal succession, and constant conjunction. When events are repeatedly perceived to be contiguous in space and time in several instances we will induce that they are causally related. The temporal information allows us to determine which event is the cause and which is the effect. Since the information defined by these principles is contained in our sensory input, we have a well-defined account of data-driven causal induction even though our senses do not directly provide us with causal knowledge.

2.2 Critique of Hume's Principles of Causal Induction

It was Immanuel Kant (1724-1804) who in his *Kritik der reinen Vernunft* (*Critique of Pure Reason*) (1781/1974) was the first and most prominent philosopher to attack Hume's empiricist account. Kant's philosophy differed from the empiricists' position as well as from traditional rationalist approaches. On the one hand, Kant denied the empiricists' claim that all our knowledge is derived from experience and rejected the idea that the acquisition of causal knowledge is a purely inductive process. He also took issue with Hume's claim that causal necessities do not exist outside our experience and rejected the attempt to reduce causal relations to experienced regularities. On the other hand, Kant's philosophy was also at variance with traditional rationalist approaches. According to Kant, the capacity to deduce new knowledge through exercises of reason alone is limited to certain subject areas such as pure mathematics.

Kant's central concern was the question of how we can derive true knowledge from empirical observations (i.e., the possibility of "synthetic a priori knowledge" that "transcends" our past experience). He takes the view that knowledge can be acquired through experience, but argues that our experiences are not only constrained by our sense organs but also by the constitution of our cognitive faculty. Kant argues that the human mind must be endowed with general conditions ("reine Anschauungen", "pure intuitions") and certain fundamental categories of thought ("reine Verstandesbegriffe", "pure categories of the understanding") which do not originate in our experience. These concepts are necessary preconditions for coherent perceptions of the world and it is only in virtue of these cognitive structures that we can learn from experience in the first place. For example, the existence of an *a priori* spatio-temporal framework is a necessary precondition to the perception of an object as being uniquely located in space and time. We cannot decouple the representation of an object from the underlying concept of space and, for example, conceive of an object without any spatial properties.

With respect to causality, Kant agreed with Hume that causal knowledge about particular causal relations is rather inductive than deductive. However, in Kant's philosophy the *general* notion of causality is one of the pure categories of the understanding and therefore not derived from experience. Even though we might induce the existence of particular causal relations from our sense experience we cannot derive the concept of causality itself empirically. Rather, a general notion of cause and effect is a necessary prerequisite to causal induction. It is this objection that connects Kant's philosophy with the current debate on psychological models of causal induction. In the tradition of Hume, associative theories of causal learning claim that causal knowledge is essentially associative and suggest that the acquisition of causal knowledge is a purely inductive process (e.g., Dickinson, Shanks, & Evenden, 1984; Shanks & Dickinson, 1987). Other accounts such as causal model theory (e.g., Waldmann, 1996; Waldmann & Holyoak, 1992) and the power PC theory (Cheng, 1997; Novick & Cheng, 2004) also assume that covariational information is important for causal learning but emphasize the role of domain-independent causal knowledge for the process of causal induction.

Kant's critique of Hume was a fundamental one; he not only objected to Hume's analysis of causality but also refuted the empiricists' philosophical position in general. Whereas one of Kant's central concerns was whether we can derive a general concept of causality from our experiences, other philosophers have rather focused on particular problems connected with Hume's epistemology and his attempt to reduce causal knowledge to experienced regularities. Hume defined a cause as "(…) *an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second.* Or, in other words, *where, if the first object had not been, the second never existed*" (Hume, 1748/1993, p. 51, his italics). Traditionally, the first part of this statement, which refers to the criterion of regular successions, has been conceived of as the core assumption of Hume's analysis of causality. The second part is rather an alternative notion of the concept of cause based on a counterfactual definition. This definition has received considerably less attention, but based on this notion some philosophers (e.g., Lewis, 1973) have developed so-called *counterfactual theories of causality*, which refrain from the principle of constant conjunction. Other philosophers such as Mackie (1974) have objected to Hume's definition of the concept of cause and defined causes as so-called INUS conditions („*I*nsufficient but *N*ecessary parts of *U*nnecessary but *S*ufficient conditions"). According to this idea causes are always only a part of a larger set of relevant conditions which are singly necessary and jointly sufficient. Conceptualizing causes as INUS-conditions provides a much more detailed account of the meaning of the concept of cause and also acknowledges the relevance of further causally relevant factors (cf. section 3.2.2).

However, it is clearly Hume's regularity criterion and the principle of constant conjunction that has been criticized most vigorously for its several shortcomings. First, the criterion of constant conjunction is *overinclusive*. Picking up a classic example, the night is invariably followed by the day but the night does not cause the day. Thus, regular successions do not necessarily imply causal relations. This difficulty is closely related to the problem of *spurious regularities*. Returning to the example given in the introduction, we can observe that the state of a barometer repeatedly covaries with the upcoming weather. Since the events are also temporally ordered and contiguous in space and time, all of Hume's criteria are met. However, the barometer clearly does not cause the weather. The underlying problem is that Hume's simple regularity criterion is not sensitive to spurious correlations arising from common causes (cf. Reichenbach, 1956). Another challenge for Hume's criterion of constant conjunction is that of *imperfect regularities*. Smoking and lung cancer might be causally connected but this does not imply that all smokers inevitably develop the disease. Thus, causes are not always followed by their effects. Even though we do not always observe perfect regularities we are nevertheless willing to induce causal relations from probabilistic relations. Moreover, to assess the causal influence of the putative cause it is also necessary to consider instances in which the effect occurs in the absence of its cause (e.g., there are also non-smokers who suffer from cancer).

The problems with Hume's original approach led to the development of so-called *probabilistic theories of causality* (Eells, 1991; Pearl, 1988; Salmon, 1980; Suppes, 1970) which tackle several of the problems associated with the principles of causal induction proposed by Hume. In contrast to Hume's criterion of constant conjunction these approaches do not require that the cause is always followed by the effect. Instead, it is only required that causes raise the probability for the occurrence of their effects, that is, constant conjunction is replaced by probabilistic conjunction.[1] In addition, probabilistic theories of causality not only consider how often the cause is followed by the effect but also take into account instances in which the effect occurs in the absence of its cause. The importance of considering the probability of the effect when the cause is absent is nicely illustrated by Salmon's (1971) example of John Jones who has been taking birth control pills regularly and successfully fails to get pregnant.

Taken together, these two consideration can be formalized in standard probability calculus as the inequality of the two conditional probabilities $P(\text{Effect} \mid \text{Cause})$ and $P(\text{Effect} \mid \neg\text{Cause})$ (conventionally abbreviated as $P(e \mid c)$ and $P(e \mid \neg c)$). For example, observing that the effect is more likely to occur in the presence of the candidate cause than in the absence of the cause indicates a generative causal connection. Formally, this is expressed as $P(e \mid c) > P(e \mid \neg c)$. Conversely, observing $P(e \mid c) < P(e \mid \neg c)$ indicates an inhibitory relation. The difference of these two conditional probabilities is also referred to as the *contingency* ΔP of cause and effect.[2] The contingency ΔP is often considered as a normative measurement of causal strength and has also been regarded as a psychological model of natural causal induction (cf. Section 3.2).

To differentiate between spurious and genuine relations the constraint is added that the probabilistic relationship between cause and effect must not vanish when taking into account further variables (Cartwright, 1983; Eells, 1991; Reichenbach, 1956; Suppes, 1970). For example, the effects X and Y of a common cause C are spuriously correlated but become statistically independent conditional on states of their common cause. The common cause C is then said to *screen off* X from Y. Returning to the barometer

[1] This assumption concerns the epistemology of causal relations rather than their ontology. For example, with reference to quantum mechanics it has been argued that causal relations are inherently probabilistic. Other authors (e.g., Pearl, 2000) have adopted Laplace's (1814/1912) quasi-deterministic conception of causality according to which we only observed imperfect regularities because we do not have knowledge of all relevant variables (e.g., unobserved inhibitors).

[2] Note that the cause-effect contingency must not equal the effect-cause contingency.

example, the impending weather is independent of barometer readings conditional on their common cause, atmospheric pressure. This kind of conditional independence relation also plays an important role in causal Bayes nets theory introduced in Section 4.

2.3 Summary

Hume's analysis revealed the fundamental problem of causal induction: our sensory input contains no direct knowledge of causal relations. Thus, causal knowledge must be inferred from other sources of information. Hume put forward three fundamental principles which he claimed are sufficient to give rise to causal knowledge: contiguity, temporal succession, and constant conjunction. In accordance with his philosophical orientation this information, which is contained in our sense experience, defines causal induction as a purely data-driven process. Even though the details of his approach have been criticized for several reasons, the idea that we can infer non-observable causal relations from observable covariations has strongly influenced research in philosophy, statistics, computer science, and psychology.

Modern probabilistic theories of causality acknowledge the shortcomings of Hume's original account and take them into account by introducing the concept of contingency as a probabilistic measurement of causal relations. However, the attempt to reduce causal relations to probabilistic regularities also has its problems because it neglects important features of causal relations. For example, whereas statistical relations are symmetric, a general feature of all causal relations is that of *causal directionality*: causes generate their effects but not vice versa.[3] This asymmetry has major consequences for our ability to control our environment. Intervening in the cause event will influence the effect event but intervening in an effect will not change the probability of its cause. For example, drinking alcohol will make you more likely to suffer from headache but producing headache by hitting on your head will probably not make you more likely to drink alcohol. Thus, even though probabilistic theories of causality have provided important insights into the analysis of causal relations the attempt to define causal relations only by means of standard probability calculus remains incomplete.

[3] In fact, the physicist and philosopher Hans Reichenbach (1956) has argued that the direction of time can be derived from the asymmetry of causal relations and the irreversibility of certain physical processes (e.g., thermodynamical processes).

3 Psychological Theories of Causal Cognition

One fundamental question has not changed since the time of Hume and Kant: which kind of information enters the process of natural causal induction? In the tradition of Hume most psychological theories of causal cognition model the process of causal induction as a purely data-driven process and consider covariational information to be the primary means by which we infer the presence of causal relations. In the literature, different kinds of rule-based contingency models have been contrasted with associative accounts (for overviews see Allan, 1993; Allan & Tangen, 2005; Shanks, 1993). Even though the two approaches differ with respect to the way the covariational information is assumed to be processed, they both propose that covariational information is the primary means by which we infer the existence of causal relations.

Whereas many accounts consider causal learning to be a bottom-up process, other authors have argued that covariational information is not the only source of information that enters the process of causal induction. These theories also consider covariation as an important cue to causality but also emphasize the importance of prior knowledge for causal learning. Some researchers have investigated the influence of domain-specific knowledge, such as assumptions about underlying mechanisms (e.g., Ahn, Kalish, Medin, & Gelman, 1995; Koslowski, Okagaki, Lorenz, & Umbach, 1989). Other theories propose that it is rather abstract knowledge such as knowledge of causal directionality that strongly influences causal learning (e.g., Waldmann, 1996; Waldmann & Hagmayer, 2001; Waldmann & Holyoak, 1992).

In general, the models of causal induction discussed in the following section have been applied to both observational learning and interventional learning (cf. Shanks, 1993). The former refers to situations in which learners passively observe cause-effect relations, whereas the latter involves an active participation. For example, a causal learning experiment could employ a medical scenario in which participants are requested to evaluate how different kinds of foods (the candidate causes) are related to allergic symptoms (the effects) in animals. A case of observational learning is then that learners simply receive information of the kind “an animal has eaten food X and effect Y has happened”. In contrast, an interventional learning design would ask participants to actively feed an animal with a certain kind of food to find out whether some allergic reaction results.

3.1 Associative Theories of Causal Induction

Associative learning theories have a long tradition in research on animal learning. Originally developed to explain the acquisition of conditioned reactions and instrumental behavior, associative learning models have also been applied to higher level processes such as category learning (e.g., Gluck & Bower, 1988) and causal induction (e.g., Chapman, 1991; Chapman & Robbins, 1990; Dickinson, 2001; Dickinson & Burke, 1996; Dickinson & Shanks, 1995; Dickinson et al., 1984; Lopez, Shanks, Almaraz, & Fernandez, 1998; Shanks, 1985; Shanks & Dickinson, 1987; Shanks, Lopez, Darby, & Dickinson, 1996; Vallée-Tourangeau, Murphy, Drew, & Baker, 1998; Van Hamme & Wasserman, 1993, 1994; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996). According to associative theories, causal learning is basically the same as learning cue-outcome relations since both tasks are characterized by multiple-cue contingency learning in which a common associative learning mechanism is assumed to operate. Applied to human causal induction it is proposed that causal learning consists of associating 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.

In accordance with Hume's original approach, early theories of associative learning assumed that spatio-temporal contiguity is sufficient to learn associations between cues and outcomes (see Domjan, 2003, for an overview). However, Rescorla (1968) showed that the acquisition of conditioned reactions not only depends on the number of instances in which the cue is followed by the outcome (e.g., a tone followed by a shock) but also on the number of trials in which the outcome occurs without the cue. Rescorla discovered that with a fixed number of cue-outcome pairings the strength of a conditioned reaction decreased as a function of the probability with which the outcome occurred in the absence of the cue. Thus, associative strength was not only a function of $P(\text{Outcome} \mid \text{Cue})$, as proposed by contiguity-based approaches, but also of $P(\text{Outcome} \mid \neg\text{Cue})$.

Another finding at variance with contiguity-based theories are cue interactions such as the *blocking effect* (Kamin, 1968). A typical blocking experiment involves two learning phases. In the first phase, a cue *C* (e.g., a tone) is constantly followed by an outcome *E* (e.g., a shock) until *C* elicits a conditioned reaction (e.g., a fear reaction). In a subsequent learning phase a second cue *X* (e.g., a light) is introduced which is always presented together with *C* and followed by *E*. Since in this phase cue *X* is also constantly paired with the outcome, contiguity based theories predict that *X* should become associated with the outcome. However, when tested on cue *X* alone very little response was observed. This cue interaction effect is referred to as blocking since the previous pairing of cue *C* with the outcome prevents the acquisition of associative strength between cue *X* and outcome *E* in the second learning phase. This finding, too, is at variance with contiguity-based theories.

3.1.1 The Rescorla-Wagner Model

The studies of Rescorla (1968) and Kamin (1968) made it necessary to revise traditional associative theories which had considered contiguity to be sufficient for the acquisition of associative strength. One prominent model is the Rescorla-Wagner model (Rescorla & Wagner, 1972), probably the best known and most influential model formalizing the acquisition of associative knowledge (see Miller, Barnet, & Grahame, 1995, for an overview). The Rescorla-Wagner model (henceforth R-W model) provides a discrepancy based learning rule which has not only been applied to animal learning but also claimed to provide an account of human causal induction (e.g., Shanks & Dickinson, 1987).

The R-W model requires binary cause and effect events which are assumed to be present or absent. The model also postulates that there is an always-present background cue *A* which can be thought of as representing unobserved alternative causes. In addition, the learning process is divided in discrete time steps ("trials"). According to the R-W model, on each trial the association of cause and effect is modified according to the discrepancy between the expected and the observed state of the outcome. For example, in trials in which the cause is followed by the effect the associative weight is increased. Conversely, when the cause is present but the effect is absent, the associative weight is decreased. Thus, the associative strength between the cause and the effect after trial $t + 1$ is a function of the existing weight and the computed discrepancy (the "error"), that is, $V_i^{t+1} = V_i^t + \Delta V_i^t$. Formally, the (positive or negative) change in associative weight, ΔV_i^t, is given by

$$\Delta V_i^t = \begin{cases} 0 & \text{if the cause is absent} \\ \alpha_i \beta_1 \left(\lambda - \Sigma V_j \right) & \text{if both the cause and the effect are present} \\ \alpha_i \beta_2 \left(0 - \Sigma V_j \right) & \text{if the cause is present but the effect is not} \end{cases} \tag{1}$$

where λ is the value of the outcome (normally assumed to be 1 for trials in which the effect is present and 0 when the effect is absent) and also indicates the maximum associative strength supported by the outcome. ΣVj is the sum of associative strength of causes A, $C_1,\ldots,C_n$ present on that trial. Thus, the expected outcome ΣV_j is an additive function of the causes present in that trial and their associative weights. Parameters α_i, β_1, and β_2 are so-called "learning rates" assumed to reflect the salience of the cause(s) and the effect. The associative weight between cause and effect is incremented or decremented according to the learning algorithm formalized in equation (1).

The R-W model can account for several phenomena which contiguity based learning theories cannot explain. For example, Rescorla's (1968) finding is accounted for, since in trials in which the outcome occurs in the absence of the actual cue the always-present background cue gains associative strength which, in turn, contributes to ΣV_j. As a consequence the prediction error and therefore also the associative strength acquired by the actual cue decreases the more often the outcome occurs without the cue. Thus, the acquired causal strength is not only a function of the probability of the outcome in the presence of the cue, but also of the probability of the outcome occurring when the cue is absent. The R-W model also accounts for a variety of cue interaction effects. For example, the model explains the blocking effect because in the first learning phase cue C is established as a perfect predictor of the outcome (i.e., $\Delta V = \lambda - \Sigma V_j \approx 0$). Since in the subsequent learning phase the presence of C perfectly predicts the effect, no error occurs and therefore the redundant cue X cannot acquire any associative strength. However, even though the R-W model successfully explains many phenomena of animal learning, there are also results which are inconsistent with the model (cf. Miller et al., 1995).

A number of researchers (e.g., Sutton & Barto, 1981) have pointed out that the R-W model is formally equivalent to Widrow and Hoff's (1960) *delta rule*. Since the delta rule can been used to train simple connectionist networks (e.g., Gluck & Bower, 1988), these models are also sensitive to learning phenomena such as the blocking effect. However, the equivalence of the Widrow-Hoff rule and the R-W model depends on the chosen parameters and thus cannot readily be generalized to all combinations of parameters even though the basic idea (error correction) is identical in both models (see Danks, 2003, for a detailed analysis).

To corroborate the claim that causal learning can be accounted for by the R-W model, it has been investigated whether human causal learning is subject to similar conditions as animal learning. For example, it has been demonstrated that estimates of causal strength decreased the longer the temporal delay between the cause (tapping a key) and the effect (an illumination of a figure on a computer screen) (Shanks & Dickinson, 1991; Shanks, Pearson, & Dickinson, 1989). This result is consistent with studies in animal learning showing that both the acquisition of conditioned reactions and instrumental behavior is affected by the temporal delay between cue and outcome.

Since cue interaction effects have been considered a hallmark of associative learning theories, many experiments have investigated whether similar phenomena also occur in human causal learning. In fact, cue interaction effects such as blocking have also been found in studies on human causal induction (e.g., Chapman & Robbins, 1990; Shanks, 1985). Other studies have provided evidence for overshadowing effects, another phenomenon well-known from research on animal learning (cf. Domjan, 2003). Overshadowing occurs in situations in which two simultaneously presented cues (e.g., a tone and a light) are followed by an outcome (e.g., a shock). It has been found that the cues receive lower associative weights (i.e., elicit weaker reactions) when presented simultaneously than when learned separately. The R-W rule explains this finding since the predicted outcome is an additive function of the cues present. Thus, when the cues are trained separately, each of them can gain the maximal associative strength supported by λ. In contrast, if the cues are presented simultaneously they can only gain half of the associative strength (provided they have equal learning rates). This effect has also been found to occur in causal learning (e.g., Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Price & Yates, 1993).

Critique of Associative Theories of Causal Induction

Learning procedures such as the R-W rule are sensitive to covariations and provide a detailed account of how covariational information is processed. However, cue and outcome may covary because they are directly causally related or because they are spuriously correlated. Associative theories neglect that identical patterns of covariation might arise from very different causal structures: the cue and the outcome may covary because there is a direct causal relation, because they are both effects of a common-cause, or because they are part of a causal chain. Models such as the R-W rule provide no means to represent causal structure, which is at variance with findings demonstrating that learners' assessment of covariational information is influenced by hypotheses about the underlying causal structure (Waldmann, 1996, 2000, 2001; Waldmann & Hagmayer, 2001; Waldmann & Holyoak, 1992).

The failure to represent causal structure is also due to the problem that associative theories fail to take into account the asymmetry of causal relations: causes generate their effects but not vice versa. Associative models do not represent causal directionality but only use event types of cue and outcome irrespective of their causal roles. However, these event categories do not adequately reflect the asymmetry of cause and effect, which, for example, is crucial when we want to intervene to bring about (or prevent) certain events. While the mapping is superficially justified by their temporal equivalence (i.e., cues precede outcomes and causes precede their effects), associative accounts are challenged when *experienced* temporal order does not match causal order. For example, it has been demonstrated that the occurrence of the blocking effect depends on the causal status of the cues, that is, whether the events observed first (i.e., the cues) are assumed to be causes or effects (Waldmann, 2000, 2001; Waldmann & Holyoak, 1992; Waldmann & Walker, 2005). However, since there are also studies in which no effect of causal status on blocking was found (Cobos, López, Cano, Almaraz, & Shanks, 2002), it recently has been argued that associative bottom-up and knowledge-based top-down processes interact with each other (Allan & Tangen, 2005; Tangen, Allan, & Sadeghi, 2005).

There are also other cue interaction effects in causal learning which are incompatible with the Rescorla-Wagner model. For example, *retrospective revaluation effects* are problematic for the model. According to the R-W model, only associative weights of cues present are modified. Inconsistent with this assumption it has been found that the associative strength of a cue might also be modified in its absence, for example in *backward blocking* (e.g., Shanks, 1985). Backward blocking is obtained when the two learning phases of the standard blocking design are reversed. For example, participants first observe that two causes C and X (which always occur together) are followed by an effect E. In the second learning phase, cause C is presented alone with the effect. According to the R-W rule, X's associative weight should not be affected by the second learning phase (cf. equation (1)). However, it has been demonstrated that learners discount the causal strength of cause X after observing that C alone is sufficient to generate the effect (e.g., Chapman, 1991; Larkin, Aitken, & Dickinson, 1998). Further evidence for retrospective evaluation effects has been provided by de Houwer and Beckers (2002a, 2002b). Since the standard R-W model cannot account for these findings modifications have been proposed which allow for modifications of associative strength in the absence of the cue (Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994).

Finally, there is also evidence that retrospective evaluations influence forward blocking. For example, it has been demonstrated that causal judgments about a to-be-blocked cue *X* are affected by retrospective inferences about its status during the first learning phase (De Houwer, 2002). Causal judgments differed depending on whether learners inferred from the second learning phase that cue *X* was really absent in the first learning phase or whether the state of *X* could only not be observed during the first learning phase. This result is not only inconsistent with the standard R-W model but also problematic for revised versions of the model which explicitly represent absent cues (Van Hamme & Wasserman, 1994) or explain backward blocking by assuming within-compounds associations (Dickinson & Burke, 1996).

3.2 Rule-based Accounts of Causal Induction

Rule-based accounts of causal induction assume that humans act as "intuitive statisticians"; estimates of causal strength are assumed to reflect the contingency of a causal relation (for reviews see Allan, 1993; Shanks, 1993; Shimazaki & Tsuda, 1991). These theories provide computational level descriptions[4] specifying which covariational information serves as input to the process of causal induction and how this information is integrated to derive causal judgments. According to these accounts, causal learning is primarily data driven but no particular reference is made to the underlying algorithmic processes. However, under certain conditions some models (e.g., the ΔP-rule) are consistent with associative learning procedures (e.g., the Rescorla-Wagner model) since the associative weights asymptotically approach the cue-outcome contingency ΔP (Chapman & Robbins, 1990; Cheng, 1997; Danks, 2003; Wasserman, Elek, Chatlosh, & Baker, 1993).

Rule-based approaches assume that learners induce causal relations from the joint frequency distributions of the cause and the effect variable. The joint frequency distribution of discrete variables is often represented as a *contingency table* with each cell referring to a specific combination of cause and effect (see Table 1). For example,

[4] Marr (1982) suggested that cognitive systems should be analyzed on three levels of descriptions. The *computational level* describes abstractly which function is being computed to solve a given problem (e.g., the contingency ΔP). The *algorithmic level* specifies the steps carried out to compute the function described on the computational level (e.g., how frequency information is processed). Finally, the *implementational level* specifies the physical properties of the underlying information processing system (e.g., the neurobiological foundations).

when both the cause and the effect might be observed to be present or absent (which is the standard experimental paradigm) a 2x2 contingency table results. Corresponding contingency tables can be constructed for more than two variables and/or more than two possible states. Conventionally, the cells of a 2x2 contingency table are labeled as *a*-, *b*-, *c*-, and *d*-cell denoting the relative frequencies of event co-occurrences.

Table 1
2x2 Contingency Table

	Effect present	Effect absent
Cause present	*a*	*b*
Cause absent	*c*	*d*

3.2.1 The Δ*P*-Rule

The oldest and most prominent model of this sort is the Δ*P*-rule (Allan & Jenkins, 1980; Allan & Jenkins, 1983; Ward & Jenkins, 1965). According to this model, learners' causal judgments are a monotonic function of the statistical contingency Δ*P* which is assumed to be estimated from frequency information. In terms of the cell entries of a 2x2 contingency table, the contingency of cause and effect is given by

$$\Delta P = P(e \mid c) - P(e \mid \neg c) = \frac{f_a}{f_a + f_b} - \frac{f_c}{f_c + f_d} \qquad (2)$$

The Δ*P*-rule produces values ranging from -1 to +1 with positive values indicating the presence of a generative causal relation and negative values indicating an inhibitory relation. According to the simple contingency model there is no need to differentiate explicitly between causal structure and causal strength, because the absence of a causal relation is indicated by a zero contingency.

The empirical evidence for the use of the Δ*P*-rule is mixed. Some studies have provided evidence that learners' causal judgments reflect the contingency Δ*P* (e.g., Ward & Jenkins, 1965; Wasserman, Chatlosh, & Neunaber, 1983; Wasserman et al., 1993), but also considerable deviations have been found. The use of the rule seems to be affected by factors such as the employed response format (Allan & Jenkins, 1980; Wasserman et al., 1983), the way the learning data is presented (Allan & Jenkins, 1983; Kao & Wasserman, 1993), the overall probability of the effect (so-called "density bias", Allan & Jenkins, 1983; Dickinson et al., 1984; Wasserman et al., 1993), and developmental stages (Shaklee & Mims, 1981).

Another finding inconsistent with the simple contingency model is the so-called "*a*-cell bias". A common finding is that the instances of the four cells are weighted in the order *a*-cell > *b*-cell ≥ *c*-cell > *d*-cell, a result at variance with the assumptions of the

Δ*P*-rule (Kao & Wasserman, 1993; Schustack & Sternberg, 1981; Wasserman, Dorner, & Kao, 1990). Models that can encompass the *a*-cell bias are linear regression models (e.g., Schustack & Sternberg, 1981) or a weighted Δ*P*-rule (e.g., Wasserman et al., 1993).

Further research has indicated that learners also tend to use alternative strategies inconsistent with the Δ*P*-rule (cf. Allan & Jenkins, 1983; Kao & Wasserman, 1993; Shimazaki & Tsuda, 1991). Especially the "sum of diagonals-strategy" (also called Δ*D*-rule) has been frequently found to be used by participants. According to this rule, learners contrast the number of confirming instances (i.e., *a*-cell and *d*-cell) with the number of disconfirming instances (i.e., *b*-cell and *c*-cell):

$$\Delta D = (f_a + f_d) - (f_b + f_c) \tag{3}$$

A proportional variant of the sum of diagonals-strategy is the *evidential evaluation model* (White, 2002; White, 2004). This model assumes that learners derive their causal judgments in accordance with the number of instances that confirm their causal hypotheses, an idea formalized by the so-called *pCI* (*p*roportion of *c*onfirming *i*nstances)-rule:

$$pCI = \frac{(f_a + f_d - f_b - f_c)}{(f_a + f_b + f_c + f_d)} \tag{4}$$

Like the Δ*P*-rule, the *pCI*-rule generates values ranging from -1 to +1. Contrary to the inductive inference process proposed by the Δ*P*-rule, the evidential evaluation model proposes that learners analyze and encode contingency information in terms of confirmatory and disconfirmatory evidence. Causal judgments are assumed to be a function of the proportion of confirmatory and disconfirmatory instances.

Critique of the ΔP-Rule

As pointed out by several authors (Cartwright, 1983; Cheng & Novick, 1992; Melz, Cheng, Holyoak, & Waldmann, 1993; Spellman, 1996) a fundamental shortcoming of the Δ*P*-rule is that the model does not allow for taking into account further variables. For example, in situations with multiple causes, the contingency for each event is computed by collapsing over the alternative causes. However, computing the contingency over the universal set of events is not appropriate because the unconditional contingency also includes the influence of alternative confounding causes. Similarly, if two events are only spuriously related because they are both effects of a common cause, the model offers no possibility to take into account the common cause event. Therefore, the model cannot distinguish causal relations from spurious correlations even in situations with a known common cause event.

A further problem of the ΔP-rule is that of *ceiling effects* (cf. Cheng, 1997). Ceiling effects refer to situations in which the effect is always present when the candidate cause is present, but also when the candidate cause is absent (i.e., $P(e \mid c) = 1$ and $P(e \mid \neg c) = 1$). The simple contingency model would indicate that there is no causal relation between the two events (because $P(e \mid c) - P(e \mid \neg c) = 0$), whereby intuitively we have a situation in which a generative cause cannot exhibit its causal powers, and therefore should neither be judged as causal nor non-causal.

3.2.2 The Conditional ΔP-Rule

According to the unconditional ΔP-rule the cause-effect contingency is computed across the universal set of events. Alternative causes are assumed to occur independently of the candidate cause; their influence on the effect event is only represented by the number of instances in which the effect occurs in the absence of the candidate cause. However, in multiple cue environments it is often necessary to control for the influence of alternative causes to give unconfounded estimates of causal strength and to detect spurious correlations. Therefore, several authors have argued for a *conditional contingency model* (e.g., Cartwright, 1983; Eells, 1991; Melz et al., 1993; Spellman, 1996; Suppes, 1970; Waldmann, 1996; Waldmann & Holyoak, 1992).

The conditional contingency model makes it possible to assess the contingency of a candidate cause conditional on states of other events. By holding constant the potentially relevant factors A_1 to A_n, learners can derive estimates of causal strength relative to a certain causal background. This idea is formalized in a modified version of the ΔP-rule:

$$\Delta P_{cond} = P(e \mid c.\, a_1 \ldots a_n) - P(e \mid \neg c.\, a_1 \ldots a_n)^{5} \qquad (5)$$

By conditionalizing on the absence of alternative causes the conditional ΔP-rule allows for unconfounded measurements of causal strength. For example, the effects of a new drug could be influenced by a person's gender. The contingency should then be computed for both men and women separately to control for this potential confound. Conditional contingencies are also sensitive to spurious correlations arising from further

[5] $P(e \mid c.\, a)$ denotes the probability of e given c *and* a., that is, the "." symbolizes the conjunction of events c and a.

variables (provided they can be observed). Returning to the barometer example, the state of the barometer and the weather are no longer correlated conditional on the air pressure. Thus, the basic logic of equation (5) corresponds to the principle of controlling for alternative causes in experimental designs.

A variant of the conditional contingency model is the *probabilistic contrast model* (Cheng & Novick, 1991, 1992). It is assumed that the context and existing knowledge determine which factors are held constant to evaluate a hypothesized causal relation. Cheng and Novick coined the term *focal sets* to refer to subsets of events in which alternative factors are held constant. The contingency of two variables is not computed over the universal set of events but in a chosen focal set in which learners judge the candidate cause to be independent in probability of alternative causes of the effect. Thus, focal sets are defined psychologically. The use of focal sets also offers the possibility to distinguish between causes and enabling conditions (cf. Cheng & Novick, 1991, 1992).

In a series of studies (Cheng & Novick, 1990, 1991, 1992; Melz et al., 1993; Spellman, 1996; Waldmann & Hagmayer, 2001) it has been shown that learners control for alternative causes and use conditional contingencies to evaluate hypothesized causal relations. The conditional contingency model has also been used to give an alternative interpretation of the blocking effect (Waldmann & Holyoak, 1992). According to associative theories, blocking occurs because the previously paired cue *C* has already acquired sufficient associative strength to predict the outcome which prevents the to-be-blocked cue *X* from gaining associative strength. From the perspective of the conditional contingency model the finding that *X* receives lower ratings compared to a control condition in which *X* is observed to occur independently of *C* is rather assumed to reflect learners' uncertainty about the causal status of *C* and not differences in associative strength. Participants cannot give specific estimates of causal strength because they never experience what happens when *C* occurs in the absence of *X*. Thus, learners cannot estimate *X*'s conditional contingency since $P(e \mid x. \neg c)$ is not defined.

Critique of the Conditional Contingency Model

The fundamental challenge for the conditional contingency model is to give an account of how we should select the factors to control for. In principle, there is an infinite number of factors we could conditionalize on. The advice only to control for factors which are potentially relevant leads to a vicious circle since such an approach presupposes that we already have causal knowledge (cf. Cartwright, 1983; Waldmann &

Hagmayer, 2001). Nancy Cartwright once vividly summarized this problem as "no causes in, no causes out".

A further problem is that conditional contingencies crucially depend on the way the subsets are constructed (Cartwright, 1983). The problem is that different partitions of a set of events can yield different contingencies, a statistical phenomenon discovered by Pearson (Pearson, Lee, & Bramley-Moore, 1899) which is nowadays generally referred to as Simpson's paradox (Simpson, 1951). A contingency that holds in the universal set of events can change, disappear, or even be reversed depending on the way the set is partitioned (see Cartwright, 1983, for a real-world example). Even though Simpson's paradox is an extreme case, it nicely illustrates how conditional contingencies depend on the chosen reference set.

A further point is that the general rule of conditionalizing on causally relevant factors is not always appropriate. As pointed out by Waldmann and Hagmayer (2001), it is important to consider the underlying causal structure. For example, if C and E are assumed to be only spuriously correlated due to a common cause X (i.e., $C \leftarrow X \rightarrow E$) it is appropriate to conditionalize on X to evaluate the influence of C on E. The conditional contingency model will correctly indicate that C and E are only spuriously correlated. However, consider a situation in which C exerts its influence on E through an intermediate event X, that is, we have a causal chain $C \rightarrow X \rightarrow E$. Even though X is clearly a causally relevant event it is not appropriate to conditionalize on X since C and E become statistically independent conditional on values of X. Thus, if learners conditionalize on X in accordance with the general strategy of holding constant causally relevant factors they would erroneously conclude that there is no causal relation between C and E. Waldmann and Hagmayer (2001) showed that learners indeed are sensitive to this crucial difference and adjusted for alternative events in accordance with the hypothesized causal model. Of course, learners might be wrong about the hypothesized causal model, but their selections were shown to be normative relative to their causal beliefs.

3.2.3 The Power PC Theory

The *power PC Theory* (Cheng, 1997; Novick & Cheng, 2004) combines the covariational approach with the notion of causal power (Cartwright, 1989). The causal power p_X of an event denotes its capacity to produce an effect: "*Causal Power* (…) is the intuitive notion that one thing causes another by virtue of the power or energy that it exerts over the other" (Cheng, 1997, p. 368, her italics). Even though Cheng agrees with other accounts that covariational information is the key to the process of causal induction she assumes that "(…) people do not simply treat observed covariations as equivalent to causal relations; rather, they interpret and explain their observations of covariations as manifestations of the operation of unobservable causal powers, with the tacit goal of estimating the magnitude of these powers." (Cheng, 1997, p. 369). According to Cheng, the general idea of causal powers is not derived from experience but is *a priori*. This domain-independent knowledge enters the process of causal induction in the form of variables which learners seek to estimate. Causal judgments are assumed to be functions of learners' estimates of causal power.

The power PC model gives a formal account of how estimates of causal power can be derived from covariational information. Cheng's analysis applies to situations with a candidate cause *C* and (known or unknown) alternative cause represented as a composite *A*. In addition, the events involved must be represented by discrete variables which can be present or absent. In accordance with the probabilistic contrast model (Cheng & Novick, 1990), it is assumed that causal relations are evaluated in a chosen focal set (cf. Section 3.2.2). According to the power PC model, the overall probability of the effect depends on the base rates of its causes and their causal powers. The notion of causal power is formalized by introducing parameters representing causal power to standard probability calculus. The probability of an effect *E* to occur is then given by:[6]

$$P(e) = P(c) \cdot p_c + P(a) \cdot p_a - P(c) \cdot p_c \cdot P(a) \cdot p_a \tag{6}$$

Equation (6) states that the overall probability of the effect is a function of the probability of the causes' base rates (i.e., $P(a)$ and $P(c)$) and their causal powers (i.e., p_c and p_a), minus their intersection. Provided the causes occur independently, the probability of the effect conditional on the candidate cause yields

$$P(e \mid c) = p_c + P(a \mid c) \cdot p_a - p_c \cdot P(a \mid c) \cdot p_a = p_c + P(a) \cdot p_a - p_c \cdot P(a) \cdot p_a \quad \text{and} \tag{7}$$

$$P(e \mid \neg c) = P(a \mid c) \cdot p_a = P(a) \cdot p_a \tag{8}$$

Equation (7) states that when *C* is present the probability of the effect is determined by i) the causal power of the candidate cause (i.e., p_c), ii) the probability of the alternative cause to occur (i.e., $P(a)$), and iii) the causal power of the alternative causes (i.e., p_a). Conversely, when *C* is observed to be absent, the probability of the effect is

[6] The following equations apply to generative causes. See Cheng (1997) for details on the computation of causal powers for inhibitors.

determined by the probability and causal power of the alternative cause alone (equation (8)). According to equations (7) and (8) the computation of the conditional probabilities $P(e \mid c)$ and $P(e \mid \neg c)$ includes non-observable parameters representing the causal power of the observed events. Substituting equations (7) and (8) into the standard contingency formula (equation (2)) and simplifying yields

$$\Delta P = p_c \cdot (1 - P(e \mid \neg c)) \qquad (9)$$

Now, by rearranging formula (9) the theoretical entity of causal power p_c can be estimated from observational data:

$$p_c = \frac{\Delta P}{1 - P(e \mid \neg c)} = \frac{P(e \mid c) - P(e \mid \neg c)}{1 - P(e \mid \neg c)} \qquad (10)$$

Since all parameters on the right-hand side of equation (10) can be estimated from observable frequency information, the unobservable causal power of an event can be estimated from covariational information. Formula (10) states that the contingency ΔP is only an appropriate estimate of causal power when no alternative causes influence the effect. Thus, if $P(e \mid \neg c) = 0$, then $p_c = \Delta P = P(e \mid c)$ holds.

