Psychological Review
2014, Vol. 121, No. 3, 277–301

0033-295X/14/$12.00 DOI: 10.1037/a0035944

# Structure Induction in Diagnostic Causal Reasoning

Björn Meder
Max Planck Institute for Human Development, Berlin, Germany

Ralf Mayrhofer and Michael R. Waldmann
University of Göttingen

Our research examines the normative and descriptive adequacy of alternative computational models of diagnostic reasoning from single effects to single causes. Many theories of diagnostic reasoning are based on the normative assumption that inferences from an effect to its cause should reflect solely the empirically observed conditional probability of cause given effect. We argue against this assumption, as it neglects alternative causal structures that may have generated the sample data. Our structure induction model of diagnostic reasoning takes into account the uncertainty regarding the underlying causal structure. A key prediction of the model is that diagnostic judgments should not only reflect the empirical probability of cause given effect but should also depend on the reasoner's beliefs about the existence and strength of the link between cause and effect. We confirmed this prediction in 2 studies and showed that our theory better accounts for human judgments than alternative theories of diagnostic reasoning. Overall, our findings support the view that in diagnostic reasoning people go "beyond the information given" and use the available data to make inferences on the (unobserved) causal rather than on the (observed) data level.

*Keywords:* diagnostic reasoning, causal reasoning, causal models, Bayesian inference, uncertainty

*Supplemental materials:* http://dx.doi.org/10.1037/a0035944.supp

Diagnostic inferences are ubiquitous not only in medicine but also in everyday reasoning. For example, we reason from effects to causes when we try to explain why our car does not start, why our significant other is angry with us, or why our job application was rejected. In the present research, we focused on the most basic type of diagnostic reasoning involving a single cause–effect relation between two binary variables. We refer to such inferences as *elemental diagnostic reasoning*.

Research on diagnostic reasoning has a long history in cognitive psychology, especially in the context of the "heuristics and biases" framework (Tversky & Kahneman, 1974). The typical result of this research is that human diagnostic reasoning is deeply flawed and not in line with statistical norms, such as Bayes' rule. Kahneman and Tversky (1973; Tversky & Kahneman, 1974, 1982) reported numerous experiments indicating that people tend to ignore prior probabilities (i.e., base rates) in belief updating and give too much weight to likelihoods (but see Edwards, 1968; Peterson & Beach, 1967). Similar findings were obtained in hypothetical medical diagnosis tasks in which participants had to assess the posterior probability of having a disease, given a positive test result and the prior probability of the disease (Eddy, 1982).

Although these findings seem to indicate that human diagnostic reasoning is often biased and error prone, more recently the scope of many of these phenomena has been questioned. Koehler (1996) and Barbey and Sloman (2007) have summarized research showing that under specific learning and testing conditions base rates are actually appreciated. For example, Gigerenzer and Hoffrage (1995; see also Cosmides & Tooby, 1996; Sedlmeier & Gigerenzer, 2001) showed that presenting information in terms of natural frequencies, rather than as conditional probabilities, substantially improves participants' diagnostic inferences. This is in line with research showing that people often perform well in real-world diagnostic tasks (Christensen-Szalanski & Bushyhead, 1981).

Causal knowledge is another important factor in diagnostic reasoning. For example, Ajzen (1977; see also Tversky & Kahneman, 1982) showed that increasing the causal relevance of base rate information improves people's capacity to reason in accordance with Bayes' rule when making diagnostic inferences. Krynski and Tenenbaum (2007) demonstrated that causal models that highlight the presence of alternative explanations of the evidence (e.g., alternative causes that can lead to a positive mammogram in breast cancer screening) help people make better diagnostic judgments. Finally, Fernbach, Darlow, and Sloman (2010, 2011) com-

Björn Meder, Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development, Berlin, Germany; Ralf Mayrhofer and Michael R. Waldmann, Department of Psychology, University of Göttingen.

Experiment 1 was presented at the 2008 Annual Meeting of the Society for Mathematical Psychology in Washington, DC, and the 2009 Annual Conference of the Cognitive Science Society in Amsterdam, the Netherlands. This research was supported by Deutsche Forschungsgemeinschaft (DFG) Grant Wa 621/20 as well as DFG Grants ME 3717/2 and Wa 621/22 as part of the priority program "New Frameworks of Rationality" (SPP 1516). We thank Jana Samland, Dana Barthels, Mira Holzer, and Philipp Ulbrich for help in preparing the experiments and collecting the data, and we thank Patricia Cheng, Keith Holyoak, Hongjing Lu, and Jonathan Nelson for helpful comments on the project. We also thank Anita Todd and Rona Unrau for editing the manuscript.