According to the power PC model, causal judgments are not determined by the contingency alone but also by the probability $P(e \mid \neg c)$. With a fixed contingency causal power increases with the number of instances in which the effect occurs in the absence of the candidate cause, a prediction confirmed in a series of experiments by Buehner, Cheng, and Clifford (2003). This finding is also in accordance with the outcome density bias (i.e., the finding that causal judgments are affected by the overall probability of the effect).

The power PC model also makes predictions about the boundary conditions of causal induction. For example, when the effect is always present in the absence of the cause (i.e., $P(e \mid c) = P(e \mid \neg c) = 1$) causal power is not defined because the denominator is zero. Therefore, the model formalizes the intuition that we cannot evaluate the causal power of a generative cause if the effect is constantly present. Indeed, there is empirical evidence that when the effect is always present learners consider covariational data as insufficient to make judgments about a putative cause (Wu & Cheng, 1999).

Critique of the Power PC Model

The power PC model is proposed as both a normative and descriptive model of causal induction. However, the model has been criticized both on grounds of empirical evidence and theoretical analyses.

As noted, the experiments by Buehner et al. (2003) demonstrate that learners take into account the probability of the effect in the absence of the candidate cause. As anticipated by the power PC model, with a fixed contingency learners' causal judgments varied depending on the magnitude of $P(e \mid \neg c)$. However, there is also evidence that causal judgments of non-contingent causes are affected by $P(e \mid \neg c)$, a finding at variance with the power PC model (Buehner et al., 2003; Lober & Shanks, 2000; Vallée-Tourangeau et al., 1998). However, this finding also challenges all other rational models of causal induction.

Recently, theoretical aspects of the power PC theory have been criticized (Luhmann & Ahn, 2005; White, 2005). Luhmann and Ahn (2005) have provided a detailed analysis of the assumptions of the power PC model (cf. Cheng, 1997, p. 373). According to their analysis, the conditions necessary to derive estimates of causal power from observable information are rarely met and, therefore, the model is too restrictive to provide an adequate account of causal induction. In addition, the power PC model tacitly assumes that causal powers are inherently probabilistic (i.e., the capacity of a cause to produce an effect is not only probabilistic because of unobserved inhibitors), an assumption Luhmann and Ahn claim to be at variance with people's intuition about causality.

While Luhmann and Ahn focus on the assumptions necessary to derive causal power from regularity information, White (2005) criticizes the claim that the power PC model successfully integrates regularity theories with the notion of causal power. In the power PC theory, causal powers are defined as the probability with which one event, the cause, produces another event, the effect. According to White, Cheng's definition of causal power is incompatible with traditional power theories which assume causal powers to be stable properties grounded in the physical nature of the entities involved (e.g., Harré & Madden, 1975). Therefore, he argues, the power PC model is incomplete and falls short of reconciling the rivaling regularity and power views.

3.3 Is Covariation all there is to Causal Induction?

The theories reviewed so far all agree with Hume's original account in that they consider statistical cues as the primary means by which we acquire causal knowledge. Even though the theories differ with respect to the way the covariational information is processed, what they have in common is that causal learning is regarded as a data-driven bottom-up process. One notable exception is Cheng's (1997) power PC model which assumes that statistical information is only a means by which we try to estimate non-observable causal powers. However, the computational mechanism which lies at the heart of the power PC theory (cf. equation (10)) also only considers covariational information as input to the process of causal induction.

Conventionally, most of the studies testing the adequacy of the rivaling models have focused on the question of how we learn about single causal relations from statistical cues. In accordance with the idea that causal induction is a purely data-driven process learners are provided with covariational information and the accuracy of their judgments is evaluated with respect to a normative-statistical criterion (e.g., ΔP) or the predictions of an algorithmic learning procedure (e.g., the R-W model). The claim that learners derive causal judgments from statistical information is corroborated by developmental studies demonstrating that even preschoolers are sensitive to covariational information (e.g., Gopnik, Sobel, Schulz, & Glymour, 2001; Shultz & Mendelson, 1975). There is also evidence that the assessment of covariational information plays an important role in real-world situations (e.g., Coups & Chapman, 2002).

The other two principles put forward by Hume, contiguity and temporal succession, are also claimed to be relevant but do not necessarily enter the process of causal induction. However, if present they can influence the assessment of causal relations as demonstrated, for example, by Michotte's (1963) classical experiments and recent studies which manipulated the temporal lag between cause and effect (e.g., Shanks & Dickinson, 1991; Shanks, Pearson, & Dickinson, 1989). There are also developmental studies providing evidence that even young children use the events' temporal ordering to determine their causal roles (Bullock & Gelman, 1979; Mendelson & Shultz, 1976). Even though spatio-temporal contiguity is not explicitly considered by all models (e.g., contingency models), the use of these cues fits neatly with the general idea that causal learning is a data-driven process.

However, a general shortcoming of such purely inductive approaches is that they neglect the influence of prior knowledge on causal induction. Typically, domain-specific knowledge has been contrasted with abstract, domain-independent knowledge. Both types of knowledge are claimed to have an impact on causal induction. For example, we can have domain-specific knowledge about certain physical mechanisms, such as that the transmission of light can be blocked by solid objects or that biological processes may take some time. Domain-specific assumptions about causal mechanisms have been shown to structure a continuous stream of events (Hagmayer & Waldmann, 2002), bridge temporal gaps between cause and effect (Buehner & May, 2002, 2003, 2004), or override covariational information (Ahn et al., 1995). Even though these results challenge covariational theories the findings must not inevitably refute covariational models. Existing domain-specific knowledge could rather be seen to impose constraints on the situations in which covariational learning mechanisms are assumed to operate. However, it has also been objected that domain-independent knowledge also influences the process of causal induction (e.g., Waldmann, 1996), which is discussed in detail in the following section.

Covariational accounts have also been criticized on theoretical grounds. For example, traditional models of causal induction lack the representational power to express the asymmetry of causal relations and provide no means to represent complex causal structures. Finally, all current models of causal induction collapse observational with interventional knowledge, thereby blurring the important distinction between merely observed states of variables and the very same states generated by external interventions.

In the following, I will discuss these aspects in detail and introduce causal model theory (Waldmann, 1996, 2000, 2001; Waldmann & Hagmayer, 2001; Waldmann & Holyoak, 1992, 1997; Waldmann & Walker, 2005) and causal Bayes nets theory (Pearl, 2000; Spirtes et al., 1993). Contrary to the inductive models discussed so far, both theories emphasize the interplay of bottom-up and top-down processes. Causal model theory can be regarded as a qualitative and psychologically plausible variant of the larger class of causal Bayes nets theories which provide a normative-statistical framework.

3.4 Causal Model Theory

Causal model theory (Waldmann, 1996, 2000, 2001; Waldmann & Hagmayer, 2001; Waldmann & Holyoak, 1992, 1997; Waldmann & Walker, 2005) emphasizes the importance of domain-independent knowledge and assumes that top-down and bottom-up processes interact in causal learning. While many studies have focused on the influence of domain-specific knowledge on causal induction, there is also abstract, domain-independent knowledge which might influence the acquisition and use of causal knowledge.

An example of domain-independent knowledge is the fact that all causal relations are inherently asymmetric: causes generate effects but not vice versa. In the spirit of Kant (1781/1974) it is assumed that our experiences are constrained by and interact with this abstract, domain-independent piece of knowledge. With reference to causal learning it is postulated that abstract causal knowledge provides the background against which we evaluate covariational information. For example, in a standard causal learning paradigm participants are provided with events precategorized as cause and effect (e.g., fertilizer and blooming). Thus, before learners are confronted with any covariational data, they already have a qualitative understanding of how the events are related to each other: a representation implying causal directionality. Data-driven accounts lack the representational power to express this directionality. For example, associative theories use cue and outcome as the two basic types of event representations with the cue defined as the event that triggers the outcome irrespective of their actual causal roles.

Contrary to covariational accounts, causal model theory (henceforth CMT) also explicitly differentiates between causal strength and causal structure. The tacit assumption of covariational approaches is that statistical information not only allows for estimates of causal strength but simultaneously provides information about the underlying causal structure (i.e., the absence of a causal relation is considered as a special case of zero causal strength). Indeed, this idea is plausible in experiments characterized by single or multiple causes which are directly related to the effect(s).

However, in both everyday and scientific causal learning we are often confronted with complex causal networks consisting of several variables. Figure 2 shows three fundamental causal structures: a common-effect model in which event *X* is generated (independently or jointly) by both *Y* and *Z*, a common-cause model in which *X* is a cause of both *Y* and *Z*, and a causal chain in which *X* causes *Y* which, in turn, causes *Z*. In principle, by combining these basic structures causal models can be constructed for

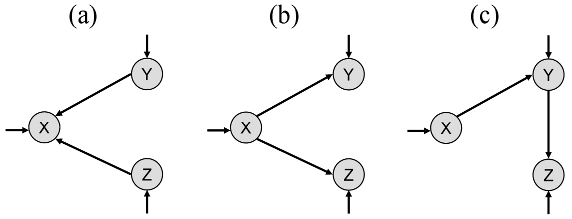


Figure 2. Basic causal models. a) Common-effect model (CE) b) Common-cause model (CC) c) Causal chain (CH)

causal relations of any level of complexity. Such causal models provide a qualitative representation of causal systems, that is, they only state the (hypothesized) existence of certain causal relations without specifying their strength. This idea fits with the intuition that we often have a rather qualitative understanding of causal relations without exact knowledge of their strength. For example, we might know that greenhouse gases affect the climate or that we might get fat from eating too much fast food without knowing the exact strength of these relations. Therefore some authors have argued for a priority of structure over strength (Griffiths & Tenenbaum, 2005; Lagnado, Waldmann, Hagmayer, & Sloman, in press; Pearl, 2000; Waldmann, 1996).

According to CMT statistical information is not treated as context-free input to the process of causal induction but is evaluated with reference to a hypothesized causal structure. Thus, covariational knowledge serves the purpose of helping to estimate a causal model's parameters. Note that causal directionality is a necessary prerequisite for representing structured models which convey more information than simply stating that certain events are correlated. For example, both the common-cause structure (Figure 2b) and the causal chain (Figure 2c) imply that events *Y* and *Z* are correlated. However, encoding causal knowledge in form of causal models also conveys the information that in the common-cause model this relation is spurious. Moreover, different causal models have statistical implications which learners can use to evaluate regularity information and decide between alternative causal models. For example, in a common effect model with independent causes (Figure 2a), events *Y* and *Z* become *de*pendent conditional on values of their common effect *X*. In contrast, the common-cause model implies that *Y* and *Z* become *in*dependent conditional on values of their common cause *X* ("so-called "explaining away" effect, in the psychological literature also known as discounting principle). Constraint-based methods of causal induction capitalize on these statistical implications to reveal causal structure from statistical information (cf. Section 4.2).

Several studies demonstrate how abstract causal knowledge influences causal learning. For example, it has been shown that causal models mediate cue interaction effects such as blocking and overshadowing. While associative accounts predict cue competition regardless of the cues' causal roles, CMT predicts cue competition only for causes but not for effect events, a prediction confirmed in a number of studies (Waldmann, 2000, 2001; Waldmann & Holyoak, 1992; Waldmann & Walker, 2005). However, there is also evidence that the effect is sometimes influenced by associative processes (Allan & Tangen, 2005; Cobos et al., 2002; Tangen & Allan, 2004; Tangen et al., 2005). It has also been shown that learners use causal models to select the events they want to control for to give unconfounded estimates of causal strength. As discussed in Section 3.2.2 it is not always appropriate to conditionalize on all causally relevant events. For example, in the common-effect structure depicted in Figure 2a it is normatively correct to conditionalize on the absence of *Y* to estimate the causal strength of the link *Z*→*X*. However, if the three events form a causal chain *X*→*Y*→*Z*, conditionalizing on the intermediate event *Y* renders *X* and *Z* independent thus erroneously indicating that the two events are not causally related. CMT predicts that learners select the variables to control for in accordance with their assumptions about the underlying causal structure. In a series of experiments, Waldmann and Hagmayer (2001) tested these predictions and demonstrated that manipulations of the suggested causal structure yield very different causal judgments derived from identical covariational information. There is also evidence that learners can integrate separately learned causal relations to more complex causal structures to predict unobserved covariations (Hagmayer & Waldmann, 2000). In addition to causal learning, CMT has also successfully been applied to categorization (Rehder, 2003a, 2003b; Rehder & Burnett, 2005; Rehder & Hastie, 2001; Waldmann, Holyoak, & Fratianne, 1995). These studies show that learners do not simply use correlated features to classify objects but take into account the internal causal structure of the entities to determine their category membership.

The main challenge for CMT is, of course, to explain how we acquire hypotheses about causal structure in the first place. One possible answer is that we use non-statistical cues such as temporal order (Lagnado & Sloman, 2004, in press), or interventions (Hagmayer, Sloman, Lagnado, & Waldmann, in press; Woodward, 2003), or that we generate hypotheses by analogy (cf. Holyoak & Thagard, 1995). Recently, algorithms have been developed which aim to uncover causal structure by analyzing the conditional dependence and independence relations found to hold in the data (Glymour & Cooper, 1999; Pearl, 2000; Spirtes et al., 1993) or use Bayesian methods to compute the likelihood of the data given a causal model (Steyvers et al., 2003). These methods are discussed in detail in Section 4.2.

4 Causal Bayes Nets Theory

Originally developed in philosophy and machine learning, causal Bayes nets theory (Pearl, 2000; Spirtes et al., 1993) has recently also attracted attention in psychology (Glymour, 2001; Gopnik et al., 2004; Hagmayer et al., in press; Lagnado & Sloman, 2004; Rehder, 2003a, 2003b; Sloman & Lagnado, 2005; Steyvers et al., 2003; Waldmann & Hagmayer, 2001; Waldmann & Hagmayer, 2005; Waldmann & Martignon, 1998). Causal Bayes nets theory combines graphical causal models and probability calculus to represent causal knowledge, to infer causal models from observations and interventions, and to formalize causal reasoning. A particularly interesting feature of this framework is the "do-calculus" developed by Judea Pearl (2000), which allows for the derivation of interventional predictions from observational knowledge. The following section gives an overview of the different aspects of causal Bayes nets theory and introduces the computational mechanisms by which the account models causal learning and causal reasoning.

4.1 Representing Causal Knowledge with Bayes Nets

Causal Bayes nets theory uses directed acyclic graphs (DAGs) to represent causal relations between variables, and parameters to express the strength of these relations (e.g., conditional probabilities). A completely parameterized causal model therefore combines qualitative assumptions about the structure of the causal model with quantitative knowledge about the parameters associated with these causal relations (e.g., base rates, causal strength, integration rules).

Formally, a graph consists of a set of discrete or continuous variables X_1 to X_n which are connected by edges.[7] If the edges are directed and the graph contains no circles it is a directed acyclic graph (DAG).[8] Thus, the three models depicted in Figure 2 are examples of DAGs. The causal arrows represent causal beliefs about the presence of (not further specified) stable causal mechanisms connecting the model's variables (Pearl, 2000; Spirtes et al., 1993; Woodward, 2003). The observable statistical dependencies among the observed events are assumed to arise from the operation of these causal mechanisms, which represent (assumptions about) invariant features of the

[7] The formalism is here described only for discrete events of cause and effect. The theory, however, can also deal with continuous variables (cf. Pearl, 2000).

[8] It is also possible to analyze cyclic graphs by the causal Bayes nets formalism. However, this requires the introduction of a time index.

world. Such graphical models provide a qualitative representation of causal knowledge which explains why we observe certain probabilistic relationships.

Associated with each DAG and its variables X_1 to X_n is a joint probability distribution $P(X_1, ..., X_n)$ encoding the probability of all possible configurations of variables. For example, with two variables C and E, the joint probability distribution consists of four data patterns (i.e., $c. e$, $c. \neg e$, $\neg c. e$, $\neg c. \neg e$), which, for instance, can be represented in a 2x2 contingency table. Such a joint distribution enables us to draw causal inferences according to the rules of standard probability calculus (e.g., to compute the probability of event X conditional on states of event Y). However, since the number of patterns grows exponentially with the number of variables and values they can take, representing causal knowledge this way is only reasonable with very few variables. For example, with only ten binary variables there are already $k^n = 2^{10} = 1024$ data patterns. Therefore, the question is how an economic representation can be achieved that facilitates causal inferences.

An alternative way to encode the joint distribution is to factorize the distribution into a product of n conditional probabilities from which the joint probabilities can be reconstructed. The product rule of probability calculus states that the joint probability of two events X and Y can always be expressed as a function of marginal and conditional probabilities, that is,

$$P(X.Y) = P(X) \cdot P(Y \mid X) = P(Y) \cdot P(X \mid Y) \qquad (11)$$

By repeated application of the product rule (i.e., the so-called chain rule) any joint probability distribution can be decomposed into a set of conditional probabilities. For example, the joint distribution of the three variables X, Y, and Z can be expressed as $P(X. Y. Z) = P(X) \cdot P(Y \mid X) \cdot P(Z \mid X. Y)$. However, this factorization will fail to provide a sparse representation because each event is conditionalized on *all* its predecessors in accordance with the chosen ordering. Moreover, the chain rule can be applied to any arbitrary ordering of events X, Y, and Z. For example, we could also choose the ordering $Z. X. Y$ and write $P(Z. X. Y) = P(Z) \cdot P(X \mid Z) \cdot P(Y \mid Z. X)$. Hence, without further constraints there is no unique decomposition of a given joint distribution.

In the causal Bayes nets framework, the joint probability distribution of a causal model is decomposed into a set of marginal and conditional probabilities by applying the *causal Markov condition* (Spirtes et al., 1993; Pearl, 2000) to the causal model. The idea is that the complete set of variables is not necessary for the computation of the probability of a variable X_i; but that a subset of variables is sufficient, namely the set of

X_i's direct causes. This set of direct causes is the set of parent nodes PA_i,which is the minimal input information required to compute the probability of node X_i taking a certain value. This set PA_i consists of what is called the *Markovian parents* of X_i, an essential concept of causal Bayes nets. According to the causal Markov condition, the state of any variable X_j in the system is then independent of all other variables (except for its causal descendants) conditional on the set of its direct causes, PA_i. Thus, for each variable X_i in the causal model the Markov condition defines a local causal process in which the state of the variable is a function of its Markovian parents. The essential point is that the graph (i.e., the causal model) provides us with a *causally based decomposition* of the joint probability distribution. Figure 3 illustrates how a joint distribution consisting of three events *X*, *Y*, and *Z* is factorized differently depending on the (hypothesized) causal model.[9]

The decomposition also specifies how the causes of a common effect combine to generate the effect (i.e., the parameter $P(x \mid y.\, z)$ in the common-effect model). If possible, this parameter can be estimated directly from frequency information. However, in some situations it might happen that no sufficient data is available, for example because the causes rarely co-occur. A physician might be confronted with an unusual combination of two rare diseases which she has not encountered before, though she has experienced the effects of both diseases separately. The question is then how the knowledge of the separate causal relations should be integrated to estimate how likely the effect occurs given several of the causes are present.

A popular integration rule modeling how multiple dichotomous causes generate a common effect are so-called "noisy-OR" gates (see Pearl, 1988, for a detailed analysis). In contrast to the logical OR, which applies to deterministic causes, the noisy-OR gate describes the disjunctive interaction of probabilistic causes. This integration rule models how likely the effect is to occur given the presence of multiple non-interacting causes. For example, in the common-effect displayed at the bottom left of Figure 3 event *X* is generated by causes *Y* and *Z* with probabilities $P(x \mid y.\, \neg z)$ and $P(x \mid \neg y.\; z)$.

[9] Figure 3 must not be understood as implying that the joint distribution is temporally prior to the causal hypotheses or that the models themselves are induced from the data.

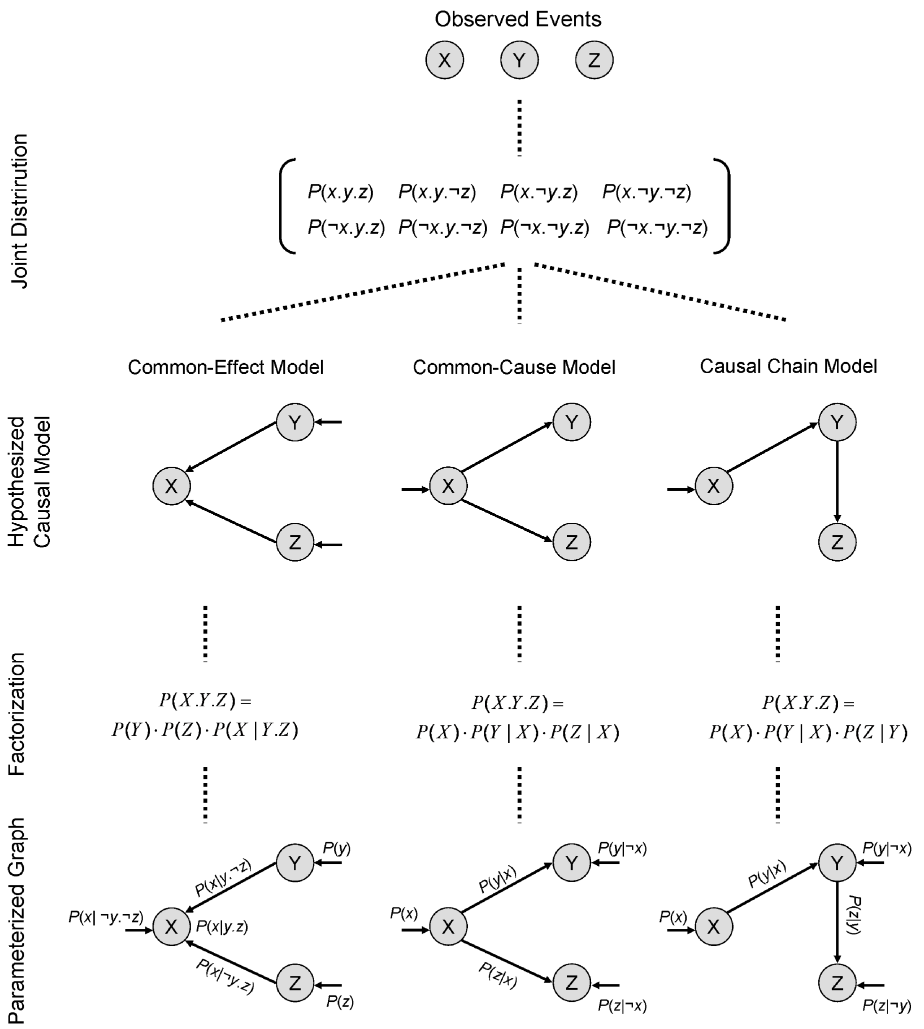


Figure 3. Basic causal models, factorization of the joint distribution, and parameterized graphs.

According to the noisy-OR integration rule, the probability of X conditional on the presence of both Y and Z (i.e., $P(x \mid y.\ z)$ is given by

$$P(x \mid y.z) = 1 - P(\neg x \mid y.\neg z) \cdot P(\neg x \mid \neg y.z) = 1 - ((1 - P(x \mid y.\neg z)) \cdot (1 - P(x \mid \neg y.z))) \quad (12)$$

For example, if the probability of developing hay fever is 0.5 in the presence of Y-pollen and 0.8 in the presence of Z-pollen (i.e., $P(x \mid y.\ \neg z) = 0.5$ and $P(x \mid \neg y.\ z) = 0.8$) the probability of hay fever given the simultaneous exposure to Y- and Z-pollen is

$$P(x \mid y.z) = 1 - ((1 - 0.5) \cdot (1 - 0.8)) = .9$$

A similar integration rule, so-called “noisy-AND” gates, exists for inhibitory causes (cf. Pearl, 1988).

Figure 3 also illustrates how the causal models determine which variables we should conditionalize upon. For example, in the common-cause model Z is conditionalized on its parent X, but in the causal chain event Z is conditionalized on its

Markovian parent *Y*. Thus, the causally based decomposition enables us to give specific estimates of causal strength since it tells us which variables we should conditionalize on and which variables are irrelevant. The conditional probabilities of the decomposed models can be derived from frequency information to parameterize the causal model. Associated with each relation in the causal model is then a conditional probability that determines the probability of an event conditional on its direct causes.

Due to the causal Markov condition, each causal model implies a specific set of unconditional and conditional dependence and independence relations. For example, the common effect model (Figure 3a) entails that *Y* and *Z* are unconditional independent but become dependent conditional on values of their common effect (so-called "explaining away effect"). Conversely, the common-cause model (Figure 3b) implies that *Y* and *Z* are unconditional dependent but become independent conditional on values of their common cause *X*. A Bayesian network satisfying the Markov condition can therefore be considered as a carrier of conditional independence information between the variables of the graph, a property also exploited by algorithms developed to induce causal structure from covariational data. Since these relations are a consequence of the structure of the graph they are structural dependency relations.

However, in some cases there might be specific combinations of parameters which can yield conditional independence relations, too. An example is given in the model in Figure 4 in which *X* causally influences *Z* via two paths, namely via a direct causal link but also indirectly via variable *Y*. For the sake of convenience, the model's parameters associated with the causal relations are denoted as *a*, *b*, and *c*. According to this causal model, *Z* is statistically dependent on *X* under virtually all parameterizations. However, there is a specific set of parameters in which *Z* is statistically independent of *X*, that is, when $a = -bc$ holds. Under this parameterization, the two alternative causal pathways cancel each other out and the consequence is that *X* and *Z* become statistically independent. Even though the realization of this condition might be quite unlikely (and would also be very unstable), it is theoretically possible and makes clear that independence relations can also occur because of specific parameterizations.

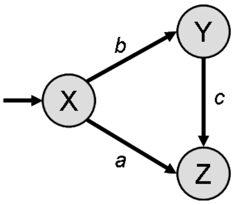


Figure 4. Graph with three nodes. Faithfulness rules out that *Z* is independent of *X*.

To account for such cases the assumption of *faithfulness* (Spirtes et al., 1993), or *stability* (Pearl, 2000), is introduced. The assumption of faithfulness states that the probabilistic dependency relations found to hold are a consequence of the causal Markov condition applied to the graph and do not result from specific parameterizations of the causal structure. In other words, faithfulness allows to move from probability distributions to graphs by constraining the set of candidate graphs. With respect to the causal model shown in Figure 4, the faithfulness condition rules out such a specific parameterization in real world situations. Thus, if we were to find out that *Z* is independent of *X* we should not consider the graph as a candidate model of the probability distribution.

4.2 Causal Learning with Bayes Nets

The preceding section has introduced causal Bayes nets as a formal means of representing causal knowledge. The following section gives an overview of how causal learning is modeled from a causal Bayes nets perspective. Most experiments on causal learning have investigated how learners derive judgments of causal strength for single causal relations or common-effect models. Typical examples are medical scenarios in which, for example, the participants' task is to evaluate the influence of a drug on a disease or to assess how different kinds of food relate to an allergic reaction. Since in such situations the qualitative causal structure is known before being presented with data, these studies are mainly concerned with the process of parameter estimation. The second key issue in causal learning is to induce the structure of complex causal models underlying patterns of covariation. Causal Bayes nets theory provides learning mechanisms for inferring causal structure from observational or interventional data, or a combination of both. In general, the formalism emphasizes the importance of structure learning because the structure determines the process of parameter estimation (e.g., by determining which variables we should conditionalize upon to give adequate estimates of causal strength). The first part of this section introduces the learning algorithms developed to infer complex causal models from observational data. The second part of this section describes how causal Bayes nets theories formalize interventions and model causal learning through interventions.

4.2.1 Causal Learning through Observations

Two kinds of learning algorithms have been developed in the context of causal Bayes nets: bottom-up constraint-based methods and top-down Bayesian methods. *Constraint-based methods* (Pearl, 2000; Scheines, Spirtes, Glymour, & Meek, 1994; Spirtes et al., 1993) try to induce causal models from the unconditional and conditional dependence and independence relations of the data. These algorithms can infer causal structures from observational data and can also integrate interventional data. Applied to human causal induction, these bottom-up approaches suppose that people examine the probabilistic dependency relations of the available data and use them to infer the underlying model. Constraint-based methods start with analyzing which probabilistic dependency relations hold between the observed variables (e.g., whether events *A* and *B* are (un)conditional (in)dependent). Algorithms such as TETRAD (Scheines et al., 1994) apply standard significance tests to determine whether a dependency relation holds. In a step-by-step procedure the algorithms then construct causal models consistent with the discovered unconditional and conditional dependence and independence relations (see Spirtes et al., 1993, for details). Thus, contingent on which dependency relations are satisfied by the data we can identify the underlying causal structure.

An alternative approach to structure induction is provided by top-down *Bayesian methods* (e.g., Heckerman, Meek, & Cooper, 1999; Steyvers et al., 2003). These approaches assume that learners start from a set of hypotheses about candidate causal models and update their hypotheses in accordance with the available data. Briefly, Bayesian learning procedures start by assigning a prior probability to each graph; either to the complete set of possible graphs or to a restricted set. The prior probabilities assigned to the causal models can be uniform, but it is also possible to incorporate prior knowledge by giving some models a higher prior probability than others. Together with assumptions about the probability functions relating the variables, the likelihood of a particular data pattern under each of the graphs can be computed. For example, a data pattern such as $x.y.z$ (i.e., all events are present) is more likely if the three variables X, Y, and Z form a common-cause model than when the events form a common effect model (because here the cause events occur independently of each other). By using Bayes theorem it is then possible to compute the posterior probability distribution over the considered causal graphs conditional on the available data. The graph with the highest posterior probability is then chosen as the one most likely to have generated the data.

Both constraint-based and Bayesian algorithms provide powerful computational methods to induce causal structure from statistical data, capitalizing on the fact that different causal structures entail different dependency relations. However, some causal models are not only observationally equivalent but also share the same set of dependency relations. For example, the finding that Y and Z are independent conditional on X is not only consistent with a common-cause model $Y \leftarrow X \rightarrow Z$ but also with the causal chains $Y \rightarrow X \rightarrow Z$ and $Y \leftarrow X \leftarrow Z$. Thus, from observational data alone these methods can only reduce the space of possible graphs to a subset of models which share the same set of probabilistic dependency relations. Such models are referred to as being *Markov equivalent* (cf. Spirtes et al., 1993). Figure 5 gives an example of the possible models we can construct from three variables *X*, *Y*, and *Z*. Shaded areas group models according to their topology, dashed lines indicate Markov equivalent models (cf. Steyvers et al., 2003).

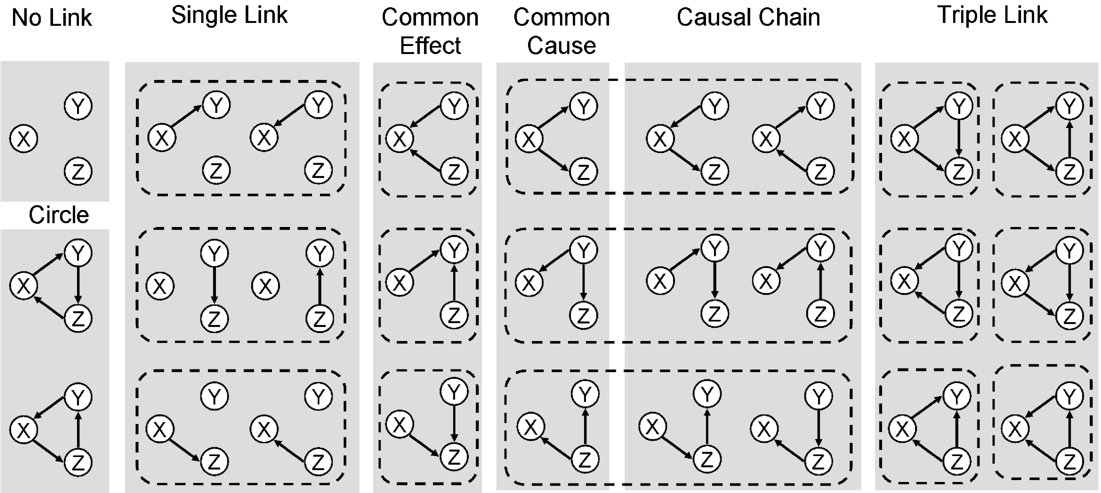


Figure 5. All possible networks with three variables. Shaded areas group models according to their topology, dashed lines indicate Markov equivalent models (cf. Steyvers et al., 2003). Cyclic graphs (bottom left) cannot be analyzed by standard causal Bayes nets analysis. See text for details.

The scope of these learning algorithms is beyond that of psychological models because they make unrealistic assumptions about the necessary information processing capacities, especially in situations with many variables. For example, constraint-based methods require learners to conduct a large number of comparisons to test which dependency relations exist in the investigated domain. Similarly, Bayesian methods have the problem that the number of possible causal models grows very fast with the number of variables (i.e., at least exponentially). Therefore, even in computer science heuristics are used which constrain the space of candidate models (cf. Heckerman et al., 1999). However, in less complex situations learners might use the analysis of dependency relations (cf. Gopnik et al., 2004). For example, learners could start from a restricted set of candidate models and analyze the available data specifically with respect to the conditional dependency relations implied by their hypothesized models

Thus far, only few studies have investigated whether learners have the capacity to induce causal structure from dependency relations alone. Gopnik and colleagues have argued that children as young as 30 months use information about dependencies to infer causal relations in accordance with constraint-based methods (see Gopnik et al., 2004, for an overview). However, research with adult reasoners suggests that this competency strongly depends on the complexity of the learning task (e.g., the number of variables, deterministic or probabilistic relations). For example, Lagnado and Sloman (2004) found that neither learners' probability judgments nor their model choices matched the predictions of constraint-based methods. Similarly, the experiments of Steyvers and colleagues (2003), who proposed a psychologically more plausible model of Bayesian inference, show that only few learners were able to identify the correct model from observational data alone. However, Hagmayer and Waldmann (2000) found an interesting dissociation between explicit and implicit sensitivity to the structural implications of different causal models (e.g., a common-cause vs. a common-effect model). Whereas learners' explicit probability judgments showed only a limited understanding of the structural implications of the different models, participants performed better in an implicit task requesting them to predict patterns of events they expected to see.

In a recent article, Lagnado and colleagues concluded that there is only little evidence that learners can uncover complex causal models from covariational data alone (Lagnado et al., in press). They point out that there are a number of other cues to causality which can be used to infer causal structure from observational data. For example, temporal information and prior knowledge can assist structure learning by providing additional cues or constraining the set of candidate models. Another effective learning tool to discover a causal system's structure is through active manipulation of the causal model's variables, which is discussed next.

4.2.2 Causal Learning through Interventions

The preceding section has outlined how causal Bayes nets theory models causal inference from observational data. However, there are crucial differences between learning from observations (observational learning) and learning from data generated through interventions (interventional learning). Whereas observations provide us with information about the operation of an undisturbed causal system, the "natural course of events", interventions inform us about a causal system's behavior conditional on active manipulations of the system's variables. Observing the consequences of our actions facilitates causal inference because we can focus on certain aspects of the investigated system and attribute observed changes to the events previously intervened in. For example, we interact with technical systems such as computers or mp3-players to find out how they work, we change our eating habits or start exercising to reduce our weight, or we try different kinds of fertilizer to make our garden's flowers grow. These kinds of informal experiments provide us with direct causal knowledge about the consequences of our actions.

In the causal Bayes nets framework, interventions which fix the state of a variable to a specific value or probability are called *atomic interventions* (cf. Pearl, 2000; Woodward, 2003). The characteristic feature of such strong interventions is that they render the variable intervened in independent of its actual causes (i.e., its Markovian parents).[10] Sloman and Lagnado (2005) referred to this induced independence of cause and effect as "undoing". For example, if we arbitrarily change the reading of the barometer our action renders the barometer independent of its usual cause, atmospheric pressure. Graphically, this can be represented by modifications of the graphical representation of the considered causal system: since the value of the variable intervened in is not any longer dependent on its actual causes, all arrows pointing at this variable are removed. Pearl (2000) has vividly called this procedure "graph surgery"; the result is a "manipulated graph" (Spirtes et al., 1993). A more technical introduction to the representation of intervention using probability calculus will be given in section 4.3.

To illustrate how interventions are modeled in the Bayes nets framework consider the following scenario. Assume we have observed a fish kill (*F*) in a pond and want to discover the causes. Analyses of the water show that there is a high level of nitrogen (*N*) in the water and a large amount of algae (*A*). Because a fish kill is a rather rare event, we cannot sample a large amount of observational data to analyze the dependency relations. However, we can focus on a set of candidate models which we then scrutinize by actively manipulating the causal system. Plausible candidate models are a common-effect model, in which both nitrogen and algae (independently or interactively) contribute to the fish kill, a common-cause model in which the nitrogen causes both an increase of algae and the fish kill, and a causal chain leading from nitrogen to algae which, in turn, causes the fish kill. To differentiate between these models we can intervene in the system's variables and examine the consequences of these actions. For example, if there exists a causal relation $A \rightarrow F$ we should observe an increase in fish kill subsequent to increasing the amount of algae.[11] Conversely, if algae and fish kill are

[10] This, obviously, is also the key idea of the experimental method in science (e.g., Fisher, 1951).

[11] Note that knowledge about the exact nature of the underlying mechanisms (e.g., a decrease in oxygen caused by the algae) is not required to infer the mere existence of causal relation.

only spuriously correlated because they are both effects of a common cause, manipulating the amount of algae should not harm the fish.

From a causal Bayes nets perspective, interventions modify the structure of the graph representing the considered causal system. Because the intervention sets the variable targeted by the intervention to a certain value, the variable becomes independent of its actual causes (i.e., its Markovian parents). Graphically, this is represented by removing all arrows pointing towards this variable (= graph surgery); the result is a manipulated graph. Figure 6 illustrates the principle of graph surgery for an intervention that fixes the amount of algae in the water to a certain level.

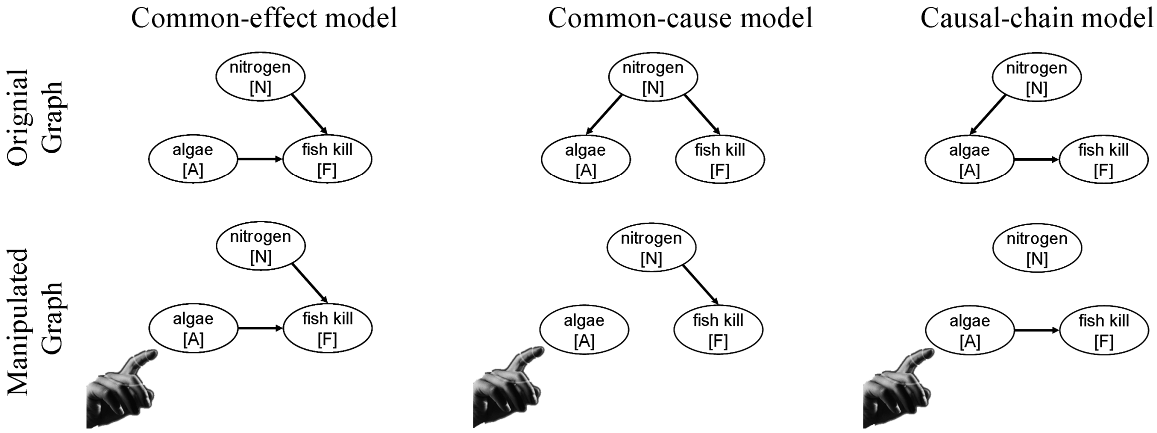


Figure 6. Graph surgery in different causal models after an intervention (symbolized by the hand).

Because intervening in a causal model modifies its structure, data obtained from interventions enable us to differentiate between otherwise observationally equivalent causal models. For example, assume that we observe a correlation between three events *X*, *Y*, and *Z* (i.e., all three events tend to be either present or absent). Provided no additional cues such as temporal order are available, several causal models are consistent with this data. By analyzing the dependency relations, we can only reduce the set of candidate models to a subset of Markov equivalent models, for example to a set consisting of a common-cause model $Y \leftarrow X \rightarrow Z$ and the causal chains $Y \rightarrow X \rightarrow Z$ and $Y \leftarrow X \leftarrow Z$. Through interventional learning, we can distinguish between these models. For example, we could manipulate event *X* (i.e., set variable *X* to a certain value) and observe the outcomes of this action. If *X* is a common cause to *Y* and *Z*, intervening in *X* should affect both variables. In contrast, if the variables form a causal chain either *Y* or *Z* should be influenced by manipulations of *X*, depending on which of the two causal chains is the true model. The example also illustrates that the advantage of interventions crucially depends on which variable we choose to intervene in (cf. Steyvers et al., 2003). Whereas interventions in *X* predict for each of the three candidate models a different outcome, this is not the case when intervening in *Y* or *Z*.

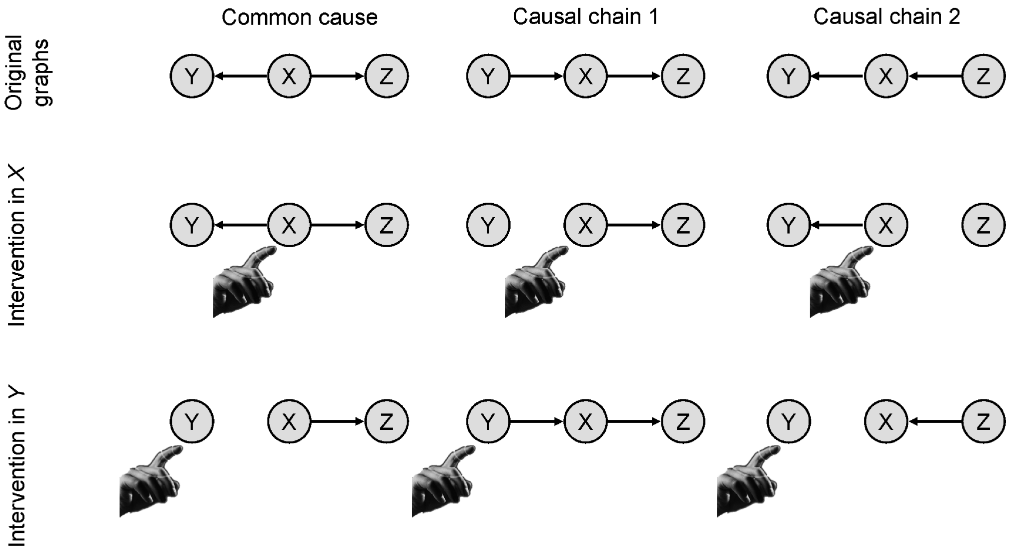


Figure 7. Interventions in Markov equivalent models.

For example, the outcomes of interventions in *Y* differentiate the causal chain *Y*→*X*→*Z* from the other two models because this chain is the only model according to which interventions in *Y* should affect both *X* and *Z*. In contrast, manipulations of *Y* cannot differentiate the common-cause model *Y*←*X*→*Z* from the causal chain *Y*←*X*←*Z* because both models imply that interventions in *Y* will influence neither *X* nor *Z*. Figure 7 illustrates the manipulated graphs subsequent to interventions in *X* and *Y*, respectively. Recent work in psychology has demonstrated that learners can use the outcomes of interventions to infer causal structure and differentiate between candidate causal models (Gopnik et al., 2004; Lagnado & Sloman, 2004, in press; Steyvers et al., 2003). For example, Lagnado and Sloman compared the learning of a simple causal-chain model from observations and interventions. Their results show that learners were more successful in identifying the causal model when they could actively intervene on the model's variables than when provided with observational data. A further experiment employed a yoked design in which learners in one condition could actively generate data by interventions, whereas participants of the yoked condition passively observed the outcomes of these interventions (intervention vs. observation of intervention). Interestingly, participants who actively intervened performed better than those who only observed the outcomes of interventions. This finding indicates that the capacity to infer causal structure is not determined only by differences in the informational content of observational and interventional data. Therefore, Lagnado and Sloman suggest that the advantage of learning through interventions mainly results from the temporal cues that accompany interventions. According to their *temporal cue heuristic* people exploit the temporal precedence of their actions and the resulting outcomes, because changes subsequent to interventions can be attributed to the variable intervened in (see Lagnado & Sloman, 2004, in press, for further details). In line with the experiments of Lagnado and Sloman, the experiments of Steyvers et al (2003) demonstrated that learners perform better when given the opportunity to actively intervene on the causal system than when only passively observing the data generated from the autonomous operation of the system. Their experiments also show that participants' choices of the variable to intervene in are sensitive to the expected information gain, that is, how well the outcomes of the interventions can discriminate between competing hypotheses about the causal system

However, the results of both the studies of Lagnado and Sloman and Steyvers and colleagues also show that many learners still had problems to infer the correct model from covariational data, even when given the opportunity to act on a causal system and observe the subsequent changes.

4.3 Causal Reasoning with Bayes Nets

The preceding sections have examined causal Bayes nets theory as a formal account of causal representation and causal learning. This chapter is concerned with the third key issue in causal cognition, causal reasoning. The characteristic feature of causal reasoning is that we aim to infer unobserved features of the world from the available information and our existing causal knowledge. For example, a physician inferring a bacterial infection from a patient's symptoms engages in diagnostic causal reasoning in which the (unobserved) cause event is inferred from its (observed) effects. Conversely, inferring the symptoms from knowledge of the bacterial infection is an example of predictive causal reasoning (i.e., reasoning from causes to effects).

An important issue is how we access and integrate our causal representations in causal reasoning. Often, people can base their inferences on causal knowledge acquired in similar situations. For example, an experienced physician might predict a patient's future course of disease from her observations of previously encountered cases. Conversely, interventional knowledge about the outcomes of actions on previous occasions can be used to decide between potential treatments of the disease.

Characteristic for both situations is that the way the causal knowledge is acquired corresponds to the type of inference: knowledge acquired from observations is used to derive observational predictions and knowledge gained through active manipulations is used to predict the consequences of future interventions.

However, how do we infer the consequences of hypothetical actions when only observational knowledge is available? How does the physician employ her observational knowledge about the natural course of a disease to predict the outcomes of actions not performed yet? The fundamental problem that has to be taken into account when deriving interventional predictions from observational knowledge is that merely observed states of variables can have different implications than the very same states generated by external interventions. For example, the physician might have observed that the fever of a patient declines before recovering from a disease. However, does this mean that future patients should be treated with a drug that lowers the fever? Probably not, because the fever is likely to be an effect of the underlying infection, too. Traditional theories of causal cognition such as contingency models fail to account for this difference because they cannot express causal directionality and lack the representational power to differentiate between merely observed states of variables from the same states generated by means of intervention. It is true that associative theories traditionally distinguish between knowledge acquired from observational learning (classical conditioning) and interventional learning (instrumental conditioning). However, these models fail when predictions for instrumental actions have to be derived from observational learning (cf. Section 5.2). In addition, neither approach provides a means to represent causal structure, which is crucial to model interventions in complex causal systems.

By contrast, causal Bayes nets theory captures the distinction between observations (seeing) and interventions (doing), and provides mechanisms for predicting the outcomes of hypothetical and counterfactual interventions from causal models parameterized by observational knowledge. The formalism models three types of causal inference which differ in their representational and computational demands: *observational inferences* (i.e., inferences based on observed states of variables), *interventional inferences* (i.e., inferences based on states of variables generated by external interventions), and *counterfactual inferences* (i.e., inferences about the outcomes of counterfactual actions). Counterfactual inferences refer to the outcomes of interventions not in the actual but in a counterfactual world, because the action contradicts factual states of the world: “The patient is in critical condition after his fever was observed to increase (observations in the actual world). What would have happened if I had intervened to lower his fever? (counterfactual intervention)”. Thus, counterfactual inferences combine factual observations with the question of what would have happened if the observed state had been modified through an intervention.

To exemplify how causal Bayes nets theory models observations, interventions, and counterfactual interventions, I will use the diamond-shaped causal model depicted in Figure 8. The graph is illustrated by an ancient communication system consisting of four watchtowers *A*, *B*, *C*, and *D*. The purpose of the system is to transmit a signal from tower *A*, which is close to the enemy border, to tower *D*. Thus, variable *A* is the initial event that can cause the final effect *D* via two causal paths, that is, either by way of *B* or *C*. For example, tower *A* can light a signal fire that can be observed by towers *B* and *C*. These towers, in turn, can light a fire to transmit the signal further to the final tower *D*. However, the communication might not always be successful, for example because of bad weather. Figure 8 depicts the communication system and the corresponding graph.

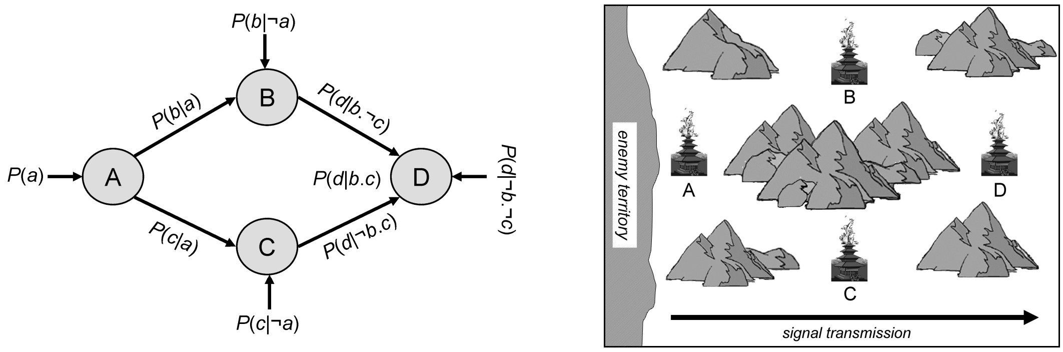


Figure 8. Diamond-shaped causal model illustrated by a primitive communication system.

By applying the causal Markov condition to the causal model the joint probability distribution is factorized into

$$P(A.B.C.D) = P(A)\cdot P(B\,|\,A)\cdot P(C\,|\,A)\cdot P(D\,|\,B.C) \qquad (13)$$

According to this factorization, the probability of each event is only dependent on its direct causes (i.e., its Markovian parents). This is mirrored in the parameters associated with the model’s causal arrows. Provided all variables can be observed, the conditional probabilities of the decomposed model can be directly estimated from the frequency information.

4.3.1 Observational Inferences

Observational inferences refer to predictions based on observed states of the causal system's variables. Based on the structure of the causal model and its parameters, the probabilities implied by the observations can be computed using standard probability calculus. For example, from observing the state of variable *C* in the causal model shown in Figure 8, the probability of *A* being present can be computed using Bayes rule. Thus, if there is a fire on tower *C*, the probability that there is also a fire on tower *A* is given by

$$P(a \mid c) = \frac{P(c \mid a) \cdot P(a)}{P(c \mid a) \cdot P(a) + P(c \mid \neg a) \cdot P(\neg a)} = \frac{P(c \mid a) \cdot P(a)}{P(c)} \tag{14}$$

Conversely, if there is no fire on tower *C* the probability of a signal fire on tower *A* is given by

$$P(a \mid \neg c) = \frac{P(\neg c \mid a) \cdot P(a)}{P(\neg c \mid a) \cdot P(a) + P(\neg c \mid \neg a) \cdot P(\neg a)} = \frac{P(\neg c \mid a) \cdot P(a)}{P(\neg c)} \tag{15}$$

These computations model diagnostic inferences from observed states of event *C* to its cause *A*. These simple diagnostic inferences only require taking into account the direct causal pathway between *A* and *C*, while the rest of the model is irrelevant.

A more interesting example is the prediction of variable *D* from observations of event *C*. Obviously there is a direct causal link connecting *C* to *D* but there is also a second causal pathway connecting *C* to *D* via *A* and *B*. Therefore, the probability of *D* given that *C* is observed to be present not only depends on the strength of the direct causal arrow *C*→*D* but also on the information *C* provides about the state of *D*'s alternative cause, event *B*. For example, observing a fire on tower *C* increases the probability of *A*, which, in turn, raises the probability that there is a fire on tower *B*, too. This, in turn leads to an increase of the probability of *D* being present (i.e., a signal fire on tower *D*). Pearl (2000) vividly called such confounding pathways *backdoors*. Formally, the probability for *D* given that *C* is observed to be present can be calculated by:

$$\begin{aligned} P(d \mid c) &= \sum P(A \mid c) \cdot P(B \mid A) \cdot P(d \mid B.c) \\ &= P(a \mid c) \cdot P(b \mid a) \cdot P(d \mid b.c) + P(a \mid c) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.c) + \\ &\quad P(\neg a \mid c) \cdot P(b \mid \neg a) \cdot P(d \mid b.c) + P(\neg a \mid c) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.c) \end{aligned} \tag{16}$$

Similarly, the probability of *D* given that *C* is observed to be absent is computed by

$$\begin{aligned} P(d \mid \neg c) &= \sum P(A \mid \neg c) \cdot P(B \mid A) \cdot P(d \mid B.\neg c) \\ &= P(a \mid \neg c) \cdot P(b \mid a) \cdot P(d \mid b.\neg c) + P(a \mid \neg c) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.\neg c) + \\ &\quad P(\neg a \mid \neg c) \cdot P(b \mid \neg a) \cdot P(d \mid b.\neg c) + P(\neg a \mid \neg c) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.\neg c) \end{aligned} \tag{17}$$

By conditionalizing *A* on *C*, these computations take into account that observed states of *C* are diagnostic for the state of *A* which, in turn, allows for the inference of the probability of *B*. The probability of the final effect *D* is estimated by adding up the different ways the occurrence of the event could be realized. Such alternative pathways are not only important to give unconfounded estimates of causal strength but are especially important with respect to interventions.

4.3.2 Interventional Inferences

A particularly interesting feature of the causal Bayes nets formalism is the derivation of interventional predictions from observational data and causal graphs. Interventional predictions refer to questions of the type "What would happen to *Y* if *X* were manipulated?". Causal Bayes nets theory makes it possible to answer such questions by means of observational knowledge and assumptions about the underlying causal model.

The notion of atomic intervention and the principle of graph surgery have already been introduced in the context of learning through intervention (cf. Section 4.2.2). The literature on causal Bayes nets has focused on these kinds of ideal interventions in which the action changes the value of a variable independent of the state of the variable's parents (for more precise characterizations of these interventions, see, for example, Woodward, 2003). I have already examined how such atomic interventions modify the graphical representation of the causal system. This stage of model manipulation is also essential when reasoning about interventions, since interventional predictions should be based on the modified graph and not on the original graph.

To formalize the idea that a variable's state is not based on the "natural course of events" but was determined by an external intervention, Pearl (2000) introduced the so-called "Do-Operator", written as Do (•). For example, the expression "Do $C = c$" ("Do *c*" for short) is read as "variable *C* is set to state *c* by means of an intervention". The Do-operator is the formal equivalent of graph surgery in terms of probability theory. Whereas the probability $P(a \mid c)$ refers to the probability of *A* being present given that *C* was *observed* to be present, the expression $P(a \mid \text{Do } c)$ refers to the probability of *A* being present given that *C* was *generated* by means of intervention.

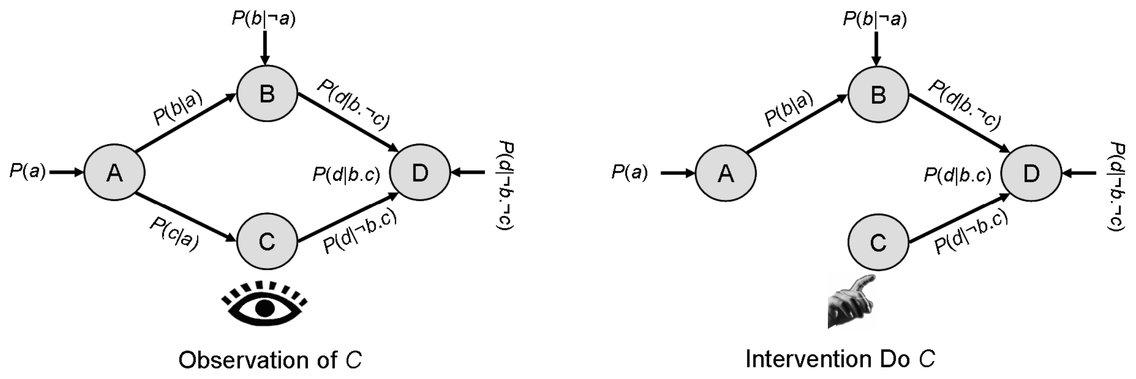


Figure 9. Observation of and intervention in variable *C*.

Because of graph surgery, interventions (in contrast to observations) do not provide diagnostic evidence for the causes of the manipulated variable. Thus, the Do-operator renders a variable independent of its direct causes, which is equivalent to deleting all causal links pointing towards the variable fixed by the intervention. Figure 9 illustrates the difference between observation of and intervention in *C* and shows the mutilated graph resulting from applying the do-Operator to variable *C*.

The Do-operator provides the formal means to represent the crucial differences between observations and interventions in the language of probability calculus. For example, the probability of $A = a$ (i.e., a signal fire on tower *A*) given that *C* is caused by an intervention (e.g., lightning that has lit the signal fire) equals the base rate of $A = a$ because the causal link connecting these two events was eliminated by the intervention, and therefore

$$P(a \mid \text{Do } c) = P(a \mid \text{Do } \neg c) = P(a). \tag{18}$$

Thus, whereas inferring the state of event *A* from observed values of *C* is modeled by conditionalizing on *C* (cf. equations (14) and (15)), this does not hold when the state of *C* is set by an intervention. Applying the Do-operator to a variable implies that the state of this event is no longer diagnostic for the state of its Markovian parents; therefore they are no longer conditionalized on the variable targeted by the intervention. For example, if the signal fire on tower *C* is lit because of an event outside of the causal system (e.g., lightning), the presence of the fire does not raise the probability that there is a signal fire on tower *A*.

In the same way, the probability of *D* can be calculated using the modified causal model. Generating a value of *C* through an intervention "closes the backdoor", since states generated by external interventions do not provide diagnostic evidence for a variable's actual causes. Nevertheless, the initial cause *A* may occur with its base rate and influence *D* via *B*. Therefore, the correct formula to calculate the probability of $D = d$ given that *C* is generated by external intervention (i.e., Do *c*) is

$$\begin{aligned} P(d \mid \text{Do } c) &= \sum P(A) \cdot P(B \mid A) \cdot P(d \mid B.c) \\ &= P(a) \cdot P(b \mid a) \cdot P(d \mid b.c) + P(a) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.c) + \\ &\quad P(\neg a) \cdot P(b \mid \neg a) \cdot P(d \mid b.c) + P(\neg a) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.c) \end{aligned} \tag{19}$$

Similarly, in case of an inhibitory intervention in *C* (i.e., Do ¬*c*) the probability of $D = d$ is given by:

$$\begin{aligned} P(d \mid \text{Do } \neg c) &= \sum P(A) \cdot P(B \mid A) \cdot P(d \mid B.\neg c) \\ &= P(a) \cdot P(b \mid a) \cdot P(d \mid b.\neg c) + P(a) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.\neg c) + \\ &\quad P(\neg a) \cdot P(b \mid \neg a) \cdot P(d \mid b.\neg c) + P(\neg a) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.\neg c) \end{aligned} \tag{20}$$

In contrast to the computations modeling the observational inferences, variable *A* is no longer conditionalized on *C* in these formulas but replaced by the base rate *P*(*A*) (cf. equations (16) and (17)). Crucially, on the right-hand side of the equations only parameters are involved which can be derived from observational data. Thus, no direct knowledge about the outcomes of interventions is necessary (i.e., parameters acquired from interventional learning).

There are two important criteria of atomic interventions which must be met to infer the consequences of interventions from causal models parameterized by passively observed events (see Pearl, 2000; Woodward, 2003, for further details). First, because interventional predictions are derived from manipulated graphs, it must be known which variables are affected by the intervention. For example, in the diamond-shaped causal model interventions in *C* render the event independent of its actual cause, event *A*. Graphically, this is represented by removing all arrows pointing at *C* while leaving the rest of the model intact (= graph surgery). However, if the intervention accidentally also fixed the value of *B*, the manipulated graph in which only *C* is disconnected from *A* would not correctly represent the structural modifications implied by the intervention (i.e., that variable *B* is also not any longer influenced by variable *A*). The second prerequisite concerns the stability of the parameters associated with the operation of the causal system's mechanisms. Since the computations used to derive interventional predictions involve these parameters, it is necessary that they are invariant against interventions. In other words, the causal strength of a causal arrow *C*→*E* must be independent of whether *C* is generated by its natural causes or whether *C* is set by means of intervention. If this is not the case, for example because the intervention not only influences *C* but also the mechanism by which C causes *E*, the predictions about the consequences of interventions are likely to be erroneous. See Hausman and Woodward (1999) and Woodward (2003) for further details; see Cartwright (2002) for a critical analysis.

4.3.3 Counterfactual Inferences

Causal Bayes nets theory not only provides computational mechanisms to infer the consequences of hypothetical interventions but also models counterfactual actions. Counterfactual interventions refer to interventions that deviate from the factual course of events in the world. For example, "If my friend had stopped me from driving too fast then I would still have my driver's license" is an inference based on a counterfactual statement, because the events "driving too fast" and "losing my driving license" are true in the actual world. Similar to the predictions for hypothetical interventions, counterfactual inferences are not invariant against causal structure.

Counterfactual inferences combine observations and interventions. A counterfactual intervention is defined as an action that alters a factual state of the world. Thus, inferences of this type refer to states in a counterfactual world. Causal Bayes nets allow to formalize questions such as "Today no signal fire was observed on tower *C* – what is the (counterfactual) probability of a fire on tower *D* if a fire on tower *C* had been lit by lightning?" Such counterfactuals comprise two pieces of information: a factual observation and a counterfactual action altering this state. For example, the counterfactual probability $P(d \mid \neg c. \text{ Do } c)$ is read as "the probability of $D = d$ given that *C* was observed to be absent but counterfactually generated". The first piece of information is that *C* was observed to be absent, a statement referring to the actual world. This information provides the basis for updating the probabilities of *C*'s Markovian parents (i.e., variable *A*). The second piece of information posits the counterfactual generation of *C*. This action refers not to the actual world (in which event *C* was absent) but to a counterfactual world in which *C* has been generated by external intervention. From these two pieces of information we can derive the state of *D* in a counterfactual world (which might or might not correspond to the factual world).

Thus, the basic logic is that the probabilities of *C*'s causes are updated in accordance with the event's factually observed state and estimates for *C*'s effects are computed conditional on the implications of the counterfactual action. We then have a three-step procedure for computing the consequences of counterfactual interventions (cf. Pearl, 2000). First, the probabilities of the observed variable's causes are updated in accordance with the event's state in the actual world. Second, the causal graph is modified according to the counterfactual intervention, that is, graph surgery is performed. The crucial point is that graph surgery is performed in the updated model, not on the original one. Otherwise, the intervention would render the variable targeted by the intervention independent of its causes which then could not be updated. Finally, the updated and truncated model is used to predict the consequences of the counterfactual intervention. Figure 10 illustrates how causal Bayes nets model counterfactual interventions.

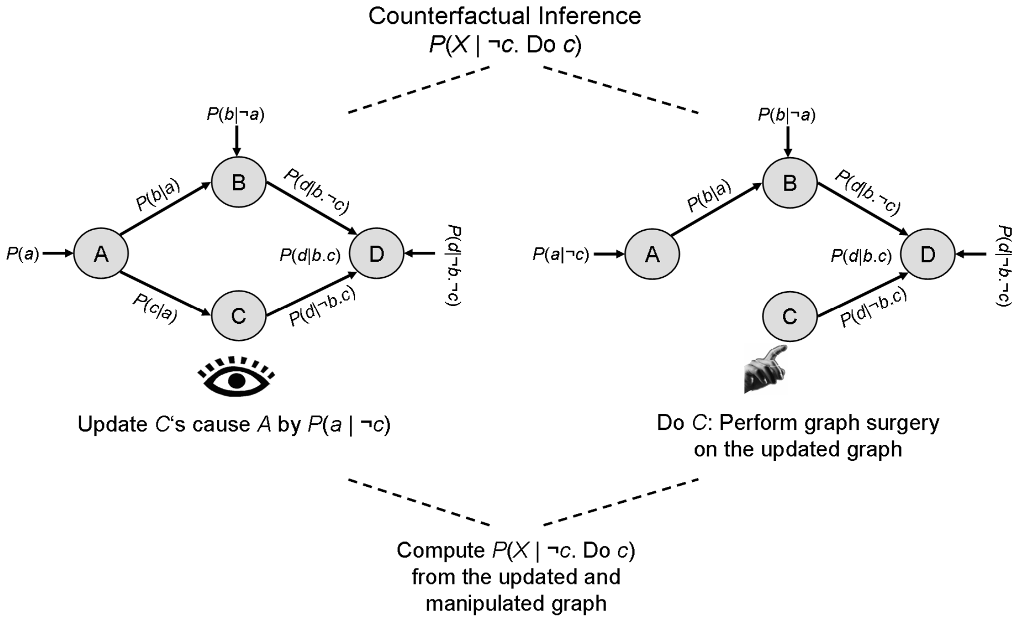


Figure 10. Modeling counterfactual interventions.

For example, in the diamond-shaped model we might observe *C* to be present (e.g., a signal fire is observed on tower *C*) but ask for the probability of a certain variable in the system conditional on a counterfactual prevention of event *C* (i.e., an intervention that would have prevented the signal fire on tower *C*). Again, I start with the simple diagnostic inference from *C* to its cause *A*. Since intervening in an effect will not influence its cause, the probability of *A* is determined by the factually observed state of *C* alone. Thus, the probability of *A* given we observe *C* to be present but counterfactually remove *C* is given by

$$P(a\,|\,c.\,\mathrm{Do}\,\neg c) = P(a\,|\,c) \tag{21}$$

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

$$P(a \mid \neg c.\, \text{Do}\ c) = P(a \mid \neg c) \tag{22}$$

Whereas for hypothetical interventions the probability of *C*'s cause *A* was given by *A*'s base rate, in the case of counterfactual interventions the probability of event *A* is updated in accordance with the observed state of *C*. Since the counterfactual intervention implies the alteration of a factual state, the probability of *A* is higher in the case of a counterfactual inhibition of *C* (which logically entails that *C* has been observed to be present) than in the case of a counterfactual generation of *C* (which implies that *C* has been observed to be absent in the actual world). For example, the probability of a signal fire on tower *A* is higher when a fire on tower *C* is observed to be present but counterfactually undone than in the case of a hypothetical intervention preventing the fire on tower *C*.

It is also possible to compute the probability of the final effect *D* conditional on counterfactual interventions in *C*. For this inference, it is necessary to integrate both the observed values of *C* and the counterfactual intervention in *C*. This is an interesting case since the counterfactual intervention will affect the direct link *C*→*D*, but the observed value of *C* provides information about *A*, and, therefore, also about the probability with which *D*'s alternative cause, *B*, occurs. Consider the counterfactual probability *P*(*d* | *c*. Do ¬*c*) which asks for the probability of *D* given that *C* is observed to be present but counterfactually removed. The counterfactual prevention of *C* will break the causal path *C*→*D* therefore *D* is not any longer influenced by *C* in the counterfactual world. However, observing *C* to be present indicates that its cause *A* has been present, too, which, in turn, raises the probability of *B* occurring. Similar to the hypothetical intervention Do ¬*c*, event *D* is then completely determined by the backdoor path. However, there is one crucial difference: whereas in case of the "normal"(hypothetical) intervention the probability of event *A* is given by its base rate, in case of a counterfactual intervention *A* is conditionalized on the observed value of *C*. Thus, the probability of *D* is higher in case of a counterfactual prevention of *C* than in case of a hypothetical prevention of *C*, provided *P*(*a*) < *P*(*a* | *c*). The corresponding formula is

$$\begin{aligned} P(d \mid c.\, \text{Do}\, \neg c) &= \sum P(A \mid c) \cdot P(B \mid A) \cdot P(d \mid B. \neg c) \\ &= P(a \mid c) \cdot P(b \mid a) \cdot P(d \mid b. \neg c) + P(a \mid c) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b. \neg c) + \\ &\quad P(\neg a \mid c) \cdot P(b \mid \neg a) \cdot P(d \mid b. \neg c) + P(\neg a \mid c) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b. \neg c) \end{aligned} \tag{23}$$

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

$$
\begin{aligned}
P(d \mid \neg c.\,\mathrm{Do}\; c) &= \textstyle\sum P(A \mid \neg c) \cdot P(B \mid A) \cdot P(d \mid B.c) \\
&= P(a \mid \neg c) \cdot P(b \mid a) \cdot P(d \mid b.c) + P(a \mid \neg c) \cdot P(\neg b \mid a) \cdot P(d \mid \neg b.c) + \\
&\quad P(\neg a \mid \neg c) \cdot P(b \mid \neg a) \cdot P(d \mid b.c) + P(\neg a \mid \neg c) \cdot P(\neg b \mid \neg a) \cdot P(d \mid \neg b.c)
\end{aligned}
\qquad (24)
$$

Note that in these formulas event *A* (*C*'s cause) is conditionalized on the observed state of *C* but the probability of *D* (*C*'s effect) is then computed from the counterfactually altered state of *C*. Again only parameters are involved which can be derived from observational learning.

4.4 Causal Bayes Nets Theory: Summary

Causal Bayes nets provide a formal account of causal representation, causal learning, and causal reasoning. By combining graph theory and probability calculus, the account explicitly differentiates between causal structure and causal strength, an important distinction blurred by most theories of causal induction. It is the formal representation of interventions and the capacity to derive interventional predictions from observations that sets causal Bayes nets apart from conventional theories of causal cognition. The representational power of the Bayes nets framework provides a formal account of interventions in causal systems which models learning through interventions and enables us to infer the consequences of hypothetical and counterfactual interventions from causal models parameterized by observational data.

The aim of the following section is to discuss causal Bayes nets theory as a psychological model of causal cognition. The formalism makes precise and testable predictions concerning which factors should influence a learner's causal inferences. The emphasis of the presented experiments is on causal reasoning and the different types of causal inference modeled by Bayes nets theory. The main questions pursued are whether learners are sensitive to the differences between observations and interventions and whether they have the capacity to derive predictions about the consequences of hypothetical and counterfactual interventions from causal models and observational knowledge.

5 Causal Bayes Nets as Models of Causal Cognition

The examples described above show that, normatively, causal inferences differ depending on whether they are based on hypothetical observations (seeing) or hypothetical interventions (doing). The interesting question is whether people are sensitive to the difference between observations and interventions as well. Thus far, only few studies have addressed this question (see Hagmayer et al., in press, for an overview).

Sloman and Lagnado (2005) have studied causal inferences in given causal structures and have compared causal with logical arguments. For example, participants were suggested a causal-chain model consisting of three events that were all described to be present. Participants were then requested to imagine that the intermediate event was either removed by an intervention or observed to be absent. In accordance with the predictions of causal Bayes nets participants inferred that the initial cause in the chain would be absent if the intermediate event was observed to be absent, but not if it was actively removed. Overall, the results of Sloman and Lagnado's "undoing" experiments were consistent with the predictions of causal Bayes nets theory. Because the focus of these studies was on comparing logical with causal reasoning, only qualitative reasoning based on the structure of causal models was investigated. In general, Sloman and Lagnado focused on people's capacity to differentiate between seeing and doing when reasoning about described causal situations. In addition, only one out of six experiments investigated reasoning about probabilistic causal relations.

Waldmann and Hagmayer (2005) wondered whether learners' inferences would be sensitive to the size of the parameters that were gleaned from observational learning data. Participants in their experiments were given instructions about the structure of causal models and subsequently received a list of cases on a single page that could be used to estimate the parameters of the models. Participants were then requested to derive predictions for new hypothetical observations and hypothetical interventions. The results showed a surprising grasp of the differences between seeing and doing, manifesting itself in predictions that took into account the size of the parameters which were estimated on the basis of the learning data.

Quite recently, it has also been demonstrated that the competency to distinguish between observations and interventions is also found in animals (Blaisdell, Sawa, Leising, & Waldmann, 2006). Previous research with animals appeared to support the view that associative processes drive learning in animals, and that animals lack a deeper understanding of causality. The experiments of Blaisdell and colleagues show that rats can derive interventional predictions subsequent to a classical conditioning phase, a finding inconsistent with the assumption that predictions of the consequences of instrumental actions require a prior phase of interventional learning (i.e., instrumental conditioning).

5.1 Research Questions

The general goal of this dissertation is to investigate learners' competency in deriving the consequences of hypothetical and counterfactual interventions from observational knowledge. The normative model against which learners' predictions are judged is causal Bayes nets theory.

Three key issues are examined in detail. Firstly, previous studies used described causal situations (Sloman & Lagnado, 2005) or provided learners with lists of aggregated data available during causal reasoning (Waldmann & Hagmayer, 2005). By contrast, the experiments presented here move one step further in the realm of learning through trial-by-trial learning procedures. Secondly, participants are presented with complex causal models containing confounding pathways. This allows to tap into learners' understanding of the causal logic of confounds because it requires them to disentangle direct causal relations from concurrent covariations arising from confounding variables. Thirdly, the experiments aim to reveal the boundary conditions of the normative causal Bayes nets formalism as a psychological model. Therefore, different types of causal inferences are examined which vary in their complexity and computational demands (e.g., the number of variables that have to be taken into account). Another factor investigated is how manipulations of the way the learning data is presented affect subsequent causal reasoning.

Seeing versus Doing in Trial-by-Trial Learning

In Waldmann and Hagmayer (2005) the parameters could be estimated on the basis of a list of cases which provided simultaneous information about the presence or absence of the variables. It can be argued that this task is still more a reasoning task than a learning task. The typical characteristics of causal learning, in which similar situations are encountered several times, are better mirrored in trial-by-trial learning than in a highly processed list available during the inference process. In fact, Shanks (1991) has hypothesized that induction on the basis of aggregated data is handled by different learning mechanisms than trial-by-trial learning. He argues that the processing of described causal situations and induction from aggregated data is strongly influenced by domain-specific and domain-general knowledge (i.e., top-down processes), whereas associative learning mechanisms (e.g., the R-W rule) are assumed to handle trial-by-trial learning. Thus, associative learning procedures are only activated when similar causal sequences are repeatedly experienced across a period of time. Only when learning takes places in individual trials can the associative weights be modified according to associative learning rules such as the R-W model.

According to this argument, results from experiments providing learners with summarized data (e.g., aggregated lists of cases) or descriptions of causal situations need not provide evidence against associative theories of causal learning. In contrast, a demonstration of the competency to distinguish seeing and doing in the context of trial-by-trial learning would further weaken associative theories as models of causal induction.

Causal Reasoning with Confounds

A further novel aspect of the experiments is the presentation of causal models that contain two parallel pathways representing mutual confounds (i.e., backdoor paths). Methodology textbooks (e.g., Keppel & Wickens, 2000) strongly recommend controlled experiments to avoid problems of confounding. Experiments involve the random assignment of participants to experimental and control groups and a manipulation of the cause variable by an outside intervention. This procedure ensures independence of the cause variable from all other potentially confounding variables. However, in some sciences (e.g., astronomy, epidemiology) and in many everyday contexts controlled experimentation is impossible. Thus, people have to deal with the problem of confounding variables quite often. It is therefore an important question whether learners are sensitive to confounded situations and can draw adequate causal inferences despite confounding.

An example of a causal model with a confounding pathway is the diamond-shaped causal model (cf. Figure 8) in which event *A* can cause the final effect *D* via intermediate variables *B* and/or *C*. Since events *B* and *C* are connected via their common cause *A*, observing either of *D*'s direct causes (i.e., *B* or *C*) also provides evidence about the state of the alternative cause. For example, observing *C* to be present makes it likely that *A* has been present, too, which, in turn, makes it likely that *B* was also present. However, intervening in *C* (i.e., Do c or Do $\neg c$) renders the event independent of its actual cause, event *A*, thereby breaking the correlation of variable *B* and *C*. One additional goal of the experiments was to test whether learners are sensitive to the fact that interventions and observations differ with respect to the way the confounding pathways need to be taken into account.

Competence and Performance in Seeing and Doing

A further issue is to investigate the boundary conditions of causal Bayes nets as a psychological model of causal cognition. The account itself is a formal model providing a normative yardstick to evaluate learners' causal judgments. However, the theory is not a psychological model *per se* because human causal induction is likely to be subject to a great many factors not included in the formalism. For example, in the preceding section diagnostic and predictive inferences exemplified the computational mechanisms of causal Bayes nets. These inferences differ in their complexity because the diagnostic inferences from *C* to *A* involve only a single causal relation whereas the predictive inferences from *C* to *D* require one to take into account the complete causal model (i.e., the confounding backdoor path). The difference in the number of variables and parameters that have to be taken into account might affect learners' competency in inferring the consequences of interventions in accordance with Bayes nets theory. Another potentially relevant factor might be the way the learning data is presented. For example, in trial-by-trial learning it is possible to manipulate the way the covariational data is presented. Whereas in predictive learning experienced order is consistent with causal order (i.e., learning from causes to effects), in diagnostic learning participants first receive information about the effects before being presented with the state of the cause variables. Such a manipulation can provide misleading cues to causality (because the causal model actually entails a different temporal order) or complicate the process of parameter estimation because learners have to mentally reverse the presented information to conditionalize the effect events on their causes. Pitting experienced temporal order against causal order makes it possible to investigate to which extent a formal model provides an adequate description of natural causal induction, or, stated more optimistically, under which conditions human causal reasoning approximates the predictions of a normative model.

5.2 General Method

Each of the experiments consists of three phases. The first stage is the *causal model phase* in which learners are presented with a hypothetical causal model. The causal model informs learners about the structure of the causal system, that is, which variables are causally related. However, they are not informed about the strength of these causal relations (i.e., the parameters of the causal graph). The introduction of the causal model is followed by an *observational learning phase* consisting of several single trials in which learners can estimate the model's parameters by passively observing the operation of the causal system. The observational learning phase is followed by a *test phase* in which participants have to answer several questions corresponding to the different types of causal inferences formalized by causal Bayes nets theory. Thus, their task is to draw causal inferences from observations of and hypothetical and counterfactual interventions in the causal system's variables.

What are the predictions of the discussed theories of causal cognition for the test phase? The fundamental problem is that most accounts (e.g., contingency models) do not distinguish whether a variable is observed or actively generated. Thus, these accounts can only either make identical predictions for observations and interventions or make no predictions about the consequences of interventions at all. However, associative theories have traditionally distinguished observational learning (classical conditioning) from interventional learning (instrumental conditioning). Thus, from an associative learning perspective learners first undergo a classical conditioning phase and are later required to predict the consequences of an instrumental action.[12] One position in the literature is to separate the two types of learning and to assume that different kinds of associations are acquired from observational and interventional learning (cf. Domjan, 2003). However, such an approach implies that without interventional learning (i.e., an instrumental conditioning phase) we cannot predict the consequences of our actions. Alternatively, one could postulate an interaction between associative weights acquired through classical and instrumental conditioning. For example, Rescorla and Solomon (1967) argued that associative knowledge acquired through classical conditioning can influence instrumental learning. Indeed, there is ample evidence that a classically conditioned stimulus can affect instrumental responses, irrespective of whether the classical conditioning phase precedes, follows, or is conducted in parallel to

[12] I here leave the question of counterfactual interventions aside, for they have no counterpart in the associative learning paradigm.

the instrumental learning phase (cf. Domjan, 2003). However, the experimental settings used to investigate the interaction of the two types of learning always involve both a classical and an instrumental conditioning phase, which is not the case in the studies presented here. One could, of course, assume that the weights acquired through observational learning are used to answer the interventional question, but this account fails when the implications of observed and actively generated states of events have different implications. Nor is it possible to postulate a general transfer mechanism because it depends on the causal model whether there is a difference between observations and interventions.

In summary, all models of causal cognition equating observational with instrumental knowledge are likely to produce erroneous predictions when the consequences of interventions have to be derived from observational knowledge. By contrast, causal Bayes nets theory specifies the conditions under which observations differ from interventions and provides computational mechanisms to derive interventional predictions from observational knowledge.

5.3 Overview of Experiments

The presented experiments are divided into three sections, each of which addresses different aspects of causal reasoning. Note that this division does not map directly onto the outlined research questions, which describe the general issues tackled by the experiments. For example, the complexity of the causal inferences is varied in all of these experiments, and each of the experiments employs causal structures with confounding pathways.

The first part, consisting of Experiments 1 to 4, investigates whether learners have the capacity to infer the consequences of hypothetical and counterfactual interventions after a trial-by-trial observational learning phase. The general claim is that causal reasoning is neither determined by top-down influences (i.e., knowledge about causal structure) nor by bottom-up factors (i.e., covariational data) alone. Therefore, Experiments 1 and 2 provide learners with identical learning data but suggest different causal models to the participants. Conversely, in Experiments 3 and 4 learners are presented with identical causal models but the learning data is varied between conditions.

The second part, which consists of Experiments 5 and 6, aims at investigating not the role of the learning data itself but the way the data is presented during observational learning. These studies pit causal order against experienced temporal order during learning (i.e., predictive learning from causes to effects vs. diagnostic learning from effects to causes). It is examined how this manipulation affects learners' competency to derive interventional predictions from observational knowledge.

The last section, consisting of Experiments 7 and 8, further investigates causal reasoning with causal models containing confounding variables. People's understanding of two different types of confounding is examined. Moreover, by presenting learners with competing causal models, both of which are plausible candidates for the available learning data, the studies combine structure induction with causal inferences from observations and interventions. Because interventional predictions crucially depend on causal structure, learners can only give adequate interventional predictions if they successfully identify the graph from which the learning data was generated.

5.4 Causal Reasoning with Observations and Interventions

The goal of Experiments 1 to 4 is to investigate whether people who have passively observed individual trials presenting the states of a complex causal model can later access their causal knowledge to derive observational, interventional, and counterfactual predictions in a fashion anticipated by causal Bayes nets. The experimental manipulations concern the suggested causal models, the learning data, and the complexity of the requested causal inferences.

5.4.1 Experiment 1

In Experiment 1, causal reasoning about a single cause-effect relation within a complex causal model is investigated. This task is especially suited to reveal whether people perform "graph surgery", the building block of more complex interventional inferences involving multiple causal relations and their associated parameters.

After a trial-by-trial observational learning phase participants are requested to draw causal inferences for observations, hypothetical interventions, and counterfactual interventions. In the causal Bayes nets framework counterfactual reasoning is the most complex type of causal inference since it requires the combination of observational inferences with a stage of model modification to predict the consequences of interventions in a counterfactual world. The comparison of the response patterns to the intervention and counterfactual intervention questions makes it possible to explore whether learners distinguish between these two types of interventions.

To examine the role of structural knowledge in causal reasoning in detail all learners receive identical learning input but are suggested different causal models. The goal is to scrutinize how identical covariational information can yield different causal judgments depending on the causal structure assumed to underlie the observed data. If learners' causal inferences are primarily driven by the learning input (i.e., the observed covariations), the response patterns are not expected to differ. In contrast, if the causal model is taken into account the causal inferences should differ in accordance with the causal model.

Method

Participants and Design

Thirty-six undergraduate students from the University of Göttingen participated. They received course credit for participation. Factor 'causal model' was varied between conditions, factors 'type of inference' and 'presence vs. absence of *C*' were varied within-subjects. All participants were randomly assigned to either of the two conditions.

Procedure and Materials

Causal model phase. In the first stage, the causal model phase, learners are presented with a hypothetical causal model. The model informs learners about the structure of the causal system, that is, which variables are causally related. However, they are not informed about the strength of these causal relations (i.e., the model's parameters). The causal graph introduced in this phase is manipulated between conditions. The remaining two phases, the learning phase and test phase, are identical for all participants. All instructions of this and the following experiments were given in German.

The two causal models and the chosen parameterizations are displayed in Figure 11. The two graphs are identical except for the causal relation between *A* and *C*. In condition *A*→*C* (Figure 11a) event *A* is a cause to *C*. In contrast, in condition *C*→*A* (Figure 11b) the causal arrow between *A* and *C* is reversed. Thus, in this model event *A* is not the cause but the effect of event *C*. The remaining causal relations are identical across conditions. With the chosen parameterizations, the two causal graphs generate identical patterns of covariation, that is, they are observationally equivalent (provided no temporal information is available).

However, due to the reversed arrow the two graphs generate different predictions about the consequences of interventions. For example, in condition *A*→*C* observed states of *C* are diagnostic for *A*, whereas intervening in *C* renders the event independent of its actual cause *A*. Thus, this causal model implies a difference between seeing and doing. Contrary to condition *A*→*C*, in condition *C*→*A* both observed and generated values of *C* are diagnostic for event *A* because here the intervention targets the cause variable. Thus, according to causal Bayes nets learners' interventional and counterfactual inferences should differ depending on the suggested causal graph.

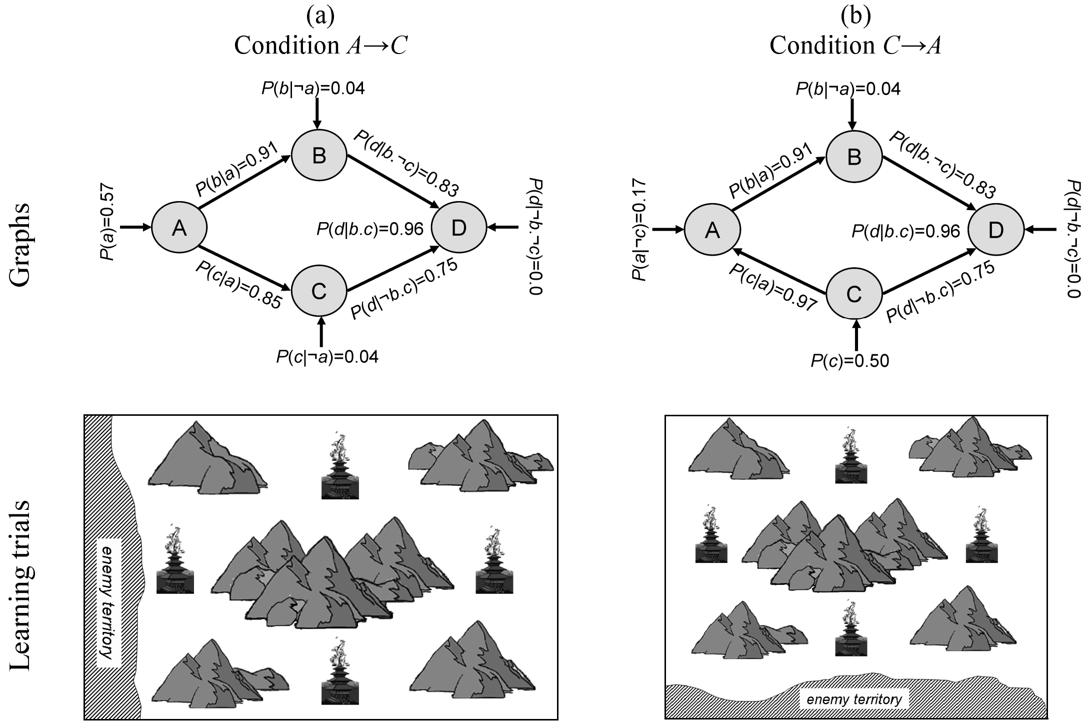


Figure 11. Parameterized causal models in Experiment 1. All parameters were set except $P(d \mid b.\ c)$ which is computed by a noisy-OR-gate. Lower images show screen-shots of learning trials.

To the participants, the causal graphs shown in Figure 11 were introduced as a medieval communication systems transmitting signal fires between four watch towers (see lower images in Figure 11). Each of the variables *A*, *B*, *C*, and *D* was represented by the image of a watch tower. Participants were told that each of the four towers *A*, *B*, *C*, and *D* can light a signal fire. The towers were labeled as eastern, northern, western, and southern tower. The crucial difference between the conditions is the position of the initial event (i.e., the tower initiating the signal transmission). In condition *A*→*C*, participants were informed that watch tower *A* is located close to the enemy's territory and is the initial event of the communication system (bottom left of Figure 11). If the guards at tower *A* observe enemy troops, they light a signal fire on the top of their tower. This fire can be seen from towers *B* and *C*, but not from tower *D*. If the guards at towers *B* and *C* observe the signal at tower *A*, they light up their own signal fires. The signal fires of tower B and C, in turn, can independently cause the final effect, a signal fire at tower D. Thus, there are two independent causal paths from A to D: either via the causal chain $A \rightarrow B \rightarrow D$ or via the alternative path $A \rightarrow C \rightarrow D$, or both. The same cover story was used in condition $C \rightarrow A$, except that here participants were told that C is the initial event of the communication system watching the enemy border (bottom right of Figure 11). Thus, in this model tower D receives the signal either directly via link $C \rightarrow D$ or through the causal chain $C \rightarrow A \rightarrow B \rightarrow D$, or both. To clarify the system's structure, in the instructional phase the paths the signal could take were illustrated by arrows. However, these arrows were not present during observational learning.

Participants were instructed to try to learn the strength of the causal relations from the observational learning phase by requesting them to learn "how well the communication system works". It was also pointed out that the causal relations are probabilistic, for example because bad weather prevents a tower's guards from detecting a signal. The kind of questions they would have to answer after the learning phase was not mentioned until the test phase. To ensure that the instructions were understood correctly participants were requested to briefly summarize the instructions before the learning phase started.

Observational learning phase. The learning phase consisted of 60 trials which implemented the probabilities of the causal graphs shown in Figure 11. The models' parameters were chosen in a way that both graphs generate identical data patterns. Thus, in both conditions participants received the very same learning data as shown in Table 2. Each trial referred to a different day on which learners could observe which towers had lit a signal fire. Information was presented in randomized order on a computer screen displaying the communication system with its four towers (cf. Figure 11). The state of all four towers was displayed simultaneously. Learners could continue at their own pace, but they could not see a trial again.

Table 2
Learning Data in Experiment 1.

Pattern	Frequency	Pattern	Frequency
$a.\ b.\ c.\ d$	*25*	$\neg a.\ b.\ c.\ d$	0
$a.\ b.\ c.\neg d$	*1*	$\neg a.\ b.\ c.\neg d$	0
$a.\neg b.\ c.\ d$	*2*	$\neg a.\neg b.\ c.\ d$	1
$a.\neg b.\ c.\neg d$	*1*	$\neg a.\neg b.\ c.\neg d$	0
$a.\ b.\neg c.\ d$	*4*	$\neg a.\ b.\neg c.\ d$	1
$a.\ b.\neg c.\neg d$	*1*	$\neg a.\ b.\neg c.\neg d$	0
$a.\neg b.\neg c.\ d$	*0*	$\neg a.\neg b.\neg c.\ d$	0
$a.\neg b.\neg c.\neg d$	*0*	$\neg a.\neg b.\neg c.\neg d$	24

Test Phase. Subsequent to the observational learning phase, participants were asked three types of causal inference questions: observational, interventional, and counterfactual questions. All questions provided only information about the state of watch tower *C*, that is, whether there was a signal fire on tower *C* or not. The remaining towers were covered by a circle with a question mark to indicate that their state was not known (Figure 12). For each of the questions learners had to estimate the probability of variable *A*.

For the *observational questions* learners were requested to imagine a new day on which the signal fire at tower *C* was observed to be present [absent] and to estimate the probability of a fire on tower *A*. Thus, these two questions required an estimation of the conditional probabilities $P(a \mid c)$ and $P(a \mid \neg c)$. For the *interventional questions* learners were asked to imagine that the state of tower *C* was either generated or prevented by an intervention. 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 fire wood and therefore no fire could be lit that day. Thus, participants had to estimate $P(a \mid \text{Do } c)$ and $P(a \mid \text{Do } \neg c)$. For the *counterfactual questions* participants were instructed to imagine a counterfactual intervention, that is, an intervention contradicting a factual observation. The questions first stated the actual state of *C*, that is, whether the signal fire on tower *C* was observed to be present or not. Then learners were requested to imagine a counterfactual intervention altering this observation. The counterfactual generative question requested learners to assume that no fire was observed this day on tower *C* (factual observation) but to imagine that on this very day lightning had caused a signal fire (counterfactual intervention). Participants then had to estimate the probability for a fire on tower *A*, that is, they had to estimate the counterfactual probability $P(a \mid \neg c \,.\, \text{Do } c)$. Conversely, the counterfactual inhibitory questions stated that a signal fire was observed to be present at tower *C*. Learners were then asked to imagine that the guards had forgotten to collect new fire wood that very day and to estimate the probability of a fire on tower *A* (i.e., learners were requested to estimate $P(a \mid c.\ \text{Do } \neg c)$).

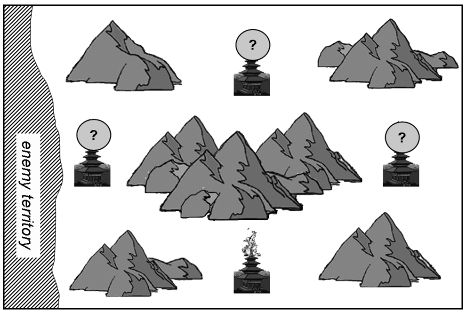


Figure 12. Screen-shot of test phase (condition *A*→*C*).

Estimates of the observational and interventional questions were given on a rating scale ranging from "0 = There definitely was no fire on tower *A*" to "100 = There definitely was a signal fire on tower *A*". The same scale was used for the counterfactual questions but labeled with "0 = There definitely would not have been a signal fire on tower *A*" and "100 = There definitely would have been a signal fire on tower *A*". Interventional, observational, and counterfactual questions were grouped into blocks; the order of blocks was counterbalanced.

Results and Discussion

The results of Experiment 1 are shown in Table 3 along with the normative values derived from causal Bayes nets. Learners' sensitivity to the difference between observations and interventions is tested by comparing their estimates of the observational and interventional questions within conditions; the influence of the causal model can be tested by between-subjects comparisons. Thus, the tests involve both within- and between-subjects comparisons which here are conducted with standard analyses of variance (ANOVA).

Table 3

Mean Probability Judgments for Diagnostic Inference Questions in Experiment 1 ($N = 36$).

		Observation		Intervention		Counterfactual Intervention	
Causal Model		$P(a \mid c)$	$P(a \mid \neg c)$	$P(a \mid \text{Do } c)$	$P(a \mid \text{Do } \neg c)$	$P(a \mid \neg c. \text{ Do } c)$	$P(a \mid c. \text{ Do } \neg c)$
Model *A*→*C*	*Bayes Nets*	*97*	*17*	*57*	*57*	*17*	*97*
	M	78.89	33.89	40.00	43.33	38.33	68.33
	SD	(15.68)	(23.30)	(27.87)	(15.72)	(31.11)	(26.84)
Model *C*→*A*	*Bayes Nets*	*97*	*17*	*97*	*17*	*97*	*17*
	M	81.11	16.67	79.44	14.44	79.44	25.00
	SD	(14.51)	(13.28)	(18.62)	(9.84)	(20.43)	(22.30)

Note. Normative values (range 0 – 100) derived from causal Bayes nets are shown in italics.

Observations vs. interventions. In both conditions, observed states of event *C* are diagnostic for event *A*, that is, *A* is more likely to be present in the presence of *C* than in the absence of *C*. In condition *A*→*C*, reasoning from *C* to *A* corresponds to giving diagnostic judgments. Since observed effects are diagnostic for their causes, learners should infer that event *A* is more likely to be present given that *C* is observed to be present than when *C* is observed to be absent. In accordance with this prediction, in condition *A*→*C* $P(a \mid c)$ received higher ratings than $P(a \mid \neg c)$, $F(1, 17) = 42.30$, $p < .001$, $MSE = 430.88$, $\eta^2 = .71$. Similarly, event *C* is also diagnostic for *A* when the arrow is reversed, that is, when *C* is not the effect but the cause variable (condition

C→*A*). Accordingly, estimates of the observational questions differed significantly in this condition, too, $F(1, 17) = 147.01$, $p < .001$, $MSE = 254.25$, $\eta^2 = .90$.

The crucial test to investigate whether learners differentiate between seeing and doing and perform graph surgery is provided by analyzing learners' responses to the interventional questions. In condition *A*→*C*, intervening in *C* and asking for *A* means manipulating an effect and asking for the probability of its actual cause. Since intervening in an effect renders the event independent of its Markovian parents, participants' estimates for *A* should remain at a constant level for both the generative and inhibitory intervention. As anticipated by causal Bayes nets, learners judged the probabilities $P(a \mid \text{Do } c)$ and $P(a \mid \text{Do } \neg c)$ to be at the same level ($F < 1$). This finding clearly differs from the responses obtained for the observational probabilities and provides strong evidence that learners performed graph surgery. Thus, in contrast to observed values of *C*, in condition *A*→*C* the state of *C* (the effect variable) was not considered to be diagnostic for the cause variable *A* when it was generated by an external intervention. This interpretation is corroborated by the finding that the probability of event *A* being present received higher ratings when *C* was observed to be present than when *C* was generated by an intervention, $F(1, 17) = 38.64$, $p < .001$, $MSE = 352.29$, $\eta^2 = .69$. In accordance with the normative probabilities, *A* was seen to be less likely when *C* was observed to be absent than when *C* was prevented by an intervention. However, due to an overestimation of the observational probability $P(a \mid \neg c)$, the comparison with the corresponding interventional probability $P(a \mid \text{Do } \neg c)$ failed to reach significance, $F(1, 17) = 2.22$, $p = .16$, $\eta^2 = .11$.

A very different response pattern was obtained in condition *C*→*A*. In this condition the variable intervened in, event *C*, is a cause to the variable asked for, event *A*. Since manipulating a cause will alter the probability of its effect(s), a difference for the interventional questions is predicted by Bayes nets theory. The results confirm this prediction. In contrast to condition *A*→*C*, learners' responses to the interventional questions were very similar to the observational question. While in condition *A*→*C* learners judged the interventional probabilities to be at the same level, in condition *C*→*A* the interventional questions differed significantly, $F(1, 17) = 149.46$, $p < .001$, $MSE = 254.41$ $\eta^2 = .90$. Consistent with the predictions and in contrast to the findings of condition *A*→*C* there was neither a significant difference between observing *C* to be present and generating *C* by means of intervention ($F < 1$), nor between $P(a \mid \neg c)$ and $P(a \mid \text{Do } \neg c)$, $F(1, 17) = 1.36$, $p = .26$. This result also refutes the idea that there is a general tendency to answer interventional questions differently from observational questions.

The results so far conformed to the predictions of causal Bayes nets theory and demonstrate that learners correctly recognized that there is a crucial difference between seeing and doing when the intervention targets an effect variable (condition $A \rightarrow C$), but not when the intervention fixes the state of a cause variable (condition $C \rightarrow A$). This interpretation is corroborated by contrasting learners' predictions for the consequences of interventions between conditions. The generative intervention question (i.e., $P(a \mid \text{Do } c)$) received lower ratings in condition $A \rightarrow C$ than in condition $C \rightarrow A$, $F(1, 34) = 24.93$, $p < .001$, $MSE = 561.60$, $\eta^2 = .42$. Conversely, when C was prevented by an intervention (i.e., $P(a \mid \text{Do } \neg c)$) event A was judged to be more likely in condition $A \rightarrow C$ than in condition $C \rightarrow A$, $F(1, 34) = 43.70$, $p < .001$, $MSE = 171.90$, $\eta^2 = .56$.

Hypothetical vs. counterfactual interventions. The findings so far demonstrate that learners derived their interventional predictions in accordance with the hypothesized causal model and distinguished between seeing and doing after a trial-by-trial observational learning phase. A further question is whether learners also correctly differentiated hypothetical interventions referring to the actual world from counterfactual actions. This is revealed by contrasting their responses to the interventional and counterfactual interventional questions within conditions, whereas the influence of manipulations of causal structure can be tested by comparing participants' estimates between conditions.

The counterfactual intervention questions comprise two pieces of information: an observation of C's state in the actual world and a counterfactual intervention which alters this state (cf. Section 4.3.3). For example, the counterfactual probability $P(a \mid c. \text{ Do } \neg c)$ is read as "the probability of A given that C was observed to be present but counterfactually prevented". However, the way observations and interventions have to be combined to infer the consequences of counterfactual actions strongly depends on the underlying causal model. For example, in condition $A \rightarrow C$ counterfactually changing the state of C will not exert any influence on A in a counterfactual world since the (counterfactual) intervention only affects C's descendants. Since the intervention in C will not affect event A, the state of A is identical in both the actual and the counterfactual world. This leads to the counterintuitive prediction that the probability of A is *lower* in the case of a counterfactual generation of C (which logically implies that C was observed to be absent in the actual world) than in the case of a counterfactual prevention of *C* (which logically implies that *C* was observed to be present in the actual world). Consistent with this prediction, in condition *A*→*C* learners judged event *A* to be less likely when *C* was counterfactually generated, $P(a \mid \neg c.\ \text{Do}\ c)$, than when *C* was counterfactually inhibited, $P(a \mid c.\ \text{Do}\ \neg c)$, $F(1, 17) = 10.59$, $p < .01$, $MSE = 764.71$, $\eta^2 = .38$. This is in clear contrast to the responses for the hypothetical intervention questions for which no difference was obtained. In accordance with the normative values, estimates for counterfactual inhibitory intervention, $P(a \mid c.\ \text{Do}\ \neg c)$, were higher than those for the hypothetical inhibitory intervention $P(a \mid \text{Do}\ \neg c)$, $F(1, 17) = 31.61$, $p < .001$, $MSE = 177.94$, $\eta^2 = .65$. However, due to the overestimation of $P(a \mid \neg c.\ \text{Do}\ c)$ no difference was obtained between the factual and counterfactual generation of *C* ($F < 1$).

A very different response pattern is expected for condition *C*→*A*. In this condition the variable intervened in is the cause event of the variable asked for. Therefore, the state of effect variable *A* in the actual world does not correspond to the state in the counterfactual world. First, learners should update the state of event *A* in accordance with the observed state of *C*. However, since *A* is an effect of *C*, the counterfactual intervention influences variable *A* in the counterfactual world (to which the question refers). Therefore, the consequences of counterfactual interventions correspond to those of factual interventions, that is, *A* is more likely to be present in case of a counterfactual generation of *C* than when *C* is counterfactually removed.

In accordance with the normative analysis, the pattern of probability judgments was reversed for condition *C*→*A*. With the alternative causal model guiding the counterfactual causal inferences, learners' estimates of the counterfactual probabilities closely resembled those obtained for the "normal" interventions. In contrast to the findings of condition *A*→*C*, in condition *C*→*A* the counterfactual generation intervention, $P(a \mid \neg c.\ \text{Do}\ c)$, received higher ratings than the counterfactual inhibition intervention, $P(a \mid c.\ \text{Do}\ \neg c)$, $F(1, 17) = 53.85$, $p < .001$, $MSE = 495.43$, $\eta^2 = .76$. Consistent with the normative values, the counterfactual generative intervention did not differ from the factual generative intervention ($F < 1$). However, there was a slight difference between the counterfactual inhibitory intervention and the factual inhibitory intervention, $F(1, 17) = 3.83$, $p = .07$. Nevertheless, the general pattern demonstrates that participants understood that the causal structure of this condition implies that there is no difference between the consequences of hypothetical actions in the actual world and the outcomes of counterfactual interventions.

The influence of the causal model is directly tested by the between conditions comparisons of learners' probability judgments. In line with the normative probabilities, the counterfactual generation question received lower ratings in condition *A*→*C* than in condition *C*→*A*, $F(1, 34) = 21.97$, $p < .001$, $MSE = 692.48$, $\eta^2 = .39$. Conversely, event *A* was judged to be more likely in condition *A*→*C* than in condition *C*→*A* conditional on a counterfactual prevention of *C*, $F(1, 34) = 27.76$, $p < .001$, $MSE = 608.82$, $\eta^2 = .45$.

Taken together, these results show that learners have a remarkable grasp of the difference between hypothetical and counterfactual interventions. Participants successfully derived the consequences of counterfactual interventions in accordance with the suggested causal model. Thus, they understood that the potential differences between hypothetical and counterfactual interventions crucially depend on the underlying causal model.

5.4.2 Experiment 2

Experiment 1 has provided strong evidence that learners distinguish between observations and interventions and perform graph surgery when giving diagnostic judgments. Moreover, the responses to the counterfactual intervention questions show that, as anticipated by causal Bayes nets, learners distinguished hypothetical from counterfactual interventions. The results of the first study also demonstrate the importance of the causal model when inferring the consequences of hypothetical and counterfactual interventions. Even though all learners received identical learning input, their estimates for the interventional and counterfactual probabilities differed depending on the causal structure assumed to have generated the data.

The capacity to perform graph surgery is the building block of more complex causal inferences which require taking into account multiple causal connections and parameters. The goal of Experiment 2 is to show that learners not only understand that intervening in an effect renders it independent of its actual causes but that this knowledge is capitalized on when drawing more complex causal inferences. Experiment 1 investigated reasoning about a single causal relation between the two events *A* and *C*, while the rest of the causal model was irrelevant for the task. Experiment 2 goes one step further and requires participants to estimate the probability of event *D* from observations of and interventions in *C*. Inferring the state of the final effect *D* from observations of and interventions in *C* allows for a profound test of the capacity to distinguish seeing and doing because learners must not only consider the direct causal link between *C* and *D* but also the backdoor path *A*→*B*→*D* (cf. Figure 11). Predicting the state of *D* is therefore much more challenging than the diagnostic inferences examined in Experiment 1.

The difference between observations and interventions in predictive reasoning about *D* is revealed most clearly by contrasting cases in which *C* is observed to be absent and those in which *C* is prevented by an intervention. Since the two causes of *D* (i.e., events *B* and *C*) are connected by their common cause *A*, observed states of *C* also provide information about the instantiation of the alternative causal path (e.g., observed values of *C* are diagnostic for *B*). For example, if *C* is observed to be absent the probability of event *A* is low; therefore it is unlikely that *D* is generated via the alternative causal chain *A*→*B*→*D*. However, the situation is different when *C* is not merely observed to be absent but actively prevented from occurring. Inhibiting *C* by means of external intervention renders variable *C* independent of *A* and ensures that *D* is not generated by *C*. However, this action leaves the backdoor path *A*→*B*→*D* intact. Thus, the probability of *D* occurring is higher when *C* is prevented by an intervention than when *C* is merely observed to be absent (provided the causal chain *A*→*B*→*D* consists of generative causal mechanisms). Correct estimates of the interventional and counterfactual probabilities are only possible when learners are sensitive to the mutual confounder *A* and the backdoor path. Experiment 2 investigates whether learners recognize that interventions and observations differ with respect to the way the second confounding pathway needs to be taken into account.

Method

Participants and Design

Thirty-six undergraduate students from the University of Göttingen participated. None of them took part in Experiment 1. They received course credit for participation. Factor 'causal model' was varied between conditions, factors 'type of inference' and 'presence vs. absence of *C*' were varied within-subjects. All participants were randomly assigned to either of the two conditions.

Procedure and Materials

Experiment 2 used the same cover story and instructions as Experiment 1. In the *causal model phase*, participants were introduced to either of the two causal models depicted in Figure 11. Again, participants were requested to attempt to learn "how well the communication system works". This stage was followed by the *observational learning phase* with the same learning input as in Experiment 1 (cf. Table 3). The observational learning phase was followed by the *test phase* in which the same set of questions (observational, interventional, counterfactual) was used as before. The only difference was that learners were not asked about the initial event *A* but had to estimate the probability of the final effect *D*. Experiment 2 also used the same rating scales as Experiment 1. Again, interventional, observational, and counterfactual questions were grouped into blocks; the order of blocks was counterbalanced across participants.

Results and Discussion

Table 4 shows the results of Experiment 2 along with the normative probabilities derived from causal Bayes nets. Again, learners' sensitivity to the difference between observations and interventions is tested by comparing their estimates of the observational and interventional questions within conditions. The influence of the causal model is tested by contrasting the probability judgments between conditions.

Table 4

Mean Probability Judgments for Predictive Inference Questions in Experiment 2 (N = 36).

		Observation		Intervention		Counterfactual Intervention	
Causal Model		$P(d \mid c)$	$P(d \mid \neg c)$	$P(d \mid \text{Do } c)$	$P(d \mid \text{Do } \neg c)$	$P(d \mid \neg c. \text{ Do } c)$	$P(d \mid c. \text{ Do } \neg c)$
Model $A{\rightarrow}C$	*Bayes Nets*	*97*	*17*	*86*	*44*	*79*	*74*
	M	73.33	33.89	68.33	46.11	68.33	52.78
	SD	(16.45)	(20.33)	(21.76)	(12.43)	(21.21)	(19.04)
Model $C{\rightarrow}A$	*Bayes Nets*	*93*	*17*	*93*	*17*	*93*	*17*
	M	81.11	16.11	83.33	13.33	76.67	22.22
	SD	(14.91)	(11.45)	(14.14)	(9.08)	(23.26)	(26.91)

Note. Normative values (range 0 – 100) derived from causal Bayes nets are shown in italics.

Observations vs. interventions. The causal model's parameters imply that in both conditions event *D* is more likely when *C* is observed to be present than when *C* is observed to be absent. Consistent with this prediction, in condition $A{\rightarrow}C$ learners judged $P(d \mid c)$ higher than $P(d \mid \neg c)$, $F(1, 17) = 27.85$, $p < .001$, $MSE = 502.78$, $\eta^2 = .62$. The same result was obtained in condition $C{\rightarrow}A$, $F(1, 17) = 153.00$, $p < .001$, $MSE = 248.53$, $\eta^2 = .90$.

Whereas both graphs entail that observed states of *C* are diagnostic for the state of variable *A*, the two causal models generate very different predictions for interventions in *C*. In condition $A{\rightarrow}C$ the variable targeted by the intervention is an intermediate event in one of two alternative causal pathways. Due to the mutual confounder, event *A*, observed values of *C* provide information about the backdoor path whereas interventions in *C* "close the backdoor". Even though *D* is still more likely when *C* is generated than when *C* is inhibited, this difference should be smaller than when *C* is passively observed to be present or absent. Consistent with the normative analysis, participants' estimates for the interventional probabilities, $P(d \mid \text{Do } c)$ and $P(d \mid \text{Do } \neg c)$, differed, $F(1, 17) = 9.87$, $p < .01$, $MSE = 450.33$, $\eta^2 = .37$, but as predicted this difference was smaller than the difference between the observational questions, $F(1, 17) = 3.67, p = .07$, $MSE = 364.13$, $\eta^2 = .18$.

The crucial tests concern the single comparisons of merely observing a state of *C* with the very same state set by external intervention. As predicted by causal Bayes nets, learners judged the probability of *D* occurring higher when *C* was prevented by an intervention, $P(d \mid \text{Do } \neg c)$, than when it was observed to be absent, $P(d \mid \neg c)$, $F(1, 17) = 5.13$, $p < .05$, $MSE = 262.09$, $\eta^2 = .23$. In accordance with the parameterization of the causal model, there was only a slight, non-significant difference between $P(d \mid c)$ and $P(d \mid \text{Do } c)$, $F(1, 17) = 1.05$, $p = .32$, $\eta^2 = .06$. These results confirm the predictions of causal Bayes nets theory and demonstrate that learners were remarkably sensitive to the confounding backdoor path.

A very different response pattern was obtained when learners were suggested the alternative causal model of condition *C→A*. Since in this condition the intervention targets the causal model's initial event, it makes no difference whether states of *C* are merely observed or actively generated. Accordingly, participants judged *D* to be more likely when *C* was generated, $P(d \mid \text{Do } c)$, than when *C* was prevented, $P(d \mid \text{Do } \neg c)$, $F(1, 17) = 249.9$, $p < .001$, $MSE = 176.47$, $\eta^2 = .94$. However, in contrast to condition *A→C* the difference between the two intervention questions was as large as the difference between the observational questions ($F < 1$). Moreover, no difference between observing *C* to be present and generating *C* by an intervention, $P(d \mid c)$ and $P(d \mid \text{Do } c)$, nor between observing *C* to be absent and preventing *C* by an intervention, $P(d \mid \neg c)$ and $P(d \mid \text{Do } \neg c)$, was found (both *F*s < 1). Thus, participants recognized that the consequences of observations of and interventions in *C* are identical when the causal arrow between *A* and *C* is reversed.

Taken together, the results convincingly demonstrate learners' capacity to distinguish seeing from doing when drawing complex predictive inferences. The data shows that the capacity to predict the consequences of interventions is not limited to simple diagnostic judgments involving only a single causal relation but was also found when the inferences required taking into account multiple causal relations and confounding backdoor paths. Participants also recognized that the implications of observations and interventions do not differ when the interventions refer to the model's initial event.

Hypothetical vs. counterfactual interventions. A further question is whether participants correctly understood the difference between factual and counterfactual interventions when drawing complex causal inferences. For condition *A*→*C*, the chosen parameterization implies that the probability of *D* is only slightly higher in case of a counterfactual generation of *C* than in case of a counterfactual inhibition of *C* (cf. Table 4). When *C* is observed to be absent but counterfactually generated (i.e., $P(d \mid \neg c.\ \text{Do}\ c)$), the probability of *D* is high because of the direct causal arrow *C*→*D*. However, the probability of *D* is only slightly lower in case of a counterfactual removal of *C* (i.e., $P(d \mid c.\ \text{Do}\ \neg c)$. Here the observed presence of *C* makes it likely that the mutual confounder *A* was present which, in turn, raises the probability that *D* is generated via the backdoor path, even when *C* is counterfactually removed.

Whereas the normative probabilities of condition *A*→*C* imply only a slight difference for the counterfactual question, the obtained probability judgments reveal a substantial difference. The counterfactual generation $P(d \mid \neg c.\ \text{Do}\ c)$ received higher ratings than the counterfactual inhibition $P(d \mid c.\ \text{Do}\ \neg c)$, $F(1, 17) = 6.36$, $p < .05$, $MSE = 342.48$, $\eta^2 = .27$. The descriptive data shows that this effect is mainly due to an underestimation of the probability of *D* in case of a counterfactual removal of *C*. This underestimation also leads to a failure to obtain a reliable difference between estimates of the factual and counterfactual inhibition of *C*. The finding that no difference was obtained between the hypothetical generation, $P(d \mid \text{Do}\ c)$, and the counterfactual generation, $P(d \mid \neg c.\ \text{Do}\ c)$ ($F < 1$) is in accordance with causal Bayes nets theory. Also in line with the normative analysis is that the factual prevention of *C* received lower ratings than the corresponding counterfactual prevention of *C*, but this difference failed to reach significance, $F(1, 17) = 1.70$, $p = .21$. These results indicate that learners had problems differentiating between hypothetical and counterfactual interventions when drawing predictive inferences, which required taking into account the confounding pathway.

In condition *C*→*A* no difference is predicted between hypothetical and counterfactual interventions because the variable intervened in is the causal system's initial event. The results conform to the normative values. In accordance with the causal model induced in this condition, learners correctly inferred that *D* is more likely given that *C* is observed to be absent but counterfactually generated, $P(d \mid \neg c.\ \text{Do}\ c)$, than when *C* is observed to be present but counterfactually removed, $P(d \mid c.\ \text{Do}\ \neg c)$, $F(1, 17) = 29.79$, $p < .001$, $MSE = 895.43$, $\eta^2 = .64$. No difference was found between the factual and counterfactual generation of *C*, $F(1, 17) = 1.58$, $p = .23$, and also no difference was obtained for the factual and counterfactual inhibition of *C*, $F(1, 17) = 2.09, p = .17$.

The data shows that learners responses to the counterfactual questions were also affected by manipulations of the causal model. In accordance with the hypothesized causal model, given a counterfactual removal of *C* (i.e., $P(d \mid c.\ \text{Do}\ \neg c)$ the final effect *D* was judged to be much more likely in condition $A \rightarrow C$ than in condition $C \rightarrow A$, $F(1, 34) = 15.47, p < .001, MSE = 543.30, \eta^2 = .31$. The chosen parameterizations imply only a small difference for the counterfactual generation of *C*. Consistently, $P(d \mid \neg c.\ \text{Do}\ c)$ received only slightly lower ratings in condition $A \rightarrow C$, $F(1, 34) = 1.26$, $p = .27$.

To sum up, in contrast to the findings of Experiment 1 learners had some problems differentiating between hypothetical and counterfactual actions. This is probably due to the increased complexity of the causal inferences, which required taking into account the confounding backdoor path.

5.4.3 Experiment 3

Experiments 1 and 2 have provided convincing evidence that learners differentiate between observations and interventions and can infer the consequences of interventions from passively observed events. This was demonstrated for both simple diagnostic inferences (Experiment 1) and the more complex predictive of Experiment 2.

However, it could be argued that reasoners derived their causal judgments mainly from the suggested causal models without adequately integrating the learning data they were provided with. This would support a top-down account of qualitative causal reasoning and refute the claim that top-down and bottom-up processes interact in a fashion anticipated by causal Bayes nets. Therefore, Experiments 3 and 4 were designed to provide unequivocal evidence that causal reasoning is not driven by the causal structure alone. Learners in Experiments 1 and 2 received identical learning input but were suggested different causal models. In contrast, in Experiments 3 and 4 participants are suggested identical causal structures but are provided with different kinds of data during observational learning. Experiment 3 manipulates base rate information whereas Experiment 4 varies causal strength within the causal model. Thus, the experimental manipulation concerns the causal system's parameters. If learners indeed consider the causal structure as well as the associated parameters then their estimates for the consequences of observations, hypothetical interventions, and counterfactual interventions should be affected by manipulations of the learning input. By contrast, if learners causal inferences are mainly driven by the suggested causal model, variations of the learning input should not affect their causal judgments.

The goal of Experiment 3 is to investigate whether learners are sensitive to manipulations of base rate information and whether they take them into account when predicting the consequences of interventions. Base rates are not only relevant for observational inferences modeled by standard probability calculus (e.g., Bayes theorem) but also have to be considered when deriving interventional probabilities. For example, in Experiment 1 the interventional questions stated that event *C* was fixed by external intervention and learners were requested to estimate the probability of its actual cause *A* (i.e., give estimates of $P(a \mid \text{Do } c)$ and $P(a \mid \text{Do } \neg c)$). The results show that participants correctly judged the probability of event *A* to be at the same level irrespective of whether *C* was generated or prevented by means of intervention. The basic principle that intervening in a variable renders the event independent of its actual causes is, of course, invariant against the absolute size of the cause's base rate. However, since the normative answer to the interventional questions is given by the unconditional probability of event *A* (i.e., $P(a)$), learners should not only judge $P(a \mid \text{Do } c)$ and $P(a \mid \text{Do } \neg c)$ to be equal, but their responses should mirror the absolute value of *A*'s base rate $P(a)$. For example, if variable *C* is fixed by an intervention participants should judge event *A* to be more likely when $P(a) = 0.6$ than when $P(a) = 0.3$.

Manipulations of base rate information should also affect the more complex predictive inferences, which require taking into account the confounding backdoor path. Since the instantiation of the alternative causal chain $A \rightarrow B \rightarrow C$ depends on the probability of the initial event *A*, manipulations of *A*'s base rate should influence causal judgments about the state of the final effect *D*. For example, if *A* is frequent (i.e., has a high base rate) it is more likely that *D* is generated via the alternative causal chain than when *A* is rare (i.e., has a low base rate). Thus, manipulations of *A*'s base rate should affect both diagnostic and predictive judgments.

Method

Participants and Design

Forty-eight undergraduate students from the University of Göttingen, Germany, participated. Factor 'base rate' was varied between conditions, factors 'type of inference' and 'presence vs. absence of C' were varied within-subjects. All participants were randomly assigned to either of the two conditions. Subjects received course credit for participation; none of them took part in Experiments 1 or 2.

Procedure and Materials

Causal model phase. Experiment 3 used the same cover story about the medieval communication system as Experiments 1 and 2. As before, participants were instructed to learn "how well the communication system works". However, in contrast to the previous studies the structure of the causal system was not varied between conditions. Similar to Experiments 1 and 2, the causal model's variables are connected by strong probabilistic causal links. As before, participants were not informed about any of the models' parameters.

The crucial manipulation of this experiment concerns the base rates of the initial event *A* and the variable later intervened in, event *C*. In Experiments 1 and 2, either tower *A* or *C* was close to the enemy territory, indicating that this was the causal model's initial event (conditions *A*→*C* and *C*→*A*, respectively, cf. Figure 11). By contrast, in Experiment 3 learners were instructed that both towers *A* and *C* are close to a border they watch (see Figure 13). If either of these two towers spots enemy troops a fire is lit, resulting in the signal transmission via the other towers (i.e., there are two possible hidden causes which 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).

In condition $A_{\text{high}}C_{\text{low}}$, the initial event *A* has a high base-rate, $P(a) = 0.62$, but the probability of *C* occurring in the absence of *A* is low, $P(c \mid \neg a) = 0.26$. This pattern is reversed in condition $A_{\text{low}}C_{\text{high}}$. In this condition the initial event's base rate is rather low, $P(a) = 0.3$, but there is a strong hidden cause which can generate *C* when *A* is absent (i.e., $P(c \mid \neg a) = 0.55$) (cf. Table 5) . Raising and lowering

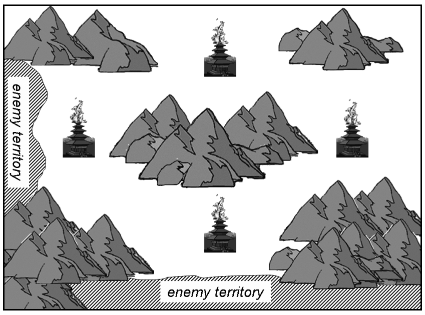


Figure 13. Example of trial in Experiment 3.

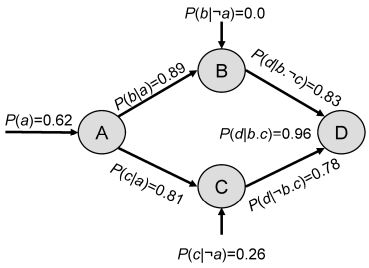
the parameter $P(c \mid \neg a)$ inversely proportional to the base rate of the initial event A allows for keeping the number of cases in which D occurs equal across the two conditions; $P(d) = 0.60$ and $P(d) = 0.58$ in condition $A_{\text{high}}C_{\text{low}}$ and condition $A_{\text{low}}C_{\text{high}}$, respectively. The kind of questions participants would have to answer after the learning phase was not mentioned until the test phase.

Table 5

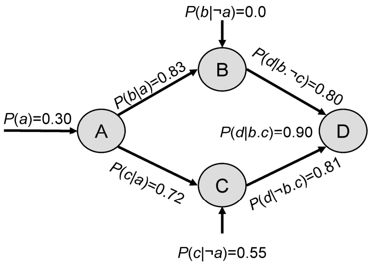
Parametrized Graphs and Learning Data of Experiment 3.

Causal Models

Condition $A_{\text{high}}C_{\text{low}}$



Condition $A_{\text{low}}C_{\text{high}}$



Learning Data

Data Pattern	$A_{\text{high}}C_{\text{low}}$	$A_{\text{low}}C_{\text{high}}$
a. b. c. d	29	9
a. b. c.¬d	1	1
a.¬b. c. d	2	2
a.¬b. c.¬d	1	1
a. b.¬c. d	5	4
a. b.¬c.¬d	1	1
a.¬b.¬c. d	0	0
a.¬b.¬c.¬d	0	0
¬a. b. c. d	0	0
¬a. b. c.¬d	0	0
¬a.¬b. c. d	4	20
¬a.¬b. c.¬d	1	4
¬a. b.¬c. d	0	0
¬a. b.¬c.¬d	0	0
¬a.¬b.¬c. d	0	0
¬a.¬b.¬c.¬d	16	18

Observational learning phase. The instruction phase was followed by the learning phase in which learners passively observed patterns of covariation. While the same hypothetical causal model was suggested to all participants, learning input during observational learning was varied in accordance with the respective model's parameters. A trial-by-trial based learning paradigm was employed with 60 trials implementing the parameters of the two causal models (cf. Table 5). Trials were presented in randomized order. As in Experiments 1 and 2, the trials presented information on a computer screen about the states of the four variables with each trial referring to a new day on which the communication system was observed. During each trial, a picture of the communication system was displayed showing the state of all four watchtowers (cf. Figure 13). Participants could continue at their own pace but were not allowed to refer back to previous trials.

Test phase. The observational learning phase was followed by the test phase in which the same questions (observational, interventional, counterfactual) were asked as in Experiments 1 and 2. In this study learners were requested to draw both diagnostic inferences from *C* to *A* and predictive inferences from *C* to *D*. The questions stated the current status of variable *C* (present vs. absent) and whether the state of tower *C* was merely observed, intervened in, or counterfactually altered. Participants first had to give estimates of the probability of *A* and were then asked about *D* before proceeding to the next question. Thus, one question was diagnostic, the other predictive. In total, each participant had to answer 12 questions (state of *C* × type of question × type of inference). All 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 grouped into blocks; the order of blocks was counterbalanced.

Results and Discussion: Diagnostic Inferences

Table 6 shows the results for the diagnostic inference questions along with the normative values derived from causal Bayes nets.

Table 6

Mean Probability Judgments for Diagnostic Inference Questions in Experiment 3 (N = 48).

		Observation		Intervention		Counterfactual Intervention	
Base Rates		$P(a \mid c)$	$P(a \mid \neg c)$	$P(a \mid \text{Do } c)$	$P(a \mid \text{Do } \neg c)$	$P(a \mid \neg c. \text{ Do } c)$	$P(a \mid c. \text{ Do } \neg c)$
	Bayes Nets	*87*	*27*	*65*	*65*	*27*	*87*
$A_{\text{high}}C_{\text{low}}$	*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)
	Bayes Nets	*35*	*22*	*30*	*30*	*22*	*35*
$A_{\text{low}}C_{\text{high}}$	*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. Normative values (range 0 – 100) derived from causal Bayes nets are shown in italics.

Diagnostic inferences: observations vs. interventions. The within-subjects comparisons of the observational probabilities provide first evidence for how learners' responses were affected by the varying base rates of events *A* and *C*. In condition

$A_{\text{high}}C_{\text{low}}$, a clear difference for the observational probabilities $P(a \mid c)$ and $P(a \mid \neg c)$ was obtained, $F(1, 23) = 6.02$, $p < .05$, $MSE = 291.21$, $\eta^2 = .21$, even though both estimates showed a strong regression tendency (i.e., $P(a \mid c)$ was underestimated and $P(a \mid \neg c)$ was overestimated). Nevertheless, learners' estimates of the observational probabilities mirror the fact that in this condition observed values of C are highly diagnostic for the initial event A. In contrast, when A has a low base-rate and C has a strong hidden cause (condition $A_{\text{low}}C_{\text{high}}$), observed values of C provide only little information about the state of A. In accordance with this prediction, participants in this condition judged probability of A being present given that C was present (i.e., $P(a \mid c)$) only slightly higher than when C was observed to be absent (i.e., $P(a \mid \neg c)$) ($F < 1$).

Learners' sensitivity to base rate information is directly tested by contrasting the probability judgments between conditions. As predicted by causal Bayes nets, $P(a \mid c)$ received higher ratings in condition $A_{\text{high}}C_{\text{low}}$ than in condition $A_{\text{low}}C_{\text{high}}$, $F(1, 46) = 7.55$, $p < .01$, $MSE = 377.81$, $\eta^2 = .14$. In line with the normative predictions only a small, non-significant difference was found for estimates of $P(a \mid \neg c)$, $F(1, 46) = 1.29$, $p = .26$. Thus, participants' responses to the observational questions were clearly affected by variations in the models' parameterizations.

The next analyses concern the question of whether learners distinguished observations from interventions. At variance with the normative predictions, the interventional probabilities in condition $A_{\text{high}}C_{\text{low}}$ differed from each other, $F(1, 23) = 5.21$, $p < .05$, $MSE = 249.91$, $\eta^2 = .19$. A similar result was obtained in condition $A_{\text{low}}C_{\text{high}}$; here the interventional probabilities were also found to differ from each other, $F(1, 23) = 4.53$, $p < .05$, $MSE = 103.53$, $\eta^2 = .16$. A closer inspection of the data revealed that these deviations are mainly due to a small number of participants who strongly underestimated the probability $P(a \mid \text{Do } c)$. However, the comparisons of the observational and interventional probabilities demonstrate that learners differentiated between seeing and doing in diagnostic reasoning. In condition $A_{\text{high}}C_{\text{low}}$, event A was judged to be more likely when C was merely observed to be present (i.e., $P(a \mid c)$) than when C was generated by an intervention, (i.e., $P(a \mid \text{Do } c)$), $F(1, 23) = 9.09$, $p < .01$, $MSE = 264.86$, $\eta^2 = .28$. Conversely, the observational probability $P(a \mid \neg c)$ received lower ratings than the corresponding interventional probability $P(a \mid \text{Do } \neg c)$, $F(1, 23) = 11.50$, $p < .01$, $MSE = 72.46$, $\eta^2 = .33$. The alternative parameterization of condition $A_{\text{low}}C_{\text{high}}$ also implies some differences between the normative values for observations and interventions. Consistent with the causal Bayes nets analysis, in this condition participants judged $P(a \mid c)$ higher than $P(a \mid \text{Do } c)$, $F(1, 23) = 5.66$, $p < .05$, $MSE = 106.43$, $\eta^2 = .20$, but the predicted difference between $P(a \mid \neg c)$ and $P(a \mid \text{Do } \neg c)$ was not found ($F < 1$).

The influence of base rate information on learners' interventional judgments is directly tested by contrasting their interventional probability judgments between conditions. As predicted by causal Bayes nets, both interventional probabilities received higher ratings when the initial event A had a high base rate (condition $A_{\text{high}}C_{\text{low}}$) than when A had a low base rate (condition $A_{\text{low}}C_{\text{high}}$). However, only the contrast for the preventive intervention question (i.e., $P(a \mid \text{Do } \neg c)$) turned out to be significant, $F(1, 46) = 6.91$, $p < .05$, $MSE = 271.56$, $\eta^2 = .13$. Even though descriptively in line with the normative values, the difference for the generative action (i.e., $P(a \mid \text{Do } c)$) failed to reach significance, $F(1, 46) = 1.97$, $p = .17$.

Taken together, participants distinguished between seeing and doing and responded differently to the observational and interventional questions. Moreover, the probability estimates were clearly affected by manipulations of base rate information. This finding refutes the hypothesis that learners' causal judgments are driven by qualitative reasoning alone. However, the finding that the interventional probabilities differed from each other is in conflict with the normative analysis and also with the results of Experiment 1, in which learners correctly judged the interventional probabilities to be at the same level. Since in the present study in both conditions event A was judged to be *more* likely when C was prevented than when it was generated, this finding is not only at variance with causal Bayes nets theory but also cannot be attributed to a failure to distinguish seeing from doing. It is, however, not clear why the generative action received lower ratings than the preventive action. One possible explanation is that some learners confused hypothetical with counterfactual intervention. This would explain why event A was seen to be more likely given C's prevention than given C's generation (see below).

Diagnostic inferences: hypothetical vs. counterfactual interventions. As in Experiments 1 and 2, learners were not only asked to predict the consequences of hypothetical interventions but also requested to estimate the probability of A given counterfactual interventions in C. These judgments, too, should reflect variations of base rate information.

Since the probability of event A has to be updated in accordance with the factual observation of C, in condition $A_{\text{high}}C_{\text{low}}$ event A is more likely given that C is counterfactually inhibited than when *C* is counterfactually generated (because the counterfactual inhibition logically implies that *C* has been observed to be present and the counterfactual generation implies that *C* has been observed to be absent). In accordance with this prediction, in condition $A_{\text{high}}C_{\text{low}}$ participants judged *A* to be less likely given that *C* was counterfactually generated (i.e., $P(a \mid \neg c. \text{ Do } c)$ than when *C* was counterfactually inhibited (i.e., $P(a \mid c. \text{ Do } \neg c)$, $F(1, 23) = 4.19$, $p = .05$, $MSE = 262.95$, $\eta^2 = .15$. In contrast, the parameterization of condition $A_{\text{low}}C_{\text{high}}$ entails only a minor difference between the two counterfactual probabilities. Consistent with this prediction, only a small, non-significant difference was obtained ($F < 1$).

With respect to the between-condition comparisons, causal Bays nets predicts higher ratings for $P(a \mid c. \text{ Do } \neg c)$ in condition $A_{\text{high}}C_{\text{low}}$, but no difference is expected for $P(a \mid \neg c. \text{ Do } c)$. Consistent with these predictions, the counterfactual probability $P(a \mid c. \text{ Do } \neg c)$ received higher ratings in condition $A_{\text{high}}C_{\text{low}}$ than in condition $A_{\text{low}}C_{\text{high}}$, $F(1, 46) = 5.25$, $p < .05$, $MSE = 431.79$, $\eta^2 = .10$, but only a small, non-significant difference was found between conditions for $P(a \mid \neg c. \text{ Do } c)$ ($F < 1$). However, participants failed to differentiate hypothetical from counterfactual interventions. In both conditions the hypothetical and counterfactual interventions received the same ratings, both for the prevention and generation of *C* (all *F*s < 1).

In general, learners' estimates of the counterfactual probabilities matched the normative values better than the responses to the hypothetical interventions. In accordance with the normative analyses, manipulations of the learning data (i.e, base rates) influenced the counterfactual probability judgments. The fact that no difference was found between estimates for factual and counterfactual interventions indicates that learners confused the two types of interventions and treated the hypothetical actions as counterfactual intervention questions.

Results and Discussion: Predictive inferences

Table 7 shows the results for the predictive inference questions along with the normative probabilities. In contrast to the diagnostic judgments, probability estimates for the final effect *D* require that one takes into account the complete model and its parameters. The chosen parameterizations of the two models allow for testing for both learners' capacity to differentiate between seeing and doing and their sensitivity to the diverging base rates of events *A* and *C*.

Table 7

Mean Probability Judgments for Predictive Inference Questions in Experiment 3 (N = 48).

		Observation		Intervention		Counterfactual Intervention	
Base Rates		$P(d \mid c)$	$P(d \mid \neg c)$	$P(d \mid \text{Do } c)$	$P(d \mid \text{Do } \neg c)$	$P(d \mid \neg c.\ \text{Do } c)$	$P(d \mid c.\ \text{Do } \neg c)$
	Bayes Nets	*92*	*23*	*88*	*50*	*80*	*67*
$A_{\text{high}}C_{\text{low}}$	*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)
	Bayes Nets	*84*	*17*	*84*	*20*	*83*	*23*
$A_{\text{low}}C_{\text{high}}$	*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. Normative values (range 0 – 100) derived from causal Bayes nets are shown in italics.

Predictive inferences: observations vs. interventions. Due to the direct causal arrow $C \rightarrow D$ in both models the final effect D is more likely when C is observed to be present than when C is observed to be absent. As predicted, the observational probabilities $P(d \mid c)$ and $P(d \mid \neg c)$ differed significantly, $F(1, 23) = 50.33$, $p < .001$, $MSE = 405.71$, $\eta^2 = .69$ in condition $A_{\text{high}}C_{\text{low}}$, as well as in condition $A_{\text{low}}C_{\text{high}}$, $F(1, 23) = 122.77$, $p < .001$, $MSE = 291.21$, $\eta^2 = .84$. Similarly, due to the direct causal link $C \rightarrow D$, estimates of D should also differ for the interventional questions. Accordingly, a significant difference between the interventional probabilities $P(d \mid \text{Do } c)$ and $P(d \mid \text{Do } \neg c)$ was obtained in condition $A_{\text{high}}C_{\text{low}}$, $F(1, 23) = 35.80$, $p < .001$, $MSE = 293.39$, $\eta^2 = .61$, as well as in condition $A_{\text{low}}C_{\text{high}}$, $F(1, 23) = 52.95$, $p < .001$, $MSE = 377.90$, $\eta^2 = .70$.

The crucial test of sensitivity to the difference between seeing and doing is provided by comparing the probability of D given observations of and interventions in C. To give adequate estimates of D, learners need to take into account the backdoor path $A \rightarrow B \rightarrow D$, especially when preventing C by an intervention. However, the difference between observations and interventions crucially depends on the base rate of the initial event A: whereas the alternative causal pathway is likely to be instantiated when A has a high base rate (condition $A_{\text{high}}C_{\text{low}}$), the influence of the backdoor path can be neglected when the initial event A has a low base rate (condition $A_{\text{low}}C_{\text{high}}$). Therefore, learners in condition $A_{\text{high}}C_{\text{low}}$ should differentiate between observing C to be absent and actively preventing C, whereas no difference is predicted for condition $A_{\text{low}}C_{\text{high}}$. The statistical analyses are in accordance with these predictions: in condition $A_{\text{high}}C_{\text{low}}$ the observational probability $P(d \mid \neg c)$ was judged lower than the corresponding interventional probability $P(d \mid \text{Do} \neg c)$, $F(1, 23) = 4.29$, $p < .05$, $MSE = 194.20$, $\eta^2 = .16$, but no difference was obtained in condition $A_{\text{low}}C_{\text{high}}$, $F(1, 23) = 1.57$, $p = .22$. Thus, participants not only proved to be sensitive to the alternative pathway in general but also understood the importance of *A*'s base rate for the instantiation of the backdoor path. Consistent with the normative analysis, in condition $A_{\text{high}}C_{\text{low}}$ learners gave similar ratings for $P(d \mid c)$ and $P(d \mid \text{Do } c)$ ($F < 1$). In conflict with the normative values is the finding that in condition $A_{\text{low}}C_{\text{high}}$ event *D* was judged to be more likely when *C* was observed to be present than when it was actively generated, $F(1, 23) = 4.95$, $p < .05$, $MSE = 152.08$, $\eta^2 = .18$.

The influence of base rate information is further revealed by contrasting learners' responses to the interventional questions between conditions. Here the crucial comparison concerns learners' estimates of event *D* given the inhibition of *C* (i.e., $P(d \mid \text{Do} \neg c)$ since this probability is most strongly influenced by the backdoor path. In contrast, due to the strong causal arrow *C→D*, no difference is expected for estimates of $P(d \mid \text{Do } c)$. As predicted by causal Bayes nets theory, learners gave equal judgments for the generative interventional question, ($F < 1$), whereas $P(d \mid \text{Do} \neg c)$ received higher ratings when *A* had a high base rate than when *A* had a low base rate, $F(1, 46) = 7.04$, $p = .01$, $MSE = 404.98$, $\eta^2 = .13$.

This finding demonstrates that learners proved to be sensitive to the relevance of the backdoor path in accordance with the base rate information acquired through observational learning. In summary, while some participants had problems to integrate base rate information in their probability judgments, the general response pattern confirms that learners successfully distinguished between seeing and doing and took into account the causal model's parameters.

Predictive inferences: hypothetical vs. counterfactual interventions. Due to the causal arrow *C→D*, event *D* is more likely to occur when *C* is counterfactually generated than when *C* is counterfactually removed, irrespective of the chosen parameterization. This is mirrored in learners' responses to the counterfactual intervention questions: in both conditions $P(d \mid \neg c. \text{Do } c)$ received higher ratings than $P(d \mid c. \text{Do} \neg c)$. In condition $A_{\text{high}}C_{\text{low}}$, the contrast yields $F(1, 23) = 15.38$, $\text{p} < .01$, $MSE = 644.84$, $\eta^2 = .40$, and, consistently, a significant difference was also found in condition $A_{\text{low}}C_{\text{high,}}$ $F(1, 23) = 59.88$, $p < .01$, $MSE = 361.96$, $\eta^2 = .72$. In line with the normative analysis, participants gave higher ratings for $P(d \mid \neg c. \text{Do } c)$ in condition $A_{\text{high}}C_{\text{low}}$ than in condition $A_{\text{low}}C_{\text{high}}$, $F(1, 46) = 5.23$, $p < .05$, $MSE = 460.69$, $\eta^2 = .10$.

Also predicted is the finding that estimates of $P(d \mid \neg c.\ \text{Do}\ c)$ did not differ between conditions ($F < 1$).

The chosen parameterizations do not imply many differences between hypothetical and counterfactual interventions. In condition $A_{\text{high}}C_{\text{low}}$, learners' causal judgments reflect that the probability of *D* occurring is only slightly lower for a counterfactual generation than for a hypothetical generation of *C*, $F(1, 23) = 1.65$, $p = .21$. However, whereas the normative analysis implies that the probability of *D* is higher in case of a counterfactual removal of *C* than when *C* is hypothetically prevented, no difference was found between the two estimates ($F < 1$). In condition $A_{\text{low}}C_{\text{high}}$, the normative values derived for the counterfactual actions are essentially the same as those for the factual interventions. Accordingly, under this parameterization no difference was obtained between hypothetical and counterfactual intervention questions (both *F*s < 1).

5.4.4 Experiment 4

The findings of Experiment 3 show that learners integrate base rate information when drawing causal inferences from observations and interventions. Experiment 4 aims at investigating further the role of variations in a causal model's parameters. While Experiment 3 showed how causal reasoning was influenced by manipulations of the events' base rates, Experiment 4 varies the strength of the causal links connecting the observed events.

The differences between observations and interventions crucially depend on the strength of a causal model's relations. For example, consider the case of an inhibitory action preventing the occurrence of variable *C*. The probability of the final effect *D* then depends on the instantiation of the alternative causal chain *A→B→D*, which is influenced by two factors. First, the base rate of variable *A* determines how likely it is for the chain's initial event to occur. Learners' sensitivity to manipulations of this parameter was investigated in Experiment 3. The second relevant factor is the strength of the causal relations in the chain. For example, with a high base rate of event *A* and a causal path made of strong causal relations, there is a high probability for *D* to be generated via this causal path. In contrast, with the same base rate but a causal chain consisting of rather weak causal arrows, the influence of event *A* on the final effect *D* is attenuated by the weak relations the path is made of. Even though the high base rate makes it likely that the chain's initial event occurs, the influence of *A* on *D* also depends on the strengths of the causal links *A→B* and *B→D*. Thus, in addition to base rates the strength of the causal model's links is also an important factor that has to be considered when predicting the consequences of interventions.

The rationale of Experiment 4 is the same as in Experiment 3. All learners are suggested the same causal model, but the graph's parameters (i.e., the learning input) are varied between conditions. If learners' causal inferences are by and large determined by causal structure, then different parameterizations should not affect the causal inferences. In contrast, if both the model and its parameters are taken into account, learners' estimates of the observational, interventional, and counterfactual probabilities should reflect manipulations of causal strength.

Method

Participants and Design

Thirty-six undergraduate students from the University of Göttingen, Germany, participated. Factor 'parameterization' was varied between conditions, factors 'type of inference' and 'presence vs. absence of *C*' were varied within-subjects. Subjects received course credit for participation. All participants were randomly assigned to either of the two conditions. None of them took part in Experiments 1 to 3.

Procedure and Materials

Causal model phase. The same diamond-shaped causal structure as in Experiments 1 to 3 was used, but this time the four variables of the causal model were introduced as chemical substances causally interacting in wine casks. Each of the substances was given a fictitious label (e.g., Renoxin, Desulfan). Participants were told that substance *A* causes the generation of substances *B* and *C*, each of which can then independently cause the generation of substance *D*. It was also pointed out that the causal relations are probabilistic. In addition, participants were shown the graph of the causal model. They were instructed to attempt to learn the strength of the causal relations from the learning data. The kind of questions they would have to answer after the learning phase was not mentioned until the test phase.

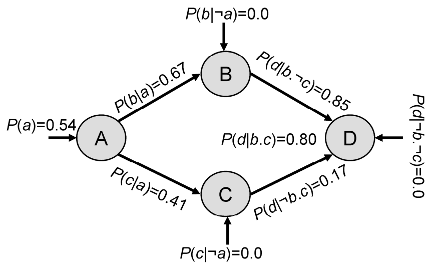
The experimental manipulation lies in the different parameterizations of the two causal chains leading from *A* to *D*. In contrast to Experiment 3, which manipulated base rates but did not vary causal strength within the model, in Experiment 4 there are both strong and weak links connecting the model's variables. Thus, in this study the base rate of the initial event *A* is identical across conditions, but the causal arrows' strengths are manipulated. Table 8 shows the two different parameterizations of the causal graphs along with the data sets generated from the two graphs. In condition *Weak*$_{A\rightarrow C\rightarrow D}$ (top left of Table 8), the causal path $A \rightarrow C \rightarrow D$ consists of weak probabilistic relations while the alternative causal chain $A \rightarrow B \rightarrow D$ is made of strong relations. For the alternative parameterization of the condition *Strong*$_{A\rightarrow C\rightarrow D}$ (bottom left of Table 8), this pattern is reversed. In this condition, the causal path $A \rightarrow C \rightarrow D$ involves strong causal arrows, but the alternative chain $A \rightarrow B \rightarrow D$ comprises only weak probabilistic relations. Because there is always one pathway consisting of strong causal arrows and one chain made of weak arrows, the unconditional probability $P(d)$ is nearly identical across conditions ($P(d) = .36$ and $P(d) = .32$ in conditions *Strong*$_{A\rightarrow C\rightarrow D}$ and *Weak*$_{A\rightarrow C\rightarrow D}$, respectively).

Table 8

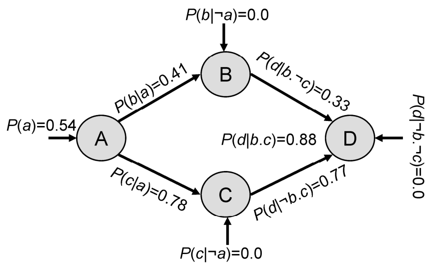
Parameterized Graphs and Learning Data of Experiment 4.

Causal Models

Condition *Weak*$_{A\rightarrow C\rightarrow D}$



Condition *Strong*$_{A\rightarrow C\rightarrow D}$



Learning Data

Data Pattern	*Weak*$_{A\rightarrow C\rightarrow D}$	*Strong*$_{A\rightarrow C\rightarrow D}$
a. b. c. d	4	7
a. b. c.¬d	1	1
a.¬b. c. d	1	10
a.¬b. c.¬d	5	3
a. b.¬c. d	11	1
a. b.¬c.¬d	2	2
a.¬b.¬c. d	0	0
a.¬b.¬c.¬d	3	3
¬a. b. c. d	0	0
¬a. b. c.¬d	0	0
¬a.¬b. c. d	0	0
¬a.¬b. c.¬d	0	0
¬a. b.¬c. d	0	0
¬a. b.¬c.¬d	0	0
¬a.¬b.¬c. d	0	0
¬a.¬b.¬c.¬d	23	23

Observational learning phase. The learning phase consisted of 50 trials in randomized order which implemented the parameters of conditions *Weak*$_{A\rightarrow C\rightarrow D}$ and *Strong*$_{A\rightarrow C\rightarrow D}$, respectively (cf. Table 8). The learning data varied according to the model's parameters. 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 the label of the corresponding substance. 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 information about the four variables was given, that is, which chemicals were present and which were absent. The presence of a chemical substance was depicted by a colored circle, its absence by a crossed-out circle. Figure 14 displays two examples of learning trials.

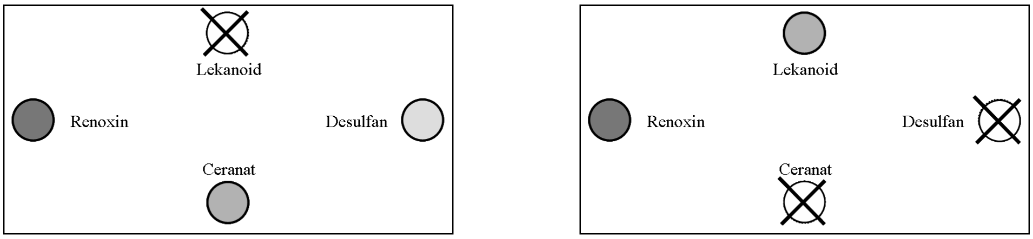


Figure 14. Example of trials in Experiment 4.

Information about the substances was given successively in the temporal order implied by the causal model. Thus, information about the initial event *A* was given first, followed by information about the presence or absence of *B* and *C*, and finally information about *D* was provided. This sequential presentation of information also conforms to the standard design of associative learning experiments. The interstimulus interval was 1 s; after the sequence had finished the information remained for another 2 s on the screen before the next trial began. Participants started each of the trials by pressing the space bar on the keyboard.

Test Phase. In this experiment, learners were only requested to draw predictive inferences. The questions first stated the current status of variable *C* (present vs. absent) and then asked to estimate the probability of variable *D*.

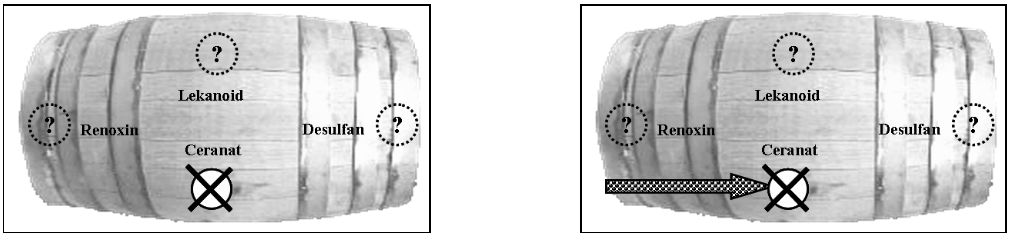


Figure 15. Screen-shots of test phase in Experiment 4. Left: Observing *C* to be absent. Right: Inhibiting *C* by intervention (i.e., Do $\neg c$).

For the *observational questions*, participants were instructed to imagine observing substance *C* in a previously unseen wine cask and then to estimate the probability that substance *D* is present, too (i.e., they estimated $P(d \mid c)$). Analogously, participants were asked to estimate the conditional probability of *D* when *C* was observed to be absent (i.e., $P(d \mid \neg c)$). For the *interventional questions* learners were asked to imagine that *C* was generated or eliminated by an intervention. For the generative intervention, learners were requested to imagine that substance *C* was added to a new wine cask (i.e., $P(d \mid \text{Do } c)$). For the inhibitory interventional question they were asked to imagine that *C* was inhibited from developing by adding a substance called "Anti-*C*" (i.e., $P(d \mid \text{Do } \neg c)$). For the *counterfactual questions* participants were asked to imagine a counterfactual intervention, that is, an intervention contradicting the factual observation of *C* being present or absent. For the counterfactual generative intervention learners were asked to imagine a previously unseen cask in which *C* was observed to be absent, but to suppose substance *C* had been added to this very cask. Thus, this question required to the estimation of the counterfactual probability $P(d \mid \neg c.\ \text{Do } c)$. Conversely, to estimate the counterfactual probability $P(d \mid c.\ \text{Do } \neg c)$, participants were requested to imagine a wine cask in which *C* was observed to be present, but to imagine that the development of substance *C* had been prevented by adding "Anti-*C*" to this cask.

Ratings for the observational and interventional questions were given on a 0 - 100 scale ranging from "0 = *D* is definitely not present" to "100 = *D* is definitely present". The same scale was used for the counterfactual questions but labeled with "0 = *D* definitely would not have been present" and "100 = *D* definitely would have been present". Interventional, observational, and counterfactual questions were grouped into blocks with the order of blocks counterbalanced across participants.

Results and Discussion

Table 9 shows the results for the observational, interventional, and counterfactual inference questions along with the normative probabilities.

Table 9

Mean Probability Judgments for Predictive Inference Questions in Experiment 4 (N = 36).

		Observation		Intervention		Counterfactual Intervention	
Causal Strength		$P(d \mid c)$	$P(d \mid \neg c)$	$P(d \mid \text{Do } c)$	$P(d \mid \text{Do } \neg c)$	$P(d \mid \neg c.\ \text{Do } c)$	$P(d \mid c.\ \text{Do } \neg c)$
	Bayes Nets	*59*	*23*	*39*	*30*	*34*	*56*
$Weak_{A \rightarrow C \rightarrow D}$	*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)
	Bayes Nets	*81*	*03*	*79*	*7*	*78*	*14*
$Strong_{A \rightarrow C \rightarrow D}$	*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. Normative values (range 0 – 100) derived from causal Bayes nets are shown in italics.

Observations vs. interventions. In both conditions the model's parameters imply that *D* is more likely to occur when *C* is observed to be present than when the event is observed to be absent. Learners' responses to the observational questions indicate that they considered observed states of *C* to be diagnostic for event *D*. A significant difference was obtained for the observational probabilities $P(d \mid c)$ and $P(d \mid \neg c)$ in condition $Weak_{A \rightarrow C \rightarrow D}$, $F(1, 17) = 5.62$, $p < .05$, $MSE = 360.29$, $\eta^2 = .25$, as well as in condition $Strong_{A \rightarrow C \rightarrow D}$, $F(1, 17) = 73.04$, $p < .001$, $MSE = 247.97$, $\eta^2 = .81$.

The capacity to differentiate seeing from doing is revealed by comparing learners' responses to the observational and interventional questions. Whereas the chosen parameterizations imply a difference between the observational probabilities in both conditions, the consequences of interventions in *C* crucially depend on the strength of causal chain intervened in. In condition $Weak_{A \rightarrow C \rightarrow D}$ there is only a weak causal relation between the variable intervened in, *C*, and the final effect, *D*. Therefore, the probability of *D* occurring is only slightly higher when *C* is generated than when *C* is prevented. As anticipated by causal Bayes nets, in this condition learners judged the probability of *D* to be at the same level independent of whether *C* was generated (i.e., $P(d \mid \text{Do } c)$) or prevented by means of intervention (i.e., $P(d \mid \text{Do } \neg c)$) ($F < 1$). A different pattern is predicted for condition $Strong_{A \rightarrow C \rightarrow D}$, in which the variable targeted by the intervention is part of the strong causal chain. Due to the strong causal arrow $C \rightarrow D$, event *D* is much more likely to occur when *C* is generated than when the event is prevented. The obtained response pattern matches the normative predictions: contrary to condition $Weak_{A \rightarrow C \rightarrow D,}$ a significant difference was obtained for the interventional probabilities $P(d \mid \text{Do } c)$ and $P(d \mid \text{Do } \neg c)$, $F(1, 17) = 31.85$, $p < .001$, $MSE = 397.43$, $\eta^2 = .65$. This finding indicates that learners differentiated between observations and interventions depending on the causal model's parameters.

This conclusion is corroborated by the comparisons of the observational and interventional probabilities. In condition $Weak_{A \rightarrow C \rightarrow D}$ a substantial difference is predicted between $P(d \mid c)$ and $P(d \mid \text{Do } c)$, but only a slight difference is expected between observing *C* to be absent (i.e., $P(d \mid \neg c)$) and preventing *C* by external intervention (i.e., $P(d \mid \text{Do } \neg c)$). Consistent with this prediction, no difference was obtained for $P(d \mid \neg c))$ and $P(d \mid \text{Do } \neg c)$ ($F < 1$). In contrast, participants judged the probability of *D* considerably higher when *C* was observed to be present $P(d \mid c)$ than when it was generated by an intervention $P(d \mid \text{Do } c)$. However, the obtained difference failed to reach significance, $F(1, 17) = 2.82$, $p = .11$, $\eta^2 = .12$. Contrary to condition

$Weak_{A \to C \to D}$, in condition $Strong_{A \to C \to D}$ no difference between observations and interventions is predicted. In accordance with this prediction learners in this condition neither differentiated between observing C to be present and generating C by an intervention, $F(1, 17) = 1.41$, $p = .25$, nor between observing C to be absent and preventing C by active manipulation ($F < 1$).

Finally, the comparisons of learners' interventional judgments between conditions test for the influence of the causal model's parameters. As anticipated by causal Bayes nets theory, learners' predictions for both the generative and preventive intervention differed depending on the causal model's parameters. The generative interventional question $P(d \mid \text{Do } c)$ received higher ratings in condition $Strong_{A \to C \to D}$ than in condition $Weak_{A \to C \to D}$, but only approached significance, $F(1, 34) = 3.12, p = .09, MSE = 636.53, \eta^2 = .09$. As predicted, the preventive interventional questions $P(d \mid \text{Do } \neg c)$ received lower ratings, $F(1, 34) = 11.67, p < .01, MSE = 225.26, \eta^2 = .26$.

Taken together, the results demonstrate that participants' interventional inferences were sensitive to the causal model's parameters.

Hypothetical vs. counterfactual interventions. Finally, learners' responses to the counterfactual intervention questions are analyzed. For condition $Weak_{A \to C \to D}$, causal Bayes nets theory again predicts that the counterfactual generative intervention, $P(d \mid \neg c. \text{ Do } c)$, should receive lower ratings than the counterfactual inhibitory intervention, $P(d \mid c. \text{ Do } \neg c)$. At variance with this prediction, $P(d \mid \neg c. \text{ Do } c)$ received slightly higher ratings than $P(d \mid c \text{ . Do } \neg c)$, though the difference was not significant ($F < 1$). It is likely that learners confused hypothetical with counterfactual interventions, which is also indicated by the comparisons of learners' estimates for the consequences of hypothetical and counterfactual interventions. The parameters of condition $Weak_{A \to C \to D}$ imply only a small difference between the hypothetical and counterfactual generation of C. In line with this prediction estimates of $P(d \mid \text{Do } c)$ and $P(d \mid \neg c. \text{ Do } c)$ did not differ, $F(1, 17) = 1.35$, $p = .26$. With respect to the prevention of event C, the causal model's parameters imply that the probability of D given a counterfactual inhibition of C, $P(d \mid c. \text{ Do } \neg c)$, is higher than the probability of D conditional on a hypothetical prevention of C, $P(d \mid \text{Do } \neg c)$. Descriptively, the data conformed to this prediction, but the comparison failed to reach significance $F\ (1, 17) = 1.00, p = .33$.

For condition $Strong_{A \to C \to D}$, the model's parameters imply only minor differences between factual and counterfactual interventions. In accordance with the parameterization, the counterfactual generation of C, $P(d \mid \neg c. \text{ Do } c)$, received higher ratings than the counterfactual prevention of *C*, $P(d \mid c.\, \mathrm{Do}\, \neg c)$, $F(1, 17) = 26.71$, $p < .001$, $MSE = 535.47$, $\eta^2 = .61$. Moreover, a difference between factual and counterfactual actions was found neither for the generation of *C* nor for the prevention of *C* (both *F*s < 1). Taken together, the findings indicate that participants failed to differentiate between factual and counterfactual interventions.

However, learners' responses to the counterfactual intervention questions were clearly affected by the parameters associated with the two causal models. Consistent with the values derived from causal Bayes nets, the counterfactual generation of *C* was seen to have more impact on the probability of *D* occurring when *C* was part of the causal pathway made of strong probabilistic relations. Accordingly, estimates of $P(d \mid \neg c.\, \mathrm{Do}\, c)$ were higher in condition $Strong_{A \rightarrow C \rightarrow D}$ than in condition $Weak_{A \rightarrow C \rightarrow D}$, though the difference failed to reach significance, $F(1, 34) = 3.14$, $p = .09$, $MSE = 553.76$, $\eta^2 = .08$. Similarly, responses to the counterfactual prevention question (i.e., $P(d \mid c.\, \mathrm{Do}\, \neg c)$) were lower in condition $Strong_{A \rightarrow C \rightarrow D}$ than in condition $Weak_{A \rightarrow C \rightarrow D}$, $F(1, 34) = 9.01$, $p < .001$, $MSE = 439.39$, $\eta^2 = .21$.

Overall, the analyses of the response patterns obtained for the counterfactual intervention questions indicate that learners were very sensitive to the causal model's parameters, although they did not differentiate between counterfactual and hypothetical interventions.

5.4.5 Experiments 1 to 4: Summary and Discussion

Taken together, the results of Experiments 1 to 4 provide clear evidence that learners successfully distinguished between merely observed states of variables and the very same states generated by interventions. The capacity to distinguish seeing from doing was demonstrated both for simple diagnostic judgments involving a single causal relation and more complex predictive inferences which required taking into account multiple variables and confounding pathways. In all studies participants had the competency to derive interventional predictions from causal models parameterized by passively observed events.

The results of Experiment 1 show that participants understood that intervening in an effect renders it independent of its actual causes, that is, learners performed graph surgery. However, when a different model was suggested in which the variable targeted by the intervention was the cause of the variable asked for, learners correctly understood that there is no difference between seeing and doing. In addition, participants successfully differentiated hypothetical from counterfactual interventions. Even though all subjects received identical learning input, learners' predictions for the consequences of actual and counterfactual interventions differed depending on minimal variations of the causal model assumed to underlie the data.

Experiment 2 extends these results by demonstrating that learners also have the capacity to differentiate between observations and interventions when a confounding backdoor path has to be taken into account. The analyses of the response patterns provide not only convincing evidence that learners correctly distinguished between observations and interventions but also demonstrate a surprising grasp of the implications of confounding pathways in a complex causal model. Participants correctly understood that interventions and observations differ with respect to the way the second confounding pathway needs to be taken into account. In line with causal Bayes nets theory and the results of the first study, assumptions about causal structure strongly influenced learners' causal inferences. Participants correctly recognized that the potentially diverging consequences of observations and interventions crucially depend on the underlying causal model. However, while the results of Experiment 1 demonstrate that participants correctly distinguished hypothetical from counterfactual interventions, this capacity was impaired for the more complex inference tasks of Experiment 2. Even though the descriptive data indicates that participants draw a distinction between hypothetical and counterfactual actions, they had problems to differentiate between the two types of interventions. In general, the results of Experiments 1 and 2 emphasize the role of top-down influences in causal cognition and challenge purely bottom-up approaches of causal induction.

Experiments 3 and 4 further investigated the role of the learning data. The findings of the two studies demonstrate that learners integrate a causal model's parameters in their causal inferences. Since identical causal models lead to very different causal judgments depending on the learning input these results refute the explanation that learners' causal inferences were mainly driven by the causal models suggested to them prior to observational learning. Experiment 3 investigated participants' sensitivity to base rate information in causal reasoning. The results show that learners' responses were clearly affected by variations of the base rate of events *A* and *C*. For example, when the initial event *A* was frequent, it was correctly understood that in case of a preventive intervention in *C* the backdoor path must be taken into account. Conversely, when the instantiation of the backdoor path was rather unlikely because the initial event was rare learners' responses reflected that there was only a slight difference between seeing and doing. However, the obtained response patterns also showed some deviations from the normative values. One potential factor contributing to this problem was that some learners seemed to have misunderstood certain aspects of the cover story (e.g., they assumed the hidden causes generating *A* and *C* not to be independent). In addition, a number of studies on judgment and decision making have shown that people often tend to neglect base rate information, a phenomenon known as *base rate neglect* or *base rate fallacy* (e.g., Eddy, 1982; Kahneman & Tversky, 1982). This might also have contributed to the deviations from the normative values.

Finally, Experiment 4 illustrates how the strength of the causal relations affects causal inferences. As in Experiment 3, participants were suggested identical causal models but the learning input entailed different parameterizations of the causal models. As in Experiment 3, participants' causal inferences varied systematically in accordance with the manipulations of the learning input. For example, learners understood that a confounding backdoor path consisting of strong causal relations will exhibit a larger influence on the probability of the final effect occurring than the same pathway consisting of weak causal relations. Thus, it was correctly recognized that the strength of the causal mechanisms connecting the model's variables crucially influences the consequences of interventions.

Taken together, the findings of the four experiments illustrate how reasoners integrate both qualitative knowledge (i.e., causal models) and quantitative knowledge (i.e., parameters) in their causal judgments. These studies corroborate the assumption that causal reasoning is neither purely data-driven nor completely determined by prior knowledge. Instead, top-down and bottom-up processes interact in causal reasoning in a fashion anticipated by causal Bayes nets theory. Alternative accounts of causal cognition (e.g., contingency models) cannot explain the fact that learners' causal inferences differed depending on whether the state of a variable was merely observed or generated by means of intervention. These accounts fail to take into account the crucial differences between observations and interventions and their diverging implications. The results are also at variance with the predictions of associative accounts. Even though learners never experienced the consequences of interventions (i.e., instrumental actions), they showed a remarkable competency to infer the consequences of hypothetical actions from their observational knowledge.

However, the responses to the counterfactual intervention questions also indicate participants' problems in distinguishing between hypothetical and counterfactual interventions, especially when making judgments that require taking into account confounding backdoor paths. Whereas the estimates of the counterfactual probabilities conformed to the normatively predicted response patterns when giving simple diagnostic judgments, this competency was impaired for the more complex predictive inferences. This is probably due to the complexity of counterfactual inferences, which require an updating of the model's probabilities prior to the stage of model manipulation.

To sum up, the studies provide convincing evidence that learners were able to derive interventional predictions from observational data subsequent to a trial-by-trial learning phase. The results show that the capacity to predict the consequences of interventions from causal models parameterized by passively observed events is not limited to tasks in which learners are provided with lists of aggregated data (Waldmann & Hagmayer, 2005) or description of causal situations (Sloman & Lagnado, 2005). The findings weaken associative theories of causal cognition and are at variance with the claim that trial-by-trial learning operates through different learning mechanisms than causal reasoning with aggregated data. Instead, the results suggest that there are different modes by which identical causal knowledge is accessed and integrated to derive observational and interventional predictions (cf. Waldmann & Hagmayer, 2005). Causal Bayes nets theory, which captures the distinction between seeing and doing and provides the computational mechanisms to derive interventional predictions from observational data, is supported by these results. Theories of causal cognition that lack the representational power to express the differences between observations and interventions and do not take into account causal structure fail to account for the empirical findings.

5.5 Pitting Causal Order against Temporal Order

Experiments 1 to 4 have provided clear evidence that learners distinguish seeing from doing and have the competency to derive interventional predictions subsequent to a trial-by-trial observational learning phase. In accordance with causal Bayes nets theory, these experiments demonstrate that learners use both the causal model and the learning data to infer the outcomes of potential actions.

Experiments 5 and 6 aim at investigating not the role of the learning data itself but the influence of the way the data is presented during observational learning. Studying causal inferences based on trial-by-trial learning allows the introduction of additional temporal cues during observational learning. Thus, the influence of temporal cues during observational learning can be examined because the typical temporal characteristics of causal learning are better mirrored in trial-by-trial learning than in a highly processed list that lacks natural temporal cues.

In psychology, a number of studies have examined the influence of temporal information on elemental causal induction. For example, Michotte's (1963) classical experiments on the perception of causality demonstrated that the longer the delay between the observed events the less likely they are experienced as being causally connected. In a similar vein, Shanks, Pearson, and Dickinson (1989) showed that learners had problems differentiating causal from non-causal actions when the temporal delay between action and outcome was increased. However, Buehner and May (2002, 2003, 2004) point out that longer temporal delays do not always result in decreased judgments of causal strength. Their studies show that the impact of temporal delays on contingency judgments is strongly mediated by the expected timeframe derived from prior knowledge about the causal relation in question (see also Hagmayer & Waldmann, 2002).

Trial-by-trial learning which presents the events in a temporal order not only allows for manipulations of the delay between cause and effect but also makes it possible to pit causal order against temporal order. A causal model not only represents structural relations between variables but also implies a natural temporal order in which we expect the events to occur (i.e., causes prior to their effects). In the real world, causal order is often signaled by the temporal order in which causes and effects are experienced. Therefore, temporal information provides important cues to causality which learners can use to infer causal structure (Lagnado & Sloman, 2004, in press; Lagnado et al., in press). However, we might also experience situations in which the expected and experienced order of events mismatches. For example, physicians often observe symptoms (i.e., effects) prior to learning about their causes. Since temporal order does not mirror causal order in these cases, it is crucial that the experienced temporal order of events is ignored as a cue to causality.

A number of experiments designed to test causal-model theory have pitted temporal order against causal order (Waldmann, 2000, 2001; Waldmann & Holyoak, 1992; Waldmann & Walker, 2005). These studies show that learners are capable of reasoning with causal models regardless of whether temporal order matches or mismatches causal order. However, this competency breaks down when complexity is increased (Waldmann & Walker, 2005). Moreover, the competency was only tested with test questions that requested observational predictions. Interventional questions are more complex because they require a stage of model manipulation (e.g., graph surgery) prior to using the manipulated model for the predictions.

To investigate the influence of misleading temporal cues on causal reasoning about observations and intervention, Experiments 5 and 6 manipulate the way the information is presented during observational learning. Presenting the data in a series of single trials allows for the introduction of potentially misleading cues, for example, by reversing the temporal order. Causal Bayes nets theory assumes that the parameters associated with a causal model reflect the asymmetry between cause and effect, that is, causal strength is encoded as the probability of the effect conditional on its cause(s) and not vice versa. However, even though categories of cause and effect might ultimately determine the way causal strength is represented, temporal cues might nevertheless interfere with the estimation of these parameters. For example, in the diamond-shaped causal model used in the previous studies, the intermediate events *B* and *C* are dependent of the initial event *A*. Similarly, the final effect *D* is dependent on patterns of its two causes *B* and *C*. However, the observed temporal order need not mirror the causal dependencies, for example when the learning order is reversed. Thus, information about the final effect *D* is provided first, then about the intermediate events *B* and *C,* and finally about the initial event *A*. Formally, learners observe the probability of *B* and *C* given *D* and the probability of *A* given *B* and *C* but have to infer the probabilities of *B* and *C* conditional upon *A*, and of *D* conditional upon *B* and *C*.

In such a learning environment, it is crucial to ignore the temporal cues and estimate the parameters according to the causal model. The question pursued in Experiments 5 and 6 is how temporal cues during observational learning influences causal reasoning. Therefore, temporal order during observational learning is manipulated. In Experiment 5, the temporal order during each trial conforms to the causal order of the events in the causal model (i.e., predictive learning from causes to effects). In contrast, in Experiment 6 the temporal order during learning is inconsistent with the order implied by the causal model. (i.e., diagnostic learning from effects to causes). In this experiment, it is necessary to ignore the temporal cues and estimate the parameters in accordance with the initially suggested causal model. Note that the patterns of covariation are nevertheless the same across the two experiments, that is, participants' judgments are based on the very same learning data. Only the order in which information about the events is given is manipulated. Although it is expected that learners will attempt to correctly parameterize the causal model regardless of temporal order during learning, and that they will differentiate between seeing and doing, this competency might be marred by performance deficits caused by the misleading temporal cues (see also Waldmann & Walker, 2005). Because the influence of misleading temporal cues might depend on the complexity of the inference task, both diagnostic judgments and predictive inferences are investigated.

These experiments also pose an interesting challenge to causal Bayes nets theory, because the formalism provides no means to express the difference in experienced learning order. Since the patterns of covariation are identical, the normative probabilities derived from causal Bayes nets are identical regardless of temporal order.

5.5.1 Experiment 5

The goal of Experiment 5 is to investigate whether learners differentiate between seeing and doing after a trial-by-trial learning phase in which learning order corresponds to causal order. Both simple diagnostic and complex predictive judgments are investigated.

Method

Participants and Design

Twenty-four undergraduate students from the University of Göttingen participated.. Factors 'intervention vs. observation' and 'presence vs. absence of *C*' were varied within-subjects. Participants received course credit for participation; none of them took part in one of the other studies.

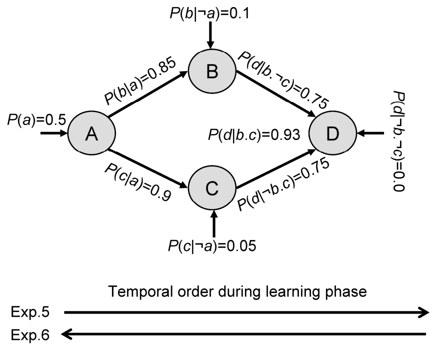
Procedure and Materials

Causal Model phase. The causal model underlying the learning data and its parameterization are shown in Table 10. As in Experiment 4, the variables of the causal model were introduced as four chemical substances causally interacting in wine casks. Participants were presented with the graph of the hypothetical causal model and instructed to attempt to learn the strength of the causal relations from the learning data. The kind of questions they would have to answer after the learning phase was not mentioned until the test phase.

Table 10

Parameterized Causal Model and Learning Data of Experiments 5 and 6.

Graph



Learning Data

Data Pattern	Frequency	Data Pattern	Frequency
a. b. c. d	14	*¬a. b. c. d*	0
a. b. c.¬d	1	*¬a. b. c.¬d*	0
a.¬b. c. d	2	*¬a.¬b. c. d*	1
a.¬b. c.¬d	1	*¬a.¬b. c.¬d*	0
a. b.¬c. d	2	*¬a. b.¬c. d*	1
a. b.¬c.¬d	0	*¬a. b.¬c.¬d*	1
a.¬b.¬c. d	0	*¬a.¬b.¬c. d*	0
a.¬b.¬c.¬d	0	*¬a.¬b.¬c.¬d*	17

Observational learning phase. The learning phase consisted of 40 trials. Table 10 displays the parameterized causal model along with the learning data implementing the probabilities of the graph. As in Experiment 4, each trial referred to a different wine cask. The trials presented information on a computer screen about the states of the four variables; the same symbols were used as in Experiment 4 (cf. Figure 14). The temporal order during each trial conformed to the causal order of the events in the causal model. Information about *A* was given first, followed by information about the presence or absence of *B* and *C*, and finally information about *D* was provided. The interstimulus interval was 1 s. After the sequence, the complete pattern remained for another two seconds on the computer screen.

Test Phase. Subsequent to the observational learning phase, participants were asked to imagine new cases in which either variable *C* was observed to be present or absent, or *C* was generated or prevented by an intervention. Learners had to estimate both the probability of *A* and *D* (i.e., diagnostic and predictive inferences) for observations of and interventions in *C*. For each question, participants were instructed to imagine 40 previously unseen wine casks and to estimate the number of casks in which substance *A* [*D*] would also be found, (i.e., judgments were given in a frequency format). Interventional and observational questions were grouped into blocks with the order of blocks counterbalanced across participants.

Results and Discussion

Diagnostic inferences. The results for the diagnostic test questions are shown in Table 11 along with the normative values derived from causal Bayes nets.

Table 11

Results of Diagnostic Inference Questions in Experiment 5 (N = 24). Numbers Indicate Means of Conditional Frequency Estimates for 40 Cases.

	Observation		Intervention	
	$P(a \mid c)$	$P(a \mid \neg c)$	$P(a \mid \text{Do } c)$	$P(a \mid \text{Do } \neg c)$
Bayes Nets	*38*	*4*	*20*	*20*
M	30.50	17.08	25.54	27.25
SD	(7.56)	(10.37)	(10.57)	(8.59)

Note. Normative probabilities derived from causal Bayes nets are shown in italics (range 0 – 40).

As anticipated by causal Bayes nets, participants gave different estimates for the two observational probabilities but judged the interventional probabilities to be at the same level. There was a significant difference between the observational questions, $F(1, 23) = 35.51$, $p < .001$, $MSE = 59.17$, $\eta^2 = .61$, but no difference between the interventional questions ($F < 1$). In addition, both interventional probabilities differed from their observational counterparts. Substance *A* was judged to be more likely when substance *C* was observed to be present (i.e., $P(a \mid c)$), than when the substance was generated by intervention (i.e., $P(a \mid \text{Do } c)$), $F(1, 23) = 4.61$, $p < .05$, $MSE = 63.93$, $\eta^2 = .17$. Conversely, $P(a \mid \neg c)$ received lower ratings than $P(a \mid \text{Do } \neg c)$, $F(1, 23) = 21.03$, $p < .001$, $MSE = 58.99$, $\eta^2 = .48$. Thus, although participants' estimates did not perfectly match the normative causal Bayes net predictions, the results provide clear evidence for participants' sensitivity to the difference between seeing and doing in diagnostic judgments.

Predictive inferences. The results of the predictive questions for the probability of the final effect *D* are shown in Table 12.

Table 12

Results of Predictive Inference Questions in Experiment 5 (N = 24). Numbers Indicate Means of Conditional Frequency Estimates for 40 Cases.

	Observation		Intervention	
	$P(d \mid c)$	$P(d \mid \neg c)$	$P(d \mid \text{Do } c)$	$P(d \mid \text{Do } \neg c)$
Bayes Nets	*36*	*5*	*33*	*14*
M	29.67	14.79	27.54	21.58
SD	(10.04	(11.56)	(11.64)	(12.55)

Note. Normative probabilities derived from causal Bayes nets are shown in italics (range 0 – 40).

This type of inference is more complicated than the diagnostic judgments. Whereas the latter only requires considering the direct causal relation between *A* and *C* (with the rest of the causal model being irrelevant for the task), predicting *D* from values of *C* requires taking into account the complete model. In particular, the alternative confounding pathway *A*→*B*→*D* needs to be considered.

The causal model's parameters entail a difference both between the observational and the interventional questions, but the difference between the interventional probabilities should be smaller than for the observational probabilities. Consistent with this prediction, there was as significant difference between $P(d \mid c)$ and $P(d \mid \neg c)$, $F(1, 23) = 39.04$, $p < .01$, $MSE = 68.01$, $\eta^2 = .29$, as well as between $P(d \mid \text{Do } c)$ and $P(d \mid \text{Do } \neg c)$, $F(1, 23) = 6.37$, $p < .05$, $MSE = 66.84$, $\eta^2 = .22$. In accordance with the normative values, the difference between the interventional questions was smaller than for the observational questions, $F(1, 23) = 8.73, p < .01, MSE = 54.65, \eta^2 = .28$

Participants' sensitivity to the difference between seeing and doing is directly tested by comparing their estimates of the observational and interventional probabilities. As predicted by causal Bayes nets, there was only a slight, non-significant difference between $P(d \mid c)$ and $P(d \mid \text{Do } c)$, $F(1, 23) = 1.0$, $p = .33$. The crucial test of the predictions of causal Bayes nets concerns the comparison of $P(d \mid \neg c)$ and $P(d \mid \text{Do } \neg c)$ (i.e., merely observing *C* to be absent versus actively preventing *C*). Participants judged the probability of the occurrence of *D* to be significantly higher when *C* was prevented by an intervention than when it was merely observed to be absent, $F(1, 23) = 9.57$, $p < .01$, $MSE = 57.83$, $\eta^2 = .29$. This test shows that learners differentiate seeing from doing and take into account the alternative causal chain *A*→*B*→*D* when estimating the probability of *D* given interventions in *C*.

5.5.2 Experiment 6

Whereas in Experiment 5 learning order matched causal order, in Experiment 6 the experienced temporal order is reversed (i.e., diagnostic learning from effects to causes). During observational learning, participants first receive information about *D*, then about the intermediate events *B* and *C*, and finally the state of the initial event *A* is presented. Thus, the experienced temporal order of events is inconsistent with the temporal order implied by the presented causal model. Instead, the temporal order is consistent with a causal model in which variable *D* is the initial event causing the final effect *A* via the intermediate variables *B* and *C*.

Reversing the temporal order also allows for a determination of whether the inconsistent temporal cues mislead learners to induce an alternative causal model or whether the modified learning procedure prevents an adequate acquisition of the model's parameters. If the temporal cues override the initially instructed causal model, learners should induce an alternative causal model in which *D* is the initial cause and *A* the final effect. Whether learners adhere to the original model or induce a model according to the experienced temporal order is revealed by examining their interventional judgments. In the initially instructed model, they should perform graph surgery for the interventional questions when estimating the state of variable *A*. In contrast, if the temporal cues induce an alternative causal graph in which *D* is a cause to *A*, participants should perform graph surgery when asked to estimate the probability of variable *D* subsequent to an intervention in event *C*.

Method

Participants and Design

Twenty-four undergraduate students from the University of Göttingen participated. They received course credits for participation. Factors 'intervention vs. observation' and 'presence vs. absence of *C*' were varied within-subjects.

Procedure and Materials

Experiment 6 used the same cover story, instructions, and learning input as Experiment 5. The crucial difference concerns the temporal order during observational learning. Whereas in Experiment 5 experienced order matched causal order, in this experiment, participants first received information about the final effect *D*, then about the states of the intermediate variables *B* and *C*, and finally information about the presence or absence of the initial cause *A*. Thus, the experienced order is inconsistent with the temporal order entailed by the instructed causal model. The same set of observational and interventional questions was asked as in Experiment 5.

Results and Discussion

Diagnostic Inferences. Table 13 shows the means of the conditional frequency estimates for the diagnostic inference questions.

Table 13

Results of Diagnostic Inference Questions in Experiment 6 (N = 24). Numbers Indicate Means of Conditional Frequency Estimates for 40 Cases.

	Observation		Intervention	
	$P(a \mid c)$	$P(a \mid \neg c)$	$P(a \mid \text{Do } c)$	$P(a \mid \text{Do } \neg c)$
Bayes Nets	*38*	*4*	*20*	*20*
M	33.46	15.38	25.50	22.42
SD	(8.59)	(11.20)	(11.16)	(10.31)

Note. Normative probabilities derived from causal Bayes nets are shown in italics (range 0 – 40).

Similar to Experiment 5, the responses to the observational questions differed significantly, $F(1, 23) = 63.88$, $p < .001$, $MSE = 61.43$, $\eta^2 = .74$, whereas learners judged the probability of *A* to be at a similar level regardless of whether *C* was interventionally generated or prevented, $F(1, 23) = 1.52$, $p = .23$. Normatively correct, both interventional probabilities differed from their observational counterparts. Learners gave higher ratings for *A* when *C* was observed to be present (i.e., $P(a \mid c)$) than when *C* was generated by intervention (i.e., $P(a \mid \text{Do } c)$), $F(1, 23) = 21.28$, $p < .001$, $MSE = 35.72$, $\eta^2 = .48$. Conversely, the observational probability $P(a \mid \neg c)$ received lower ratings than the interventional probability $P(a \mid \text{Do } \neg c)$, $F(1, 23) = 13.15$, $p < .001$, $MSE = 45.24$, $\eta^2 = .36$. Thus, despite the misleading temporal cues during observational learning, participants' estimates show that they differentiate between observed values of variables and the very same states generated by interventions.

Predictive inferences. As in the previous studies, participants were also asked to estimate the probability of *D* conditional on observations of and interventions in variable *C*.

Table 14

Results of Predictive Inference Questions in Experiment 6 (N = 24). Numbers Indicate Means of Conditional Frequency Estimates for 40 Cases.

	Observation		Intervention	
	$P(d \mid c)$	$P(d \mid \neg c)$	$P(d \mid \text{Do } c)$	$P(d \mid \text{Do } \neg c)$
Bayes Nets	*36*	*5*	*33*	*14*
M	30.54	18.33	29.71	20.33
SD	(11.00)	(13.26)	(10.90)	(11.74)

Note. Normative probabilities derived from causal Bayes nets are shown in italics (Range 0 – 40).

In contrast to Experiment 5, the results for the predictive inferences deviated from the causal Bayes net predictions (see Table 14). In accordance with the normative analysis, $P(d \mid c)$ received higher rating than $P(d \mid \neg c)$, $F(1, 23) = 17.88$, $p < .001$, $MSE = 100.04$, $\eta^2 = .43$, and $P(d \mid \text{Do } c)$ received higher rating than $P(d \mid \text{Do } \neg c)$, $F(1, 23) = 16.88$, $p < .001$, $MSE = 62.47$, $\eta^2 = .42$, but the predicted interaction was not found ($F < 1$).

In accordance with the normative values, learners judged *D* to be equally likely conditional on observing *C* to be present and generating *C* by means of intervention ($F < 1$). However, the crucial test concerns participants' causal judgments about *D* given that *C* is observed to be absent or actively prevented. Even though the estimates for observing *C*'s absence, $P(d \mid \neg c)$, are slightly lower than the estimates for the prevention of *C,* $P(d \mid \text{Do } \neg c)$, the difference failed to reach significance ($F < 1$). Thus, participants failed to differentiate between seeing and doing in the predictive task.

However, because learners correctly assumed that intervening in *C* would influence *D,* the general pattern of results shows that they reasoned in accordance with the initially instructed graph instead of inducing a new causal model.

5.5.3 Experiments 5 and 6: Summary and Discussion

Experiments 5 and 6 aimed at investigating how temporal cues during observational learning influence participants' performance when drawing causal inferences with varying complexity. The results show that learners correctly distinguished between observations and interventions when drawing diagnostic inferences, irrespective of the experienced learning order. Thus, learners successfully recognized that intervening in an effect renders the variable independent of its actual causes.

However, Experiment 6 also shows that the competency of learners only displays itself when the complexity of the task does not exceed learners' information processing capacity (see also Waldmann & Walker, 2005). A popular strategy to deal with such impairments is to postulate two systems, a rule-based component that handles summarized data and an associative learning component that is specialized for trial-by-trial learning (e.g., Price & Yates, 1995; Shanks, 1991). Although the results of Experiments 1 to 5, which also used a trial-by-trial learning procedure, already weaken this account, it might still be speculated that learners fell back on an associative mode in Experiment 6. However, the data of the study are inconsistent with this theory, too. Learners were not generally impaired, only the predictive inferences were affected. The less complex diagnostic inferences showed a remarkable grasp of the differences between seeing and doing despite the misleading temporal cues. The estimates show that basic inference procedures (i.e., graph surgery) were not impaired by the misleading temporal cues during observational learning. Only the more complex predictive inferences were negatively affected.

The reason for the differences between Experiment 5 on the one hand and Experiments 6 on the other is likely to be located in the parameter estimation processes. For example, learners in Experiment 5 observed the probability of *D* given *B* and *C* but participants in Experiment 6 observed the probability of *B* and *C* given *D*. Since correct answers to the predictive inference questions require an estimation of *D* given *B* and *C* as a parameter of the causal model, only learners in Experiment 5 could estimate this conditional probability directly from their learning experience. Therefore, the learning process in the studies with the misleading temporal cues may have led to inadequate estimates of the model's parameters. The diagnostic questions could be correctly answered by recognizing that interventions render the manipulated variables independent of their actual causes, which implies that solely the base rate *P(A)* needs to be accessed in order to give a correct response. In contrast, the predictive questions can only be answered adequately if both the model is correctly altered for the intervention questions *and* the parameters of the full causal model are correctly estimated. Thus, if the parameters are not acquired correctly during learning, the inferences are likely to be wrong.

5.6 Understanding the Causal Logic of Confounds

Experiments 1 to 6 have demonstrated that people have the capacity to infer the consequences of interventions from causal models parameterized by passively observed events. Participants' responses showed that they not only differentiated between observations and interventions, but they also took into account spurious relations that were generated by a confound. These findings provided first evidence that participants understand the causal logic of confounding and are able to separate a direct causal influence from a concurrent spurious relation.

The goal of Experiments 7 and 8 is to investigate further learners' understanding of and reasoning with different types of confoundings. Two basic types of confounding are investigated, common-cause confounding and causal-chain confounding, which differ in terms of the underlying causal model and their implications for the consequences of interventions.

Spurious Correlations and Confounds

An important task in causal induction is to separate spurious correlations from causal relations. A statistical relation observed between *C* and *E* not only may reflect a direct causal relation but a spurious relation due to other, confounding variables. For example, in the 1950's, a series of studies (e.g., Doll & Hill, 1956) with non-experimental data was published showing that lung cancer was found to be more frequent in smokers than in non-smokers. This data was interpreted as evidence that smoking is a cause of lung cancer. However, some prominent statisticians (e.g., Fisher, 1958) argued that such a conclusion was not justified on the basis of the available data. Fisher offered an alternative causal model in which the observed covariation was not interpreted as a direct causal relation but as a spurious correlation generated by a hidden common cause, a genotype causing both a craving for nicotine and the development of lung cancer. This model, too, implies that smoking and cancer covary but denies that there is a direct causal relation.

The example illustrates the necessity of taking into account common causes to distinguish between spurious and causal relations. However, the existence of a common cause does not rule out that there is a direct causal relation as well. This is a particularly interesting situation, because the common cause generates a spurious correlation that superimposes itself upon and distorts the genuine causal relation. The causal relation is then said to be confounded. The detection and analysis of such confounded causal relations is especially challenging because unconfounded estimates of causal strength and correct interventional predictions require us to disentangle the direct causal relation from the concurrent spurious correlation.

Confounds

Confounding variables are statistically related to both the potential cause *C* (independent variable) and the presumed effect *E* (dependent variable). It is the relation between the confounding variable and the cause that creates serious problems. In the most extreme case, the cause and the other variable are perfectly confounded, that is, they are either both present or both absent all the time. In this case it is impossible to tell whether the effect is generated by the cause or by the confounding variable. Therefore it seems to be necessary to eliminate the relation between the candidate cause and the confounding variable. Note that the problem of confounding does not originate in the relation between the extraneous variable and the effect. Even if the extraneous variable has a very strong influence, the impact of the cause variable can be detected as long as the extraneous variable is not permanently present and the cause variable and the extraneous variable are independent of each other. Under these circumstances, the impact of the cause variable can be seen as an increase (generative influence) or decrease (inhibitory influence) of the probability of the effect given the presence of the cause.

The recommended method to avoid confounding and discriminate between candidate causal models is to run randomized experiments, in which the putative cause is manipulated by external intervention (e.g., Fisher, 1951). Accordingly, to test whether a correlation indicates a genuine causal relation we should intervene in the putative cause and scrutinize whether this intervention exerts an influence on the effect event. This procedure ensures the independence of the cause variable from all other potentially confounding variables. However, controlled experiments are not the only way to avoid confounding. Observational studies in combination with other control techniques (e.g., holding constant hypothesized confounds) may also allow us to draw valid causal inferences, which is especially important when controlled experimentation is not possible (e.g., in astronomy or epidemiology). Causal Bayes nets theory provides a formal means to represent confounding, and specifies under which conditions causal inferences can be drawn from observational data in spite of confounding variables.

Common-Cause and Causal-Chain Confounding

Two basic causal structures may underlie confounding. One possibility is that the confounding variable *X* is a cause of both the candidate cause variable *C* and the effect variable *E* (*common-cause confound*, left-hand side of Figure 16). A second type of confounding is a causal-chain model in which the cause variable *C* not only directly influences the effect *E* but also generates the confounding variable *X* which, in turn, influences *E* (*causal-chain confound*, right-hand side of Figure 16). The crucial point is that both models imply a spurious relation between cause and effect even when there was no direct causal relation between them. If the confounding variable *X* is present, both the cause and the effect should tend to be present; if *X* is absent both *C* and *E* should tend to be absent. In addition to the causal relations connecting the confounding variable to *C* and *E*, there is a direct causal relation between *C* and *E* whose existence and strength has to be identified.

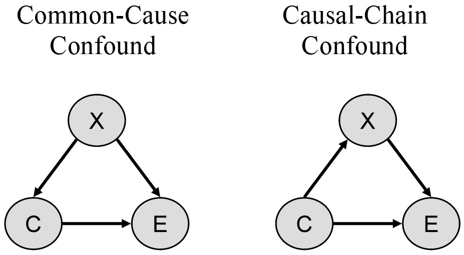


Figure 16. Two types of confounding.

The two models represent two different kinds of confounding. The common-cause confound model represents the situation in which some extraneous variable is causally affecting both the cause and the effect. The hypothesis that smoking and lung cancer are both caused by a specific genotype exemplifies this type of confounding. There are several possibilities to eliminate the causal relation among the common cause *X* and the candidate cause *C*. For example, *X* might be eliminated or held constant (e.g., only people without the carcinogenic genotype are studied). In addition, *C* might be manipulated independently of *X* (which, in the case of smoking, would not be possible for ethical reasons). Such an independent manipulation is equivalent to a randomized experiment (Fisher, 1951).

However, controlled experimentation cannot eliminate causal-chain confounding. This type of confounding calls for other controls because a manipulation of *C* would directly affect *X*. Thus, other ways have to be found to block the causal relation connecting the cause *C* to the confound *X*. For example, aspirin (*C*) might not only have a direct influence on headache but also make your blood thinner (*X*), which, in turn, might also have an impact on your headache (*E*). One way to get rid of confounding in this case is to administer aspirin to people who have thin blood anyway or who are resistant against the side effect, which is equivalent to holding the confound constant. Another possibility is to manipulate the confounding variable in addition to the cause variable and thereby eliminate their causal relation.

In summary, there are two different fundamental types of confounding which call for different measures of control. Manipulations of the putative cause (*C*) eliminate common-cause confounding because the intervention disconnects the event intervened in from the common-cause. However, this is not true for causal-chain confounding because manipulations of the candidate cause will also affect the confounding variable.

Causal Bayes nets allow for representing causal structures with confounding variables. Provided certain conditions are met, causal Bayes nets also enable valid inferences about the existence and strength of confounded causal relations. Experiments 7 and 8 intend to investigate participants' understanding of the two types of confounding. Whereas Experiment 7 confronts learners with common-cause confounding, Experiment 8 focuses on their understanding of causal-chain confounding.

5.6.1 Experiment 7

The goal of Experiment 7 is to investigate people's causal reasoning with common-cause confounds. In contrast to the previous experiments, in which the true causal model was known prior to observational learning, in this study learners are presented with two competing candidate models, a common-cause model and a common-cause confound model (cf. Figure 17). Participants' task is to find out whether there is a direct relation between events *C* and *E*, which are known to be causally connected by a common cause *X*. Since the consequences of interventions crucially depend on which of the two models underlies the learning data, correct interventional predictions are only possible if the participants identify the causal structure underlying their observations. In particular, learners have to differentiate between spurious correlations and causal relations to decide which causal model underlies the observed phenomena. The observational data can then be used to parameterize the chosen causal model and to infer the consequences of interventions.

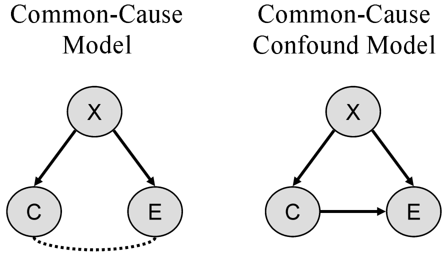


Figure 17. Causal models in Experiment 7.

The common-cause and the common-cause confound models depicted in Figure 17 can be decomposed by applying the causal Markov condition to the graphs. According to the causal Bayes nets formalism, the joint distribution of the common-cause model is factorized into

$$P(X.C.E) = P(X) \cdot P(C \mid X) \cdot P(E \mid X) \qquad (25)$$

whereas the common-cause confound model is decomposed into

$$P(X.C.E) = P(X) \cdot P(C \mid X) \cdot P(E \mid X.C). \qquad (26)$$

These factorizations mirror the causal structure of the two models, which differ in terms of the existence or non-existence of the causal relation *C*→*E*. Accordingly, in the common-cause model event *E* is only conditionalized on *X*, while in the decomposition of the common-cause confound model event *E* is conditionalized on both *X* and *C*.

Thus, the crucial difference between the two factorizations is whether the probability of *E* only depends on the occurrence of *X* (as is the case in the common-cause model) or on the influence of both *C* and *X* (as is the case in the confound model). This, in turn, has consequences for the computation of the interventional probabilities. In the common-cause model, interventions in the candidate cause *C* render the event independent of its cause *X* but will not influence *E* because *C* and *E* are only spuriously correlated. For example, in the common-cause model the probability of *E* given that *C* is generated by an intervention is formalized by

$$P(e \mid \text{Do } c) = P(x) \cdot P(e \mid x) + P(\neg x) \cdot P(e \mid \neg x) \qquad (27)$$

Equation reflects the fact that interventions in *C* do not affect *E*, because the two events are only spuriously related due to their common cause *X*. Thus, the probability of *E* occurring conditional on an intervention in *C* is solely determined by the base rate of the confound, *P*(*x*), the strength of the causal relation between *X* and *E*, *P*(*e* | *x*), and the probability of *E* occurring in the absence of *X*, *P*(*e* | ¬*x*). In contrast, in the common-cause confound model intervening in *C* also renders the event independent of *X* but furthermore influences *E*, because there is a direct causal relation *C*→*E*. Therefore, the probability of *E* given that *C* is actively generated is formalized by

$$P(e \mid \text{Do } c) = P(x) \cdot P(e \mid x.c) + P(\neg x) \cdot P(e \mid \neg x.c) \qquad (28)$$

According to this computation, the probability of event *E* is not only determined by the base rate of *X* and the strength of the causal relation *X*→*E*, but is also influenced by the causal arrow *C*→*E*.

If the parameters of the decomposed model can be estimated from the available data, it is possible to give unconfounded estimates of causal strength and predict the consequences of interventions from observational data. However, certain conditions have to be met in order to infer the model's parameters from observational data. First, the state of the confounding variable *X* must usually be observable.[13] However, the crucial condition is that there is some variation in the confounding variable *X*, that is, *X* must not always be present, and that the confounding variable *X* is not the only cause of *C*. This is essential because the influence of *C* on *E* cannot be evaluated when the confound and the candidate cause are perfectly correlated.[14] In contrast, if there are cases in which *C* occurs in the absence of *X*, the two models can be differentiated by observing whether *C* can generate *E* in the absence of the confound.

Returning to the smoking/cancer debate, if no experimental data is available the adequacy of Fisher's common-cause model can be tested if and only if a) the genotype can be measured and b) there are people who do not have the genotype but smoke (i.e., $P(\text{smoking} \mid \neg\text{genotype}) > 0$). If there exists a direct causal mechanism relating smoking to cancer, then the probability of getting cancer should be higher for smokers than for non-smokers in the population of people without the gene. This example also points out the relation between the conditional contingency model (cf. Section 3.2.2) and causal Bayes nets theory. Whereas both the common-cause and the common-cause confound model entail that there is a positive unconditional contingency between events *C* and *E*, the conditional contingency between *C* and *E* is only positive in the common-cause confound model. Formally, $P(e \mid c) > P(e \mid \neg c)$ holds for both models, but $P(e \mid \neg x.\ c) > P(e \mid \neg x.\ \neg c)$ holds only if there exists a direct causal relation $C \rightarrow E$ (i.e., in the common-cause confound model).

To sum up, it is possible to differentiate the two candidate models by controlling for the confounding variable *X*. If learners recognize this, they can differentiate the two models and estimate the model's parameters from the available observational data. This, in turn, allows for the derivation of interventional predictions.

[13] Pearl (2000) shows that valid causal inferences are sometimes possible even when the confounding variable cannot be observed.

[14] Note that the strength of the causal relation $X \rightarrow C$ is irrelevant. For example, if there is only a weak link connecting the two events, *X* and *E* will often be present in the absence of *C*, but this does not allow one to assess whether there is a causal relation $C \rightarrow E$. Conversely, even when *X* deterministically causes *C*, the influence of *C* on *E* can be evaluated as long as there are cases in which *C* occurs in the absence of *X*.

Method

Participants and Design

Thirty-six students from the University of Göttingen, Germany, participated in this experiment. They were randomly assigned to the common-cause or the common-cause confound condition. Factor 'learning data' was varied between conditions, factors 'type of inference' and 'presence vs. absence of *C*' were varied within-subjects. Subjects received course credit for participation, none of them took part in the previous studies.

Procedure and Materials

Causal model phase. Participants were told that ornithologists had recently discovered a new species of birds. While investigating the new species the researchers noticed that not all birds breed. Since it is known from some other species that birdsong is a relevant factor for mating and breeding, the biologists hypothesized that in this species singing (*C*) is causally related to reproduction (*E*), too. Thus, the causal hypothesis is that birds that sing breed but those that do not sing do not reproduce. In addition to the verbal descriptions of the assumed causal relation, participants were shown a graphical representation (Figure 18a). It was also pointed out that the factors determining whether a bird sings are not known yet.

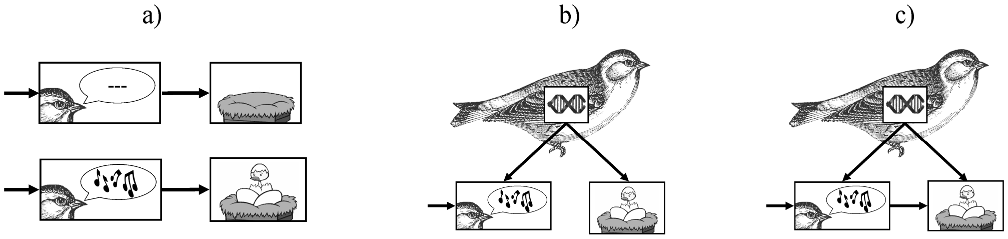


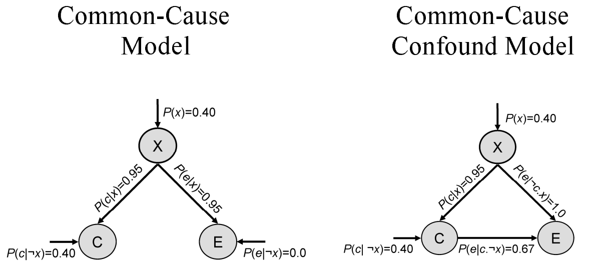
Figure 18. Instructed causal relations in Experiment 7. a) The hypothesized causal relation. b) Common-cause model. c) Common-cause confound model.

After introducing the hypothesized causal relation *C*→*E* participants were presented with the confounding variable *X*, a gene which is known to influence both the birds' capacity to sing and their fertility. Learners were then suggested two candidate causal models. The common-cause model represents the hypothesis that birdsong (*C*) and breeding (*E*) are only correlated because of their common-cause, the gene, and that there is no direct causal relation between the two variables. By contrast, the common-cause confound model represents the hypothesis that birdsong (*C*) and breeding (*E*) are not only spuriously related because of their common cause (*X*), but that there is an additional direct causal relation between singing and breeding. Learners were shown a graphical representation of the two causal models (Figure 18b, c) and requested to find out which of the two models was correct. The kind of questions they would have to answer after the learning phase was not mentioned until the test phase. To avoid misunderstandings, participants were asked to summarize the instructions and the two models in question after reading the instructions.

Learning Phase. To assess whether there is a direct causal relation between birdsong and breeding, learners received 50 index cards with each card referring to a different bird. Each card displayed the state of the three variables (i.e., whether the bird sings, whether it breeds, and whether is possesses the gene). The two data sets either implemented a common-cause model without a direct causal relation between *C* and *E* or a common-cause confound model. Table 15 shows the two parameterized models along with the observational data generated from these graphs. In both data sets, there are 12 birds in which the gene is absent but the bird sings. Normatively, these are the relevant cases which indicate whether there is a causal relation between birdsong (*C*) and breeding (*E*). In the confounder condition, 8 out of these 12 birds breed, which indicates a direct causal relation between *C* and *E*. In contrast, in the common cause condition none of these 12 birds breed. There was no time limit for inspecting the data. Participants were simply asked to signalize when they felt confident that they had determined whether there was a direct causal relation between birdsong and breeding. No feedback was provided.

Table 15

Causal Models and Learning Data of Experiment 7.



Data Pattern			Frequencies	
X	*C*	*E*	Common-cause	Confound
yes	yes	yes	18	18
yes	yes	no	1	1
yes	no	yes	1	1
yes	no	no	0	0
no	yes	yes	0	8
no	yes	no	12	4
no	no	yes	0	0
no	no	no	18	18

Note. X = gene, *C* = birdsong, *E* = reproduction.

Test phase. Finally, learners were asked three blocks of two questions each. The blocks consisted of observational, interventional, and counterfactual questions with each question referring to a new case (cf. Figure 19). The order of blocks was counterbalanced. Participants were allowed to refer back to the index cards and instructions while answering the questions.

The *observational questions* stated that the ornithologists had captured a new bird and observed that this bird sings [does not sing] (cf. Figure 19a). Based on this observation, learners were asked to estimate the probability that this bird would breed (i.e., participants gave estimates of the conditional probabilities $P(e \mid c)$ and $P(e \mid \neg c)$). The *generative interventional questions* stated that the biologists had attached a miniature speaker to a bird which imitates birdsong (i.e., Do c) (cf. Figure 19b). The *inhibitory interventional questions* stated that researchers had surgically modified a bird's vocal cords, thereby preventing this bird from singing (i.e., Do $\neg c$). Participants were requested to estimate the probability that these birds would breed (i.e., learners gave estimates of the interventional probabilities $P(e \mid \text{Do } c)$ and $P(e \mid \text{Do } \neg c)$).

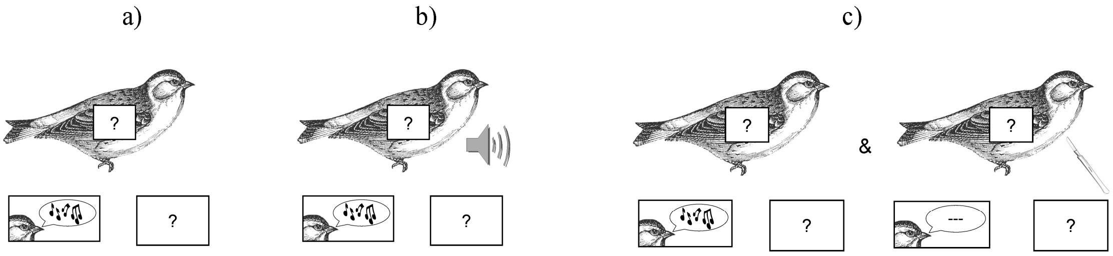


Figure 19. Examples of test questions in Experiment 7. a) Observation of singing bird. b) Generative Intervention (Do *c*). c) Counterfactual prevention combining observation and intervention

The *counterfactual inhibitory question* first stated that the researchers had trapped a new bird which had been observed singing. Participants were then asked to imagine that this very bird's vocal cords had been modified by surgery, and requested to estimate the likelihood that it would have bred (i.e., give estimates of the counterfactual probability $P(e \mid c. \text{ Do } \neg c)$). The *generative counterfactual question* first stated a non-singing bird had been trapped. Learners were then asked to imagine that a speaker imitating birdsong had been attached to this bird and were requested to estimate the probability this bird would have bred (i.e., give estimates of the counterfactual probability $P(e \mid \neg c. \text{ Do } c)$).

The ratings for the observational and interventional questions were given on a scale ranging from "0 = this bird will definitely not breed" to "100 = this bird will definitely breed". For the counterfactual questions, the scale was labeled "0 = this bird would definitely not have bred" to "100 = this bird would definitely have bred".

The test phase ended with a *model selection task*. Learners were given a graphical representation of the two alternative causal models (cf. Figure 18) and requested to select the correct one.

Results and Discussion

Probability judgments. Table 16 shows learners' probability estimates for observations, interventions, and counterfactual interventions along with the normative probabilities derived from causal Bayes nets.

Table 16

Mean Probability Judgments in Experiment 7 (N = 36).

		Observation		Intervention		Counterfactual Intervention	
Causal Model		$P(e \mid c)$	$P(e \mid \neg c)$	$P(e \mid \text{Do } c)$	$P(e \mid \text{Do } \neg c)$	$P(e \mid \neg c. \text{ Do } c)$	$P(e \mid c. \text{ Do } \neg c)$
Common-cause	*Bayes Nets*	*58*	*05*	*38*	*38*	*05*	*58*
	M	63.89	22.78	50.00	41.39	32.50	43.89
	SD	(16.14)	(27.56)	(27.01)	(22.48)	(23.03)	(20.04)
Common-cause confound	*Bayes Nets*	*84*	*05*	*78*	*40*	*68*	*61*
	M	58.33	14.44	63.06	20.56	54.72	25.28
	SD	(22.49)	(20.64)	(20.08)	(22.81)	(24.04)	(28.10)

Note. Normative values (range 0 – 100) derived from causal Bayes nets are shown in italics.

Observations versus interventions. The analysis of the responses to the observational questions shows that learners were clearly sensitive to the fact that both models imply that observed values of *C* are diagnostic for *E*. An analysis of variance with 'presence vs. absence of *C*' as within-subjects factor and 'learning data' as between-subjects factor yielded only a significant main effect for the presence of *C*, $F(1, 34) = 63.73$, $p < .001$, $MSE = 510.38$, $\eta^2 = .62$, but no interaction effect ($F < 1$) and no main effect of condition, $F(1, 34) = 1.86$, $p = .18$. This result indicates that in both conditions participants correctly referred to the unconditional probabilities to infer the state of *E* from observations of *C*.

While both models imply that observed values of *C* provide evidence for the state of *E*, interventions in *C* should only exert an influence on *E* if there is a direct causal link between *C* and *E* (i.e., between birdsong and breeding). In accordance with the causal Bayes nets analysis, the interventional predictions differed depending on the causal model from which the learning data was generated. In the common-cause condition, only a small, non-significant difference was obtained for learners' estimates of the interventional probabilities, $P(e \mid \text{Do } c)$ and $P(e \mid \text{Do } \neg c)$, $F(1, 17) = 1.09$, $p = .31$. In contrast, in the confounder condition the probability of E being present was judged higher when C was generated, $P(e \mid \text{Do } c)$, than when C was prevented, $P(e \mid \text{Do } \neg c)$, $F(1, 17) = 71.67$, $p < .001$, $MSE = 226.84$, $\eta^2 = .81$. Further evidence for the influence of the underlying causal model comes from the between condition comparisons. Given that C was generated by an intervention (i.e., Do c), event E was judged to be more likely in the confounder than in the common-cause condition, though the difference failed to reach significance, $F(1, 34) = 2.71$, $p = .11$. However, consistent with the normative analysis, the interventional probability $P(e \mid \text{Do } \neg c)$ received lower ratings in the confounder condition than in the common-cause condition, $F(1, 34) = 7.61$, $p < .01$, $MSE = 512.79$, $\eta^2 = .18$. These findings indicate that the participants successfully identified the causal structure from which the learning data was generated and based their interventional predictions on the inferred causal model.

Learners' sensitivity to the differences between seeing and doing is corroborated by contrasting the responses to the observational and interventional questions within conditions. Normatively, in the common-cause condition both observational probabilities should differ from their interventional counterparts. In contrast, in the common-cause confound condition, a substantial difference is predicted only between observing C to be absent and actively preventing C. The data conforms to these predictions. In the common-cause condition, both interventional questions were answered differently than the corresponding observational questions. Participants judged event E to be more likely when C was merely observed to be present than when C was generated by an intervention (i.e., $P(e \mid c)$ against $P(e \mid \text{Do } c)$), $F(1, 17) = 8.91$, $p < .01$, $MSE = 194.94$, $\eta^2 = .34$. Conversely, the observational probability, $P(e \mid \neg c)$, received lower ratings than the corresponding interventional probability $P(e \mid \text{Do } \neg c)$, $F(1, 17) = 5.44$, $p < .05$, $MSE = 573.24$, $\eta^2 = .24$.

Consistent with the normative values, in the confounder condition, event E was judged to be more likely in the case of actively preventing C than in the case of merely observing C to be absent, $F(1, 17) = 4.43$, $p = .05$, $MSE = 75.82$, $\eta^2 = .21$, though the difference was not as large as normatively predicted. The crucial test concerns the generative intervention. In the common-cause condition, learners correctly judged E to be more likely when observing C to be present than when C was generated by an intervention, because in this model intervening in C will not affect E. In contrast, in the confound condition participants recognized that generating *C* will influence *E* because of the direct causal relation. Consistent with this prediction, only a small, non-significant difference was found for estimates of $P(e \mid c)$ and $P(e \mid \text{Do } c)$ ($F < 1$).

Taken together, these findings provide strong evidence that the participants had the capacity to separate the genuine causal relation from the concurrent spurious correlation, and that learners recognized the importance of the confounding variable *X* and the backdoor path when deriving the consequences of hypothetical interventions in the putative cause *C*.

Hypothetical vs. counterfactual interventions. The two models differ not only in regard to the consequences of hypothetical interventions, but also with respect to the outcomes of counterfactual actions. The common-cause model implies that interventions in *C* will not affect *E*. Therefore, the probability of *E* given that *C* is counterfactually altered is determined by the factually observed state of event *C*. By contrast, if there is a direct causal relation $C \rightarrow E$, as is the case in the confound model, the counterfactual inferences require the combination of observation and intervention.

For the common-cause condition, causal Bayes nets theory predicts lower ratings for the counterfactual generation of *C* than for the counterfactual prevention of *C*, a difference which results from the logic of counterfactual inferences. The counterfctaul generation of event *C* implies that *C* was observed to be absent in the actual world, which makes it likely that the confounding variable *X* and, in turn, event *E* is absent. Conversely, the counterfactual prevention of *C* entails that *C* was present in the actual world, which raises the likelihood that *X* and *E* are present. In accordance with this prediction, learners gave lower ratings for $P(e \mid \neg c.\ \text{Do } c)$ than for $P(e \mid c.\ \text{Do } \neg c)$, though the difference failed to reach significance, $F(1, 17) = 2.27$, $p = .15$. Moreover, no difference was found between the hypothetical prevention of *C*, $P(e \mid \text{Do } \neg c)$, and the counterfactual prevention of *C*, $P(e \mid c.\ \text{Do } \neg c)$ ($F < 1$). However, in accordance with the normative analysis, participants judged *E* to be more likely in the case of a hypothetical generation than in the case of a counterfactual generation of *C*, $F(1, 17) = 5.81$, $p < .05$, $MSE = 473.90$, $\eta^2 = .26$. Taken together, there is only weak evidence that learners' counterfactual inferences in the common-cause condition conformed to predictions of causal Bayes nets.

Similar deviations from the normative predictions are found in the confounder condition. Normatively, the probability of *E* is only slightly higher when *C* is counterfactually generated, $P(e \mid \neg c.\ \text{Do } c)$, than when *C* is counterfactually inhibited,

$P(e \mid c.\,\mathrm{Do}\,\neg c)$, but a large difference was obtained, $F(1, 17) = 18.82$, $p < .001$, $MSE = 414.54$, $\eta^2 = .53$. The comparisons with participants' responses to the hypothetical intervention questions also indicate substantial deviations from the normative probabilities. Consistent with the predictions, event *E* received slightly higher ratings in the case of a hypothetical generation of *C* than in the case of a counterfactual generation of *C*, $F(1, 17) = 3.9$, $p = .07$, $MSE = 160.29$, $\eta^2 = .19$. But the crucial test concerns learners' estimates for the hypothetical and counterfactual prevention of *C*, because here a large difference is predicted. However, no reliable difference was obtained ($F < 1$).

Model selections. The results for the model selection task are shown in Table 17. In total, 27 out of 36 participants (75%) managed to identify correctly the causal model from which the learning data was generated. A 2x2-chi-square test on learners' model choices yields a highly reliable result, $\chi^2\,(1, N = 36) = 9.26$, $p < .01$. However, even though in both conditions a majority of participants chose the correct model, further analyses reveal that learners had more problems identifying the confound model than the common-cause model. The proportion of participants who chose the correct model was significantly greater than chance in the common-cause condition, $\chi^2\,(1, N = 18) = 8.00$, $p < .01$, whereas the proportion in the confound condition was not, $\chi^2\,(1, N = 18) = 2.00$, $p = .16$. A possible explanation is that in the common-cause confound condition some learners were led astray by the spurious correlation implied by the confounding variable *X*.

Table 17

Model Selections in Experiment 7.

	Selected Model	
Condition	Common-cause	Common-cause confound
Common-cause	15	3
Common-cause confound	6	12

Comparing model selections with probability judgments. Finally, Table 18 depicts the probability judgments for those learners who selected the correct causal model. The data indicates that some of the judgments conform better to the normative probabilities (e.g., the counterfactual probabilities in the common-cause condition), but the general picture is very similar to the aggregated data (cf. Table 16). However, the estimates which deviated most strongly from the normative values, namely the responses to the counterfactual questions in the confounder condition, are also not in line with the normative predictions for those people who selected the correct model. Thus, even learners who successfully identified the confounder model from the learning data had problems differentiating hypothetical from counterfactual actions.

Table 18

Mean Probability Judgments in Experiment 7 for Participants who Selected the Correct Model (N = 27).

		Observation		Intervention		Counterfactual Intervention	
Causal Model		$P(e \mid c)$	$P(e \mid \neg c)$	$P(e \mid \text{Do } c)$	$P(e \mid \text{Do } \neg c)$	$P(e \mid \neg c. \text{ Do } c)$	$P(e \mid c. \text{ Do } \neg c)$
Common-cause (n = 15)	*Bayes Nets*	*58*	*05*	*38*	*38*	*05*	*58*
	M	62.00	24.67	48.00	45.67	28.33	48.00
	SD	(16.56)	(29.91)	(27.57)	(21.95)	(21.69)	(19.35)
Common-cause confound (n = 12)	*Bayes Nets*	*84*	*05*	*78*	*40*	*68*	*61*
	M	56.67	9.17	59.58	14.17	47.92	24.58
	SD	(24.53)	(9.00)	(20.94)	(13.62)	(23.50)	(29.03)

Note. Normative values (range 0 – 100) derived from causal Bayes nets are shown in italics.

To sum up, the results of Experiment 7 provide clear evidence that learners had the capacity to differentiate the "normal" common-cause model from the common-cause confound model on the basis of the available observational data and, in turn, to distinguish observations from interventions. However, participants had only a limited understanding of understanding the implications of counterfactual inferences, which require us to combine observations and interventions.

5.6.2 Experiment 8

Whereas Experiment 7 focused on learners' understanding of common-cause confounding, the goal of Experiment 8 is to investigate reasoning with causal-chain confounding. As in the previous study, learners' task is to evaluate whether the observational data indicates the presence of a direct causal relation $C \rightarrow E$. The two candidate models they are presented with are a causal-chain model and a causal-chain confound model, as depicted in Figure 20. As in Experiment 7, both graphs imply that events C and E are correlated even if there is no direct causal relation between C and E. Thus, similar to the case of common-cause confounding, learners need to separate the direct causal relation from the spurious correlation to decide which of the two candidate models underlies the observational

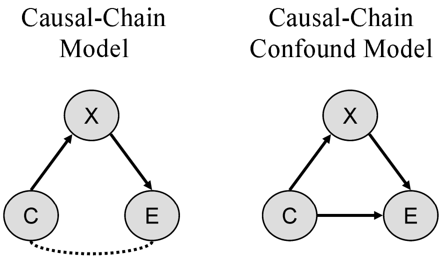


Figure 20. Causal models in Experiment 8.

data. Interestingly, the presence of a direct causal relation $C \rightarrow E$ can be evaluated in the same manner as in the case of common-cause confounding. The crucial observations which are diagnostic for the existence of a direct causal arrow are instances in which the candidate cause C is present but the confound X is absent. However, whereas in Experiment 7 the crucial condition was that X is not the only cause of C (i.e., $P(c \mid \neg x) > 0$), here the necessary condition is that C does not deterministically cause X (i.e., $P(x \mid c) < 1$). Otherwise, there will be no cases in which C is present but the confounding variable X is absent.

Applying the causal Markov condition to the two graphs factorizes the associated probability distributions. The causal-chain model is decomposed into

$$P(X.C.E) = P(C) \cdot P(X \mid C) \cdot P(E \mid X), \tag{29}$$

while the distribution of the causal-chain confound model is factorized into

$$P(X.C.E) = P(C) \cdot P(X \mid C) \cdot P(E \mid X.C). \tag{30}$$

These two factorizations reflect the structural difference between the simple causal-chain and the causal-chain confound models. According to the structure of the causal-chain model, event E is only influenced by X, therefore E is only conditionalized on X. In contrast, in the causal-chain confound model E not only depends on X but is also directly influenced by C. Therefore, event E is conditionalized on both X and C.

Provided the parameters of the causal models can be estimated from the available observational data, it is possible to predict the consequences of interventions. In Experiment 7, the consequences of observations of and interventions in C differed because interventions in C rendered the event independent of the confounding variable X. Normatively, this implies a difference between observations and interventions; and the empirical findings show that learners were sensitive to this difference.

However, if the confounding variable X is not a cause but an effect of C, as is the case in the causal-chain confound model, interventions in C do not disconnect C from X. Since C is the cause of both E and X, the dependence of X and C is not eliminated by an intervention on C. In other words, because event C is the causal models' initial event, interventions in C do not result in a manipulated graph. Therefore, the interventional and observational probabilities are equal in both the simple causal chain and the causal-chain confound model.

For example, in the causal-chain model the probability of $E = e$ given that C is merely observed to be present or actively generated is formalized in the same way

$$P(e \mid c) = P(e \mid \text{Do } c) = P(x \mid c) \cdot P(e \mid x) + P(\neg x \mid c) \cdot P(e \mid \neg x). \tag{31}$$

The same is true for the causal-chain confound model,

$$P(e \mid c) = P(e \mid \text{Do } c) = P(x \mid c) \cdot P(e \mid x.c) + P(\neg x \mid c) \cdot P(e \mid \neg x.c) \tag{32}$$

In other words, neither the causal-chain nor the causal-chain confound model implies a difference between observations of and interventions in *C* (i.e., *P*(*e* | *c*) = *P*(*e* | Do *c*) and *P*(*e* | ¬*c*) = *P*(*e* | Do ¬*c*) holds for both models). Unfortunately, this implies that participants' estimates of the direct causal influence of the cause variable *C* on the effect *E* cannot be estimated on the basis of the two conditional interventional probabilities. In order to assess only the cause's direct causal influence the causal relation between the cause and the confounding variable has to be eliminated by a second intervention. This aim could be achieved by eliminating the confounding variable or by blocking the causal pathway connecting the cause and confound.

For example, consider an intervention that simultaneously manipulates *C* and interrupts the causal mechanism by which *C* generates *X*. For this kind of *combination of interventions*, the consequences for the probability of *E* depend on the structure of the causal system. In the causal-chain model, interventions in *C* will not affect *E* if the intervention simultaneously breaks the causal arrow *C*→*X*, because the influence of *C* on *E* completely depends on the intermediate event *X*. In contrast, in the causal-chain confound model this kind of double intervention will have an impact on *E* because of the direct link *C*→*E*. In this model, the double intervention only disrupts the indirect causal path, but *C* can still influence *E* through the direct causal relation.

In the causal-chain model, the probability of *E* conditional on a combination of interventions that generates *C* and simultaneously breaks the causal path *C*→*X* is formalized by

$$P(e \mid \text{Do } c.\,\textit{break } C \to X) = P(x \mid \neg c) \cdot P(e \mid x) + P(\neg x \mid \neg c) \cdot P(e \mid \neg x). \tag{33}$$

In the causal-chain model, this double intervention renders *C* independent of *E*. Therefore, the probability of *E* occurring is determined by the probability of *X* occurring without *C* and the strength of the causal relation between *X* and *E*. In contrast, in the causal-chain confound model, in which there is also a direct causal relation between *C* and *E*, the probability of *E* conditional on the double intervention is given by

$$P(e \mid \text{Do } c.\,\textit{break } C \to X) = P(x \mid \neg c) \cdot P(e \mid x.\,c) + P(\neg x \mid \neg c) \cdot P(e \mid \neg x.\,c). \tag{34}$$

To sum up, whereas common-cause confounding entails a difference between observations of and (simple) interventions in *C*, this does not hold for causal-chain confounding. Both in the causal-chain and the causal-chain confound model the interventional probabilities include the confounding causal relation and therefore equal the observational probabilities. To test whether participants are able to extract the direct causal relation in this case, in Experiment 8 participants are not only requested to infer the consequences of simple interventions but are also asked about combinations of interventions (i.e., double interventions that simultaneously block the causal relation to the confounding variable). If participants understand the causal logic of confounding, the estimated probabilities should reflect the direct impact of the cause upon its effect.

Method

Participants and Design

Thirty-six students from the University of Göttingen, Germany, participated in this experiment. They were randomly assigned to the causal-chain or the causal-chain confound condition. Factor 'learning data' was varied between conditions, factors 'type of inference' and 'presence vs. absence of *C*' were varied within-subjects. Subjects received course credit for participation.

Procedure and Materials

Causal model phase. The same scenario was used as in Experiment 7. However, now participants were told that ornithologists were investigating whether a specific gene (*C*) has a direct causal impact upon the birds' reproduction (*E*). As before, participants were informed about the presence of a confounding variable. They were told that the gene was known to affect the birds' ability to sing (*X*). Learners were also informed that the gene affects the birds' ability to sing by a (non-observable) hormone mechanism (*H*). Moreover, singing has, according to the instructions, a causal influence upon reproduction. Participants were then presented with two competing causal hypotheses, a causal-chain model and a causal-chain confound model. The hypothesized causal relation between the gene (*C*) and reproduction (*E*) as well as the candidate model were visualized graphically (Figure 21). The unobservable mechanism (*H*) was not depicted in these graphical representations. The causal-chain confound model represents the assumption that the gene has both an immediate and an indirect causal impact upon reproduction, whereas the causal-chain model represents the hypothesis that the gene affects reproduction only via singing. As in Experiment 7, participants were asked to find out which model was correct. They were not informed about the kind of questions they would have to answer.

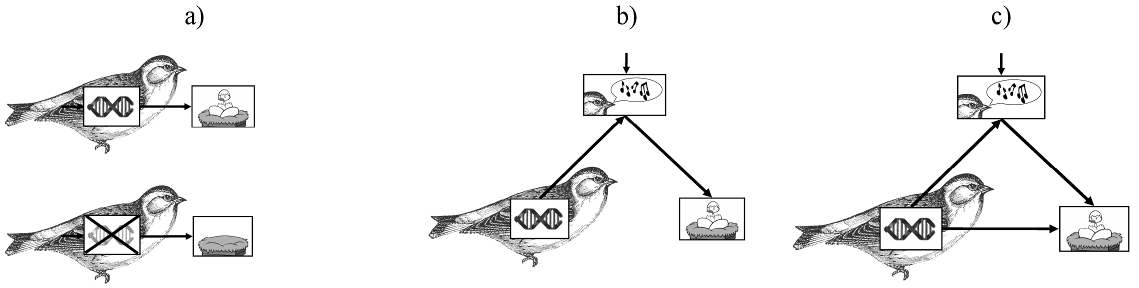


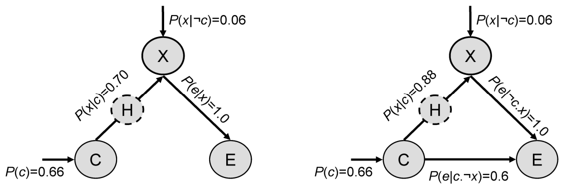
Figure 21. Instructed causal relations in Experiment 8. a) The hypothesized causal relation. b) Causal-chain model. c) Causal-chain confound model.

Learning Phase. As in the first experiment, learners received 50 index cards depicting observational data from individual birds. The models used to generate the two sets of data and the resulting distributions of event patterns are shown in Table 19. Note that participants were never informed about the state of *H*, the mechanism connecting *C* to *X*. The causal-chain data indicated that the observable relation between *C* and *E* was merely spurious, while the data corresponding to the causal-chain confound model pointed to a fairly strong direct relation between the gene and reproduction. The unconditional relation between *C* and *E* was identical in both data sets ($P(e \mid c) = .88$ and $P(e \mid \neg c) = .06$) As before, participants were free to explore the data at will.

Table 19

Causal Models and Learning Data of Experiment 8.

Causal-Chain Model	Causal-Chain Confound Model



Data Pattern			Frequencies	
X	*C*	*E*	Causal-chain	Causal-chain Confound
yes	yes	yes	29	23
yes	yes	no	0	0
yes	no	yes	1	1
yes	no	no	0	0
no	yes	yes	0	6
no	yes	no	4	4
no	no	yes	0	0
no	no	no	16	16

Note. X = gene, *C* = birdsong, *E* = reproduction.

Test Phase. In this phase, participants were given three blocks of questions with the order of blocks being counterbalanced. Examples of the test questions are shown in Figure 22. The *observational questions* asked participants to estimate the probability that a new bird possessing the gene [not possessing the gene] would breed (i.e., learners were requested to give estimates of $P(e \mid c)$ and $P(e \mid \neg c)$). The *generative interventional question* stated that the researchers had activated the gene of a new bird by means of an intervention (i.e., Do c). The *inhibitory interventional question* mentioned that the gene was deactivated by an outside intervention (i.e., Do $\neg c$). Participants had to estimate the probability that these new birds would breed (i.e., participants were asked to give estimates of $P(e \mid \text{Do } c)$ and $P(e \mid\text{Do } \neg c)$). The first question referring to a *combination of interventions* informed participants that researchers had activated the gene of a newly caught bird while simultaneously blocking the generation of the hormone affecting singing. The second combination question stated that both the gene and the hormone production had been deactivated by inhibitory interventions. For both questions, participants were asked about the probability of procreation (i.e., $P(e \mid \text{Do } c.\ \text{Do } \neg h)$ and $P(e \mid \text{Do } \neg c.\ \text{Do } \neg h)$). In both cases, participants received no information about whether the individual birds had the capacity to sing or not. As in Experiment 7, the test phase ended with a *model selection task* in which participants had to select the correct model from a graphical representation of the two alternative causal models (cf. Figure 21).

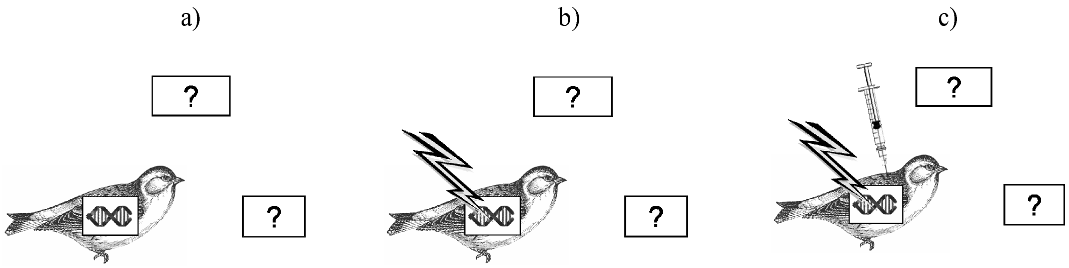


Figure 22. Examples of test questions in Experiment 8. a) Bird observed having the gene. b) Generative intervention activating the gene (i.e., Do c). c) Combination of interventions activating the gene and inhibiting the hormone mechanism (i.e., Do c & Do $\neg h$).

Results and Discussion

Table 20 shows the mean probability estimates for the six questions along with the normative values derived from causal Bayes nets. Again, participants gave on average the same ratings to the observational questions in both conditions and judged the effect to be more likely in the presence than in the absence of the observed cause. Consistent with the normative predictions, an analysis of variance with 'presence versus absence of C' as within-subjects factor and 'learning data' as between-subjects factor yielded only a main effect for the presence of C, $F(1, 34) = 317.25$, $p < .001$, $MSE = 231.19$, $\eta^2 = .90$, but neither a main effect of condition, $F(1, 34) = 1.78$, $p = .19$, nor an interaction between conditions, $F(1,34) = 1.05$, $p = .31$.

Table 20

Mean Probability Judgments in Experiment 8 (N = 36).

		Observation		Intervention		Combination of Interventions	
Causal Model		$P(e \mid c)$	$P(e \mid \neg c)$	$P(e \mid \text{Do } c)$	$P(e \mid \text{Do } \neg c)$	$P(e \mid \text{Do } c. \text{ Do } \neg h)$	$P(e \mid \text{Do } \neg c. \text{ Do } \neg h)$
	Bayes nets	*88*	*06*	*88*	*06*	*06*	*06*
Causal-chain	*M*	76.89	16.72	77.08	17.28	29.03	7.83
	SD	(18.98)	(12.67)	(16.63)	(13.48)	(31.45)	(9.87)
	Bayes nets	*88*	*06*	*88*	*06*	*62*	*06*
Causal-chain confound	*M*	76.11	8.61	81.11	14.72	62.22	15.28
	SD	(18.20)	(3.35)	(13.67)	(21.59)	(26.25)	(23.92)

Note. Normative values (range 0 – 100) derived from causal Bayes nets are shown in italics.

In contrast to Experiment 7 and in line with the Bayesian causal analysis, participants' estimates for the simple interventional questions did not differ between conditions. An analysis of variance resulted in a significant main effect for the presence of *C*, $F(1, 34) = 250.20$, $p < .001$, $MSE = 286.42$, $\eta^2 = .88$, but neither an effect of condition nor an interaction effect was found (both *F*'s < 1). Participants seemed to have understood that intervening in *C* would generate *E* no matter whether the underlying causal model was a causal-chain or a causal-chain confound model, and that there is no difference between merely observing the state of *C* and actively generating the value of *C*.

However, participants' answers to the combination of interventions questions showed that they differentiated between the two models. Consistent with the normative probabilities, an analysis of variance yielded a main effect of presence of *C*, $F(1, 34) = 54.62$, $p < .001$, $MSE = 382.55$, $\eta^2 = .62$, a main effect of 'learning data', $F(1, 34) = 9.39$, $p < .01$, $\eta^2 = .22$, and, most important, the expected interaction, $F(1, 34) = 7.80$, $p < .01$, $\eta^2 = .19$. This result indicates that learners inferred the consequences of the double intervention with respect to the model from which the learning data was generated. However, even though the difference between the double intervention questions was much smaller in the causal-chain than in the causal-chain confound condition, a difference was also obtained between the double intervention questions in the causal-chain condition, $F(1, 17) = 8.87$, $p<.01$, $MSE = 457.55$, $\eta^2 = .34$. A closer look at individual ratings revealed that 10 out of the 18 participants in the causal-chain condition judged *E* to be equally likely when *C* was generated by an intervention and when it was prevented while the causal mechanism linking *C* to *X* was blocked. In contrast, all participants in the causal-chain confound condition assumed that an intervention in *C* would increase the probability of *E* despite the blocked link. Thus, a majority of participants seemed to have grasped the causal logic of causal-chain confounding.

Model selections. The results for the model selection task are shown in Table 21. In total, 31 of the 36 participants (86%) picked the correct causal model. Thus, like in Experiment 7, a majority was able to separate the causal relation between *C* and *E* from the spurious relation. A 2x2-chi-square test on learners' model choices yielded a highly reliable result, χ^2 (1, N = 36) = 19.31, p < .001. The proportion of participants who chose the correct model was significantly greater than chance in the causal-chain condition, χ^2 (1, N = 18) = 5.56, p < .05, as well as in the causal-chain confound condition, χ^2 (1, N = 18) = 14.22, p < .001.

Table 21

Model Selections in Experiment 8.

	Selected Model	
Condition	Causal-Chain	Causal-Chain confound
Causal-Chain	14	4
Causal-Chain confound	1	17

5.6.3 Experiments 7 and 8: Summary and Discussion

The goal of Experiments 7 and 8 was to further investigate learners' causal inferences with confounding variables. Two basic causal structures containing confounds were examined: a common-cause confound model, in which a cause and an effect are directly and spuriously related, and a causal-chain confound model, in which a cause both directly and indirectly influences its effect. Manipulating the cause by an external intervention eliminates common-cause confounding but not causal-chain confounding, which requires blocking the second causal pathway. The results of the two experiments show that participants understand the causal logic of these two types of confounding. A majority of participants in both experiments were able to disentangle the direct causal relation from the additional spurious relation.

How did people achieve this? Previous research on causal judgments has shown that participants tend to control for extraneous causal variables when estimating the causal impact of a target variable (e.g., Spellman, 1996; Waldmann & Hagmayer, 2001). Confounding variables are such extraneous variables. Controlling for these variables, for example by only considering cases in which they are absent, enables us to derive causal inferences about the consequences of interventions. Observations of participants' behavior during the experiment and their written comments indicate that participants used the strategy of focusing on the events in which the confound was absent when assessing whether a direct causal relation was present or not. The findings also demonstrate that people understand that observational inferences require the use of unconditional contingencies, whereas interventional predictions require controlling for the confounding variables.

How do the experiments relate to findings showing that people occasionally fail to understand confounding? One critical factor might be the data that is shown to participants. In some studies (e.g., Luhmann, 2005), participants did not receive data that allowed them to focus on the absence of the confounding variable (i.e., holding it constant). As a consequence, participants in his study tended to overestimate how informative confounded data is, even though people seem to be sensitive to situations in which the candidate causes are perfectly confounded (Ford & Cheng, 2004). However, even when participants receive the necessary information, they may not succeed if the critical cases are rare (see Waldmann & Hagmayer, 2001). Common-cause confounding may serve as an example: If the common-cause confound has a high base rate and strongly affects the target cause only very few cases will occur in which the target cause will be present in the absence of the confounding variable. In the two experiments reported here, participants were presented with data that contained a relatively large number (> 10) of such critical cases. In a pilot study (not reported here), learners were presented with fewer of these critical observations, and participants consequently failed to arrive at correct conclusions.

In summary, people are able to understand the implications of common-cause confounding and causal-chain confounding, and they proved capable of deriving their interventional predictions in accordance with the inferred causal model. Thus, the results of Experiments 7 and 8 further corroborate the claim that learners have the capacity to reason with causal models containing confounds. This remarkable capacity may fail, however, with more complex models or less salient data.

6 General Discussion

The intention of this final section is to summarize and discuss the empirical findings. Not surprisingly, a central emphasis is placed on how the results of the experiments relate to the predictions of causal Bayes nets theory and competing theories of causal inference. The section will end with a review of the causal Bayes nets formalism as a psychological model and an outline of some future research questions.

6.1 Summary: Causal Bayes Nets as Models of Causal Reasoning

The thesis started with the question of how we can infer the consequences of our actions when no direct knowledge about the potential outcomes of these actions is available. The capacity to derive interventional predictions from observational knowledge is a touchstone of true causal reasoning, because it goes beyond the mere ability to detect and represent structure-free contingencies. Causal Bayes nets theory was introduced as a normative account of causal representation, causal learning, and causal inference. The theory formalizes interventions in causal systems and provides the computational mechanisms for inferring interventional predictions from causal models parameterized by passive observations.

The results of the experiments presented here provide strong evidence that people have the competency to engage in this kind of causal reasoning. Consistent with causal Bayes nets theory, learners distinguished between merely observed states of variables ("seeing") and the very same states generated by external interventions ("doing"). Participants correctly recognized that interventions in a variable render the event independent of its actual causes, but also understood that the potential differences between observations and interventions crucially depend on the structure and the parameters of the investigated causal system. Thus, participants' inferences about the consequences of interventions differed systematically in accordance with the instructed causal graphs and the models' parameters.

These findings challenge alternative accounts of causal inference, which provide no means to represent whether a variable's state was merely observed or actively generated. The presented experiments also refute the idea that the capacity to distinguish between seeing and doing might be restricted to descriptions of causal situations or causal reasoning with aggregated data. Learners successfully derived the consequences of hypothetical interventions after passively observing the behavior of a causal system in a trial-by-trial learning procedure. This finding is inconsistent with the predictions of associative learning theories, which fail to derive interventional predictions from a purely observational learning phase. The results are also incompatible with the assumption that causal learning is driven by associative learning mechanisms and that causal judgments are a function of associative strength. The competency to distinguish between the diverging implications of observations and interventions was demonstrated for simple diagnostic judgments as well as for predictive judgments that require taking into account multiple variables and causal relations.

The experiments also show a surprising grasp of the implications of confounding variables when reasoning with complex causal models. Learners not only proved capable of disentangling the influence of a direct causal relation from a concurrent spurious correlation, but they also took into account confounding pathways when deriving the consequences of interventions. Moreover, they were sensitive to the different causal structures that might underlie confounding, as well as to the fact that these types of confounding differ with respect to the consequences of interventions.

Finally, the studies also point out some of the boundary conditions for the causal Bayes nets formalism as a psychological model. For example, learners had problems distinguishing hypothetical interventions from counterfactual interventions. Even though the findings indicate that learners do not treat hypothetical interventions and counterfactual actions in an identical manner, their responses did show substantial deviations from the normative predictions. The problem is likely rooted in the complexity of counterfactual interventions, because this type of inference combines observations with interventions that require an updating of the model's probabilities prior to the stage of model manipulation.

In summary, the results provide clear evidence that learners are sensitive to the differences between seeing and doing, and that they have the capacity to infer the consequences of hypothetical interventions from observational knowledge. The experiments strongly support causal Bayes nets as a theoretical account of causal reasoning. Alternative theories of causal cognition lack the representational power to express the crucial differences between observations and interventions and therefore fail to account for the data. Causal Bayes nets theory is currently the only model that provides a formal account of interventions and allows for deriving interventional predictions from causal models parameterized by observational knowledge.

6.2 Seeing versus Doing in Trial-by-Trial Learning

An important feature of the studies presented was the use of trial-based learning. Previous studies investigating people's sensitivity to seeing and doing either focused on qualitative reasoning (Sloman & Lagnado, 2005) or provided participants with aggregated lists of data which were available during causal reasoning (Waldmann & Hagmayer, 2005). However, some authors have argued (e.g., Price & Yates, 1995; Shanks, 1991) that learning on the basis of aggregated data is handled by different processes than trial-by-trial learning. According to this position, associative learning mechanism are only activated in trial-based learning; therefore, experiments that do not use trial-by-trial learning do not necessarily provide evidence against associative accounts of causal cognition. Thus, trial-based learning not only provides a more naturalistic learning environment but is also an important condition for the comparison of learners' causal judgments with the predictions of associative models of causal cognition.

The fundamental problem associative theories face is to give an account of how associations acquired from observational learning relate to causal judgments about the outcomes of possible interventions. As discussed in detail in Section 5.2, there are three lines of argument, but none of them can explain learners' capacity to derive interventional predictions subsequent to an observational learning phase. One position is to separate learning from observations from interventional learning completely. Unfortunately, this would imply that interventional predictions are not possible without prior instrumental learning. Another position is to assume that interventional predictions are a direct function of the observationally acquired associations. This approach, however, fails when observed states of variables have different implications than the same states generated by external intervention. Finally, one could assume an interaction between classical and instrumental conditioning; but since learners never undergo an instrumental learning phase (i.e., experience the outcomes of interventions), this argument does not apply to the experiments presented here.

Price and Yates (1995) further specify the conditions under which causal judgments are assumed to be a function of associative strength. They advocate a more detailed model of contingency learning and causal judgment, one which comprises both an associative learning mechanism and a rule-based component. In accordance with the argument put forward by Shanks (1991), Price and Yates assume that in causal induction associative mechanisms are only activated when the data is presented in single trials; otherwise, the data is processed by the rule-based component.

According to Price and Yates, associations are unidirectional from cues to outcomes; and causal judgments are only a function of associative strength when the direction of the inference matches the direction of the acquired association.

> If a participant is asked to make a judgment that is directionally consistent with an existing cue-outcome association (e.g., an estimate of the conditional probability of one of the outcomes given one of the cues), that association serves as the basis of the judgment. However, if the participant is asked to make a judgment that is not directionally consistent with any existing association (e.g., an estimate of the conditional probability of one of the cues given one of the outcomes), the judgment is based on some other process. (Price & Yates, 1995, p. 1651)

This claim, however, bears several problems. Firstly, it severely restricts the explanatory power of an associative account of causal inference, because many (everyday and experimental) situations would fall outside the boundaries of the theory. Secondly, the assumption also raises questions about the necessity of the associative component in general: if the rule-based component can handle situations in which learning order and the direction of the causal inference mismatch, why should these processes not also operate when the judgment is directionally consistent with learning order? In addition, because it is not known during learning which kind of judgment will later be demanded, the associative and rule-based processes would have to run in parallel. Thirdly, the empirical findings are incompatible with the claim that causal judgments are a function of associative strength when the inferences are directionally consistent with a previously acquired cue-outcome association. Experiments 5 and 6, which only differ in the temporal order during observational learning, best illustrate this. In Experiment 5, learning order matched causal order, and thus cause events are mapped onto cues and effect events correspond to outcomes. In contrast, in Experiment 6 learning order was reversed; therefore, in this experiment effects correspond to cues whereas causes are mapped onto outcomes. Thus, in Experiment 5 the acquired associations lead from causes to effects, whereas in Experiment 6 the associations are directed from effects to causes. Now, because learners were requested to give diagnostic judgments (from effect to cause) as well as predictive judgments (from cause to effect) in each experiment, there are judgments that are directionally consistent as well as judgments that are directionally inconsistent with the cue-outcome association (see Figure 23). According to the model of Price and Yates, in

Experiment 5 the predictive judgments from *C* to *D* should be a function of associative strength, whereas in Experiment 6 the diagnostic judgments from *C* to *A* should be a function of associative strength, because these are the inferences directionally consistent with the acquired cue-outcome association.

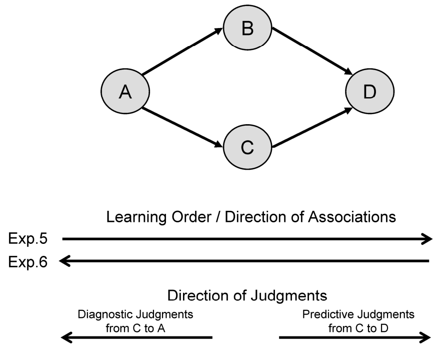

B
A
D
C
Learning Order / Direction of Associations
Exp.5
Exp.6
Direction of Judgments
Diagnostic Judgments from C to A
Predictive Judgments from C to D

Figure 23. Directionality of associations and causal judgments in Experiments 5 and 6.

The empirical findings and the obtained differences between observations and interventions refute this assumption. In Experiment 5, learners' predictive judgments from *C* to *D* differed depending on whether *C* was merely observed or actively generated. Thus, causal judgments about the consequences of the intervention were not derived from the association between the cue (event *C*) and the outcome (event *D*). However, Price and Yates point out that the paradigmatic cases in which causal judgments are derived from associative strength are those that concern the prediction of the outcome given that the cue is present, but the obtained difference between seeing and doing primarily concerned the probability of *D* given that *C* was *absent*. Apart from the problem that the assumption of directional consistency constrains the applicability of the model even further, this conjecture is refuted by the results of Experiment 6. Due to the reversal of the learning order, in this experiment the diagnostic judgments from *C* to *A* are directionally consistent with the acquired cue-outcome associations and should therefore be a function of associative strength. Inconsistent with this prediction, the probability of *A* was judged differently depending on whether *C* was merely observed to be present or actively generated by an intervention. Thus, even with the additional constraint that causal judgments are only determined by associative strength when the causal inference is directionally consistent with the acquired cue-outcome association, associative models fail to account for the empirical results.

Finally, the model of Price and Yates also illustrates that a combination of associative and rule-based learning mechanisms cannot explain the data. According to the model, causal judgments that are not directionally consistent are handled by rule-based mechanisms (e.g., contingency models) that derive the judgments from frequency information. However, because these models, too, are insensitive to the difference between observation and intervention, a combined model cannot account for the findings.

Taken together, the results of the experiments convincingly demonstrate that people can infer the outcomes of potential actions without prior instrumental learning. These findings contradict traditional associative learning theories, which fail to account for causal-model learning, and which are incapable of deriving correct predictions for actions after purely observational learning. On a more general level, the findings raise doubts about the common distinction between representations acquired from observational learning (classical conditioning) and those acquired through interventional learning (instrumental conditioning). The separation of observational and interventional learning is further challenged by a recent study of Blaisdell and colleagues (2006) showing that rats can infer the outcomes of instrumental actions after a passive classical conditioning phase.

6.3 Causal Reasoning with Confounds

A further goal of the experiments was to investigate learners' understanding of and reasoning with confounds. Confounding variables are an important issue in both causal learning and causal reasoning. In structure induction, learning through interventions (i.e., experiments) is considered the prime method for avoiding confounds, although manipulations of the candidate cause can only eliminate common-cause but not causal-chain confounding (cf. Section 5.6). Anyway, other control techniques are required when only observational data is available. Because here confoundings cannot be avoided through external manipulations of the candidate cause, it is necessary to use other strategies to differentiate between candidate causal models. One important strategy is to focus on situations in which potentially confounding variables are absent. This strategy requires selecting the crucial cases from the available data and analyzing them in an adequate manner (e.g., estimating conditional contingencies). In line with previous research (e.g., Cheng & Novick, 1992; Spellman, 1996; Waldmann & Hagmayer, 2001), the experiments presented here indicate that people understand the importance of controlling for confounds when assessing causal relations. For example, in Experiments 7 and 8 the presence of the hypothesized causal relation could be evaluated by holding constant the confounding variable; moreover, the results showed that people succeeded in doing this.

Sensitivity to confounding variables is not only crucial in causal learning but also in causal reasoning. In particular, confoundings are an important issue when reasoning about possible interventions in complex causal systems, because in such systems the outcomes of our actions might be not determined solely by the variable intervened in. However, mere knowledge of the potentially confounding variables is not sufficient; the actual influence of the confounding variable with reference to the kind of causal inference has to be taken into account. The experiments demonstrate that people are sensitive to confounds and have the capability to disentangle a genuine causal relation from a concurrent spurious correlation. Moreover, in accordance with the predictions of causal Bayes nets theory, learners recognized that observations and interventions differ with respect to the ways confounds have to be taken into account. Participants proved not only sensitive to backdoor paths in general, but also took into account the likelihood of the backdoor path's instantiation and the strength of the causal connections constituting the alternative pathway. Finally, they also differentiated between the implications of different types of confounding (i.e., common-cause confounding and causal-chain confounding), and they successfully inferred the consequences of observations, interventions, and combinations of interventions with reference to the induced causal model.

These findings support the view that everyday causal reasoning conforms to similar principles as those underlying scientific studies. Further research has to investigate in more detail learners' capacity to take into account confounding variables. For example, an important factor seems to be the number of critical cases which allow for distinguishing between candidate models. Another factor is that in the presented experiments learners already had knowledge of the existence of the confounding variable and could capitalize on this knowledge to distinguish the models in question. However, this competence might be marred if the complete model has to be induced from the observational data.

6.4 Are Bayes Nets the "Grand Unifying Theory" of Causality?

At the beginning of this thesis, I introduced causal Bayes nets theory as "a normative formal account of causal representation, causal learning, and causal reasoning" (p. 2). This description suggests that the theory provides a comprehensive framework for analyzing and investigating the three key issues of causal cognition.

Both descriptive and normative formal models have a long history in psychology. For example, in research on hypothesis testing and the famous Wason selection task (Wason, 1966), formal logic provided the yardstick against which human strategies of hypothesis testing were evaluated. Similarly, in research on judgment and decision making, the normative standards come from game theory and expected utility theory. However, progress in the development of psychological theories often does not result from asserting that human behavior conforms to certain standards that are thought of as normative, but rather by examining the conditions under which human behavior deviates from these predictions. For example, the strategies of testing descriptive rules in the Wason selection task strongly differed from what formal logic suggested. The deviations from what was considered as the normative standards then led to the development of more refined approaches that could account for the empirical findings, but still maintained a normative core (cf. von Sydow, 2006). Likewise, the development of prospect theory in judgment and decision making (Kahneman & Tversky, 1979) was driven by systematic deviations of people's choices from the predictions of expected utility theory.

In a similar vein, causal Bayes theory can be considered as serving two functions in research on causal cognition. First, the formalism constitutes a consistent theoretical framework for research on human causal cognition (cf. Danks, in press, for a similar view). Some of the theory's aspects are also found in other approaches, such as the emphasis of causal structure in causal model theory (e.g., Waldmann, 1996; Waldmann & Holyoak, 1992) or the use of conditional probabilities in probabilistic theories of causality (Cartwright, 1983; Eells, 1991; Suppes, 1970). As pointed out by Glymour (2003), there is also a close relation between the Bayes nets formalism and Cheng's (1997) power PC theory. However, whereas other accounts address only certain aspects, causal Bayes nets theory is the only model that integrates issues of causal representation, causal learning, and causal reasoning in a coherent formal framework. For example, the account specifies the relation between unobservable causal structures and observable patterns of data. In addition, the theory introduces a number of novel aspects not previously addressed in detail. The basic distinction between observations and interventions is not a new one, but only causal Bayes nets theory formalizes the notion of intervention and relates structural modifications of causal model representations to changes in the associated probability distribution. These are some of the normative aspects of the theory that provide the standards for evaluating different kinds of human causal judgment.

Some authors have questioned the adequacy of the formalism as a means of causal analysis. For example, Nancy Cartwright (2001, 2002) has raised doubts about the theoretical assumptions of the approach, such as whether the causal Markov condition holds in real-life causal systems. Although these issues clearly deserve a more thorough analysis, they concern first and foremost the theory's normative status. More relevant to the psychological debate is whether the formalism provides an adequate description of how people reach their causal beliefs and how these beliefs are used in causal reasoning. Thus, from the perspective of psychology the critical question primarily concerns the model's descriptive validity.

Some of the theory's most important predictions have been tested in the experiments of this thesis, such as whether learners' inferences about the consequences of hypothetical interventions conform to the predictions of the Bayes nets formalism, or whether reasoners take into account confounding variables in a normative fashion. The results support the claim that the formalism is not only a normative model of causality but also captures important aspects of human causal reasoning. In addition, the formalism specifies the aspects which any theory of causal inference has to address in order to account for the empirical findings. A major point is the need to express the differences between merely observed features of the world and states generated by active manipulations. The notion of intervention is intrinsically linked with the need to represent causal structure, because observations and interventions do not differ with respect to some internal feature, but they do indeed differ with respect to their structural implications. In particular, interventions are linked to structural modifications of causal model representations. Finally, computational mechanisms are demanded that specify how knowledge about causal structure is combined with quantitative knowledge, for example to derive the outcomes of potential actions from observational knowledge. Conventional theories of causal cognition, such as contingency models or associative accounts, currently fail to meet these requirements.

Nevertheless, an important question is whether causal Bayes nets theory in its current state already represents a genuine psychological model. First, some of the empirical results call for further research. One critical finding concerns learners' reasoning about counterfactual actions. The experimental data indicate that learners sometimes had problems to correctly assess how hypothetical interventions differ from counterfactual interventions. The crucial difference between the two types of interventional inferences is that predictions about counterfactual interventions demand a specific integration of two pieces of information. The factual observation requires an update of the causal graph's probabilities, and this must be followed by a stage of model manipulation in accordance with the counterfactual action. As learners proved capable of performing these two steps individually in accordance with the normative predictions, the failure to cope with the counterfactual intervention questions is likely to be rooted in the necessary combination of observations and interventions. This hypothesis also receives support from the results of Experiment 1, which focused on reasoning with a single causal relation. In this study answering the counterfactual questions did not require an integration of observation and intervention, and consequently learners' responses conformed to the normative predictions. Future research has to further investigate whether reasoning about counterfactual interventions does not obey the predictions of causal Bayes nets in general, or whether the deviations are rather due to the specific experimental setting (e.g., probabilistic relations, confounds) or limited information processing capacities. For example, the studies of Sloman and Lagnado (2004) indicate that people perform better when reasoning with descriptions of single causal episodes or deterministic relations.

A further issue that should be addressed in future research concerns people's sensitivity to the causal Markov condition. This assumption is a defining principle of causal Bayes nets theory, but there is some evidence that reasoners' inferences do not always conform to this condition (Rehder & Burnett, 2005). However, other studies (Lagnado & Sloman, 2004) have found that a majority of participants were sensitive to the conditional independence relations entailed by a causal model. More data is necessary to specify under which conditions learners' inferences obey the Markov condition and which factors lead to violations of the principle.

Finally, there are also some theoretical aspects of causal Bayes nets that should be considered in more detail, such as the theory's level of description. In its current version, the approach is a computational level description of how a causal model's parameters have to be combined in causal reasoning to infer the states of a system's variables conditional on observations, interventions, and counterfactual interventions. Thus, the theory describes which information a cognizer needs in order to perform certain computations, but how these parameters are acquired in the first place is only partially explained. If learners are provided with aggregated lists of data, people can estimate the model's parameters directly from the available frequency information. However, things get complicated in the case of trial-by-trial learning. Here learning mechanisms are required to explain how a model's parameters are acquired during the course of learning.

One position is to assume that the frequency information is somehow stored in memory and later provides the basis for estimating the graph's parameters. This account has the advantage that the data is available in a "raw format" and new information is easily added to the existing knowledge base. Therefore, the approach can account for phenomena such as retrospective evaluation effects. However, further clarifications are clearly needed with regard to the involved cognitive processes and memory systems. The major disadvantage of such an approach is that it does not specify the processes during the actual course of learning and, furthermore, that the account provides no means to express differences in the way the learning data is experienced. For example, Experiments 5 and 6 demonstrate that manipulations of the temporal order during observational learning can affect learners' causal inferences. Explaining this result requires a process explanation that spells out the details of how learning from single trial takes place.

An interesting approach would be to integrate trial-based learning mechanisms into the causal Bayes nets framework. This is where the strength of associative learning theories comes into play. Learning models such as the Rescorla-Wagner rule provide a precise description of the processes assumed to take place in trial-by-trial learning. Since the long-run estimates of such algorithms often converge to probability parameters, they offer the possibility to explain how people acquire probabilities without performing explicit calculations. Recently, an iterative learning algorithm has also been proposed that converges to the predictions of the power PC model (Danks, Griffiths, & Tenenbaum, 2003). This is particularly interesting since in the case of a single cause-effect relation the power PC theory is formally equivalent to a noisy-OR gate parameterization of a common-effect model (cf. Glymour, 2003). Thus, introducing such trial-based learning procedures could offer an opportunity for combining the computational level description of causal Bayes nets with a genuine learning model.

In conclusion, I suggest that the causal Bayes nets formalism provides a comprehensive and coherent formal framework of causal representation, causal learning, and causal reasoning. Causal Bayes nets may not be the Grand Unifying Theory of causality, but the account is clearly a great leap towards a deeper understanding of causal cognition.

7 References

Ahn, W.-K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299-352.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin, 114*(3), 435-448.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Psychology, 34*, 1-11.

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation, 14*, 381-405.

Allan, L. G., & Tangen, J. M. (2005). Judging relationships between events: How do we do it? *Canadian Journal of Experimental Psychology, 59*(1), 22-27.

Baker, A. G., Mercier, P., Vallée-Tourangeau, A. F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: The presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19*, 414-432.

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*, 1020-1022.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1119-1140.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning, 8*(4), 269-295.

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *Quarterly Journal of Experimental Psychology, 56A*(5), 865-890.

Buehner, M. J., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *Quarterly Journal of Experimental Psychology, 57B*(2), 179-191.

Bullock, M., & Gelman, R. (1979). Preschool children's assumptions about cause and effect: Temporal ordering. *Child Development, 50*, 89-96.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon Press.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Clarendon Press.

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist, 84*(2), 242-264.

Cartwright, N. (2002). Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science, 53*(3), 411-453.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(5), 837-854.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18*(5), 537-545.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367-405.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*(4), 545-567.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40*, 83-120.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99*(2), 365-382.

Cobos, P. L., López, F. J., Cano, A., Almaraz, J., & Shanks, D. R. (2002). Mechanisms of predictive and diagnostic causal induction. *Journal of Experimental Psychology: Animal Behavior Processes, 28*(4), 331-346.

Coups, E. J., & Chapman, G. B. (2002). Formation and use of covariation assessments in the real world. *Applied Cognitive Psychology, 16*(1), 51-71.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology, 47*, 109-121.

Danks, D. (in press). The Supposed Competition Between Theories of Human Causal Inference. *Philosophical Psychology.*

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical Causal Learning. In S. Becker, S. Thrun & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 67-74). Cambridge, Mass: MIT Press.

De Houwer, J. (2002). Forward blocking depends on retrospective inferences about the presence of the blocked cue during the elemental phase. *Memory and Cognition, 30*, 22-33.

De Houwer, J., & Beckers, T. (2002a). Second-order backward blocking and unovershadowing in human causal learning. *Experimental Psychology, 49*, 27-33.

De Houwer, J., & Beckers, T. (2002b). Higher-order retrospective revaluation in human causal learning. *Quarterly Journal of Experimental Psychology, 55B*, 137-151.

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

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology, 49B*, 60-80.

Dickinson, A., & Shanks, D. (1995). Instrumental action and causal representation. In D. Sperber, D. Premack & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate.* (pp. 5-25). Oxford: Clarendon Press.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A*(1), 29-50.

Doll, R., & Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking. A second report on the mortality of British doctors. *British Medical Journal, 233*, 1071-1076.

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

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge, England: Cambridge University Press.

Eells, E. (1991). *Probabilistic causality*. Cambridge: Cambridge University Press.

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

Fisher, R. A. (1958). Smoking, cancer, and statistics. *Centennial Review, 2*, 151-166.

Ford, E. C., & Cheng, P. W. (2004). Sensitivity to confounding in causal inference: From childhood to adulthood. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 398-403). Hillsdale, NJ: Erlbaum.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

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.

Glymour, C., & Cooper, G. F. (1999). *Computation, causation, and discovery*. Menlo Park, MA: MIT Press.

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., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620-629.

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

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (in press). Causal reasoning through intervention. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.*Oxford: Oxford University Press.

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 214-219). Mahwah, NJ: Erlbaum.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30*(7), 1128-1137.

Harré, R., & Madden, E. H. (1975). *Causal powers: A theory of natural necessity.* Oxford, England: Blackwell.

Hausman, D., & Woodward, J. (1999). Independence, invariance, and the causal Markov condition. *British Journal for the Philosophy of Science, 50*, 521-583.

Heckerman, D., Meek, C., & Cooper, G. F. (1999). A Bayesian approach to causal discovery. In C. Glymour & G. Cooper (Eds.), *Computation, Causation, and Discovery* (pp. 143-167). Cambridge, MA: MIT Press.

Holyoak, K. J., & Thagard, R. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Hume, D. (1739/2000). *A treatise of human nature*. Oxford: Oxford University Press.

Hume, D. (1748/1993). *An enquiry concerning human understanding* (2nd ed.). Indianapolis: Hackett Publishing Company.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-291.

Kahneman, D., & Tversky, A. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9-33). Miami: University of Miami Press.

Kant, I. (1781/1974). *Kritik der reinen Vernunft*. Frankfurt: Suhrkamp.

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(6), 1363-1386.

Keppel, G., & Wickens, T. D. (2000). *Design and analysis*. Upper Saddle River, NJ: Pearson.

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method and sample size in causal reasoning. *Child Development, 60*, 1316-1328.

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

Lagnado, D., & Sloman, S. A. (in press). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (in press). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.*Oxford: Oxford University Press.

Laplace, P. S. (1814/1912). *Essai philosophique sur les probabilities.*New York: Wiley.

Larkin, M. J., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgements under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*, 1331–1352.

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

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review, 107*(1), 195-212.

Lopez, F. J., Shanks, D. R., Almaraz, J., & Fernandez. (1998). Effects of trial order on contingency judgments: A comparison of associative and probabilistic contrast accounts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 672-694.

Luhmann, C. C. (2005). Confounded: Causal inference and the requirement of independence. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 1355-1360). Mahwah, NJ: Erlbaum.

Luhmann, C. C., & Ahn, W.-K. (2005). The meaning and computation of causal power: Comment on Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 112*(3), 685-692.

Mackie, J. L. (1974). *The cement of the universe*. Oxford, UK: Clarendon.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2005). Doing after Seeing. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 1461-1466). Mahwah, NJ: Erlbaum.

Melz, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner learning rule? Comments on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1398-1410.

Mendelson, R., & Shultz, T. R. (1976). Covariation and temporal contiguity as principles of causal inference in young children. *Journal of Experimental Child Psychology, 22*(3), 408-412.

Michotte, A. E. (1963). *The perception of causality.* London: Methuen & Co.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin, 117*, 363-386.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*(2), 455-485.

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

Pearl, J. (2000). *Causation: Models, reasoning and inference*. Cambridge, MA: Cambridge University Press.

Pearson, K., Lee, A., & Bramley-Moore, L. (1899). Genetic (reproductive) selection: Inheritance of fertility in man. *Philosophical Transactions of the Royal Society, Series A, 73*, 534-539.

Price, C. P., & Yates, J. F. (1993). Judgmental overshadowing: Further evidence of cue interaction in contingency judgment. *Memory & Cognition, 21*(5), 561-572.

Price, P. C., & Yates, J. F. (1995). Associative and rule-based accounts of cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(6), 1639-1655.

Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141-1159.

Rehder, B. (2003b). Categorization as causal reasoning. *Cognitive Science, 27*, 709-748.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264-314.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130*, 323-360.

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

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*, 1-5.

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relations between Pavlovian conditioning and instrumental learning. *Psychological Review, 74*, 151-182.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II. Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Salmon, W. C. (1971). *Statistical explanation and statistical relevance*. Pittsburgh: University of Pittsburgh Press.

Salmon, W. C. (1980). Probabilistic causality. *Pacific Philosophical Quarterly, 61*, 50-74.

Scheines, R., Spirtes, P., Glymour, C., & Meek, C. (1994). *TETRAD II*. Hillsdale, NJ: Lawrence Erlbaum.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110*(1), 101-120.

Shaklee, H., & Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development, 52*, 317-325.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology, 37B*, 1-21.

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

Shanks, D. R. (1993). Human instrumental learning: A critical review of data and theory. *British Journal of Psychology, 84*(3), 319-354.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In *The psychology of learning and motivation: Advances in research and theory, Vol. 21.* (pp. 229-261). San Diego, CA: Academic Press.

Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition, 19*(4), 353-360.

Shanks, D. R., Lopez, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In D. R. Shanks, K. Holyoak & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning* (pp. 265-311). San Diego, CA, US: Academic Press.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *Quarterly Journal of Experimental Psychology, 41B*(2), 139-159.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgment of causality by human subjects. *Quarterly Journal of Experimental Psychology, 41B*(2), 139–159.

Shimazaki, T., & Tsuda, Y. (1991). Strategy changes in human contingency judgments as a function of contingency tables. *The Journal of General Psychology, 118*(4), 349-360.

Shultz, T. R., & Mendelson, R. (1975). The use of covariation as a principle of causal analysis. *Child Development, 46*(2), 394.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B (Methodological), 13*, 238–241.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29*(1), 5-39.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7*(6), 337-342.

Spirtes, P., Glymour, C., & Scheines, P. (1993). *Causation, prediction, and search.* New York: Springer-Verlag.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453-489.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North Holland.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 135–170.

Tangen, J. M., & Allan, L. G. (2004). Cue interaction and judgments of causality: Contributions of causal and associative processes. *Memory & Cognition, 32*(1), 107-124.

Tangen, J. M., Allan, L. G., & Sadeghi, H. (Eds.). (2005). *Assessing (in)sensitivity to causal asymmetry: A matter of degree*. Mahwah, NJ: Lawrence Erlbaum Associates.

Vallée-Tourangeau, A. F., Murphy, R. A., Drew, S., & Baker, A. G. (1998). Judging the importance of constant and variable candidate causes: A test of the power PC theory. *The Quarterly Journal of Experimental Psychology, 51A*(1), 65-84.

Van Hamme, L. J., & Wasserman, E. A. (1993). Cue competition in causality judgments: The role of manner in information presentation. *Bulletin of the Psychonomic Society, 31*(5), 457-460.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*(2), 127-151.

von Sydow, M. (2006). *Towards a flexible Bayesian and deontic logic of testing descriptive and prescriptive rules.* Unpublished doctoral dissertation, University of Göttingen.

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, US: Academic Press.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 53-76.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8*(3), 600-608.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: the role of structural knowledge and processing effort. *Cognition, 82*(1), 27-58.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 216-227.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*(2), 222-236.

Waldmann, M. R., & Holyoak, K. J. (1997). Determining whether causal order affects cue selection in human contingency learning: comments on Shanks and Lopez (1996). *Memory & Cognition, 25*(1), 125-134.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124*, 181-206.

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

Waldmann, M. R., & Walker, J. M. (2005). Competence and performance in causal learning. *Learning & Behavior, 33*(2), 211-229.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19*, 231-241.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135-151). Harmondsworth, Middlesex, UK: Penguin.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation, 14*(4), 406-432.

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 509-521.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 174-188.

Wasserman, E. A., Kao, S.-F., Van Hamme, L. J., Katagiri, M., & Young, M. E. (Eds.). (1996). *Causation and association*. San Diego, CA,US: Academic Press.

White, P. A. (2002). Causal attribution from covariation information: The evidential evaluation model. *European Journal of Social Psychology, 32*(5), 667-684.

White, P. A. (2004). Causal judgement from contingency information: A systematic test of the pCI rule. *Memory and Cognition, 32*, 353-368.

White, P. A. (2005). The power PC theory and causal powers: Comment on Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 112*(3), 675-682.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *1960 IRE WESCON Convention Record (Pt. 4)*, 96-104.

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

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science, 10*(2), 92-97.

Acknowledgments

In concluding this thesis, I would like to acknowledge the help of the many people who supported me in the past three years during the course of my PhD program.

First, I would particularly like to thank my academic supervisors, Michael Waldmann and York Hagmayer, for proposing such an interesting and rewarding research topic. I have been working for and with them since I was as an undergraduate, and I am grateful to them for introducing me to the field of causality and causal cognition. It has always been a pleasure working with them, and I have not only benefited from their academic advice and the many fruitful discussions we have had, but I also enjoyed the friendly atmosphere they created. I also thank Momme von Sydow, who has not only been an excellent colleague in the past three years, but also became a true friend during this time. We started our PhD programs together, and it is great that we are now finishing them together.

There are a number of other people whose contributions to this thesis should also be acknowledged. I would like to thank my second examiner, Professor Gerd Lüer, and the other members of the thesis committee for their assistance. I also would like to thank Julia Iwen and Irene Warnecke for collecting some of the data, and I appreciate very much the support of Shane Denson, whose assistance in proofreading the thesis was a great help. Further, I am very thankful to my brother, to my friends, and especially to my flatmates, who were kind enough not to ask too many questions after a long day pondering about my thesis. My endurance through the writing of this dissertation also owes a great debt to Susanne for her support and encouragement, and for filling my heart with her smile.

Finally, I am deeply grateful to my parents, Ilona and Kurt, whose love and support has been with me throughout my life.

Curriculum Vitae

Name: Björn Meder

Born in Hannover, Germany, on September 30th, 1976.

since 2003	DFG-Project "Kategorisierung und induktives Lernen" Georg-August-Universität Göttingen, Department of Psychology
2003	*Diplom* in Psychology
1997 – 2003	Georg – August – Universität Göttingen
2001 – 2002	University of Warwick, UK
1996 – 1997	Civil Service (Paracelsus Hospital, Langenhagen)
1996	Abitur
1987 – 1996	Integrierte Gesamtschule Langenhagen
1983 – 1987	Adolf-Reichwein Elementary School, Langenhagen