Correspondence concerning this article should be addressed to Björn Meder, Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: meder@mpib-berlin.mpg.de

pared predictive reasoning from cause to effect with diagnostic reasoning from effect to cause, using basic causal Bayes nets (Cheng, 1997; Pearl, 2000) as the normative benchmark. They showed that people are more sensitive to the presence and strength of alternative causes when making diagnostic inferences from effects to causes than when making predictive inferences from causes to effects.

## Scope and Goals

The main focus of the current work is the normative and descriptive adequacy of alternative computational models of diagnostic reasoning. The primary scope of these models concerns situations in which a data sample about the covariation of a single binary cause (e.g., a virus) and a single binary effect (e.g., a substance in the blood of patients) provides the basis for making diagnostic judgments (e.g., probability of virus given substance) (i.e., elemental diagnostic reasoning).

Whereas the traditional normative benchmark for such inferences is provided by purely statistical models such as Bayes' rule, we analyze diagnostic reasoning from the perspective of causal inference under uncertainty. In particular, we focus on uncertainty with respect to the causal structure that may have generated the observed data. We will show that a simple causal inference theory that neglects uncertainty, *power PC theory* (Cheng, 1997; Fernbach et al., 2011), makes identical predictions to a purely statistical approach (i.e., Bayes' rule). However, modeling diagnostic reasoning as a causal inference under uncertainty yields predictions that diverge from the commonly accepted norm. We will present a novel model, the *structure induction model of diagnostic reasoning*, which assumes that diagnostic judgments are constrained by assumptions about alternative causal structures that may have generated the observed data. This approach takes into account the uncertainty regarding the causal structure of the environment and the uncertainty associated with parameter estimates, such as the cause's base rate and its causal strength.

A key prediction of our model is that diagnostic judgments should not only reflect the empirical conditional probability of cause given effect in the sample data. Rather, judgments should also take into account to what extent the available data support the existence of a causal relation between the candidate cause and the candidate effect, even when the empirical probability of cause given effect is fixed. Take, for example, the case of a causal relation between a disease and a symptom. The model predicts that the observation of the symptom should lead to higher diagnostic inferences the stronger the belief is in the existence of a causal relation between the disease and the symptom. By contrast, if the available data provide only limited evidence for the existence of a causal link, the model predicts that the diagnostic inference should be lower in this case, even when the conditional probability of cause given effect in the data sample is the same. Thus, even if one observes an identical probability of the disease given the symptom in different data sets, this does not mean that the diagnostic judgments should be invariant.

We tested this prediction in two experiments using a learning paradigm in which participants were provided with frequency information about the covariation of a binary cause event and a binary effect event. The base rate of the cause event was always set to 50%. Thus, in the present experiments we did not focus on the appreciation of base rates but rather on the role of causal structure in diagnostic inferences. We analyzed and tested our theory against a number of alternative models, such as Bayesian variants of power PC theory (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008) and different models of causal attribution (Cheng & Novick, 2005; Holyoak, Lee, & Lu, 2010).

## Models of Elemental Diagnostic Reasoning

In the following sections, we will discuss three different computational models of elemental diagnostic reasoning (we will later consider additional models). We will start off with a basic statistical model, *simple Bayes*, which for many years served as the normative benchmark. Whereas this model does not make any assumptions about the relation between data and an underlying generative causal structure, the other two, power PC theory (in this case equivalent to a simple causal Bayes net; see Cheng, 1997; Fernbach et al., 2011) and our structure induction model, share the assumption that diagnostic inferences operate over causal structure representations that are inferred from data. That is, diagnostic inferences take place on the causal, rather than on the data level. The observed data are assumed to be noisy and are used only as a proxy for inferring the existence and strength of the underlying, not directly observable causal relations.

A core feature of causal representations is that they mirror a characteristic property of our environment, namely, that some events—causes—have the power to generate or prevent other events—their effects (Cheng, 1997; Waldmann, Hagmayer, & Blaisdell, 2006). Causal representations allow us to make different types of probabilistic inferences (e.g., Buehner, Cheng, & Clifford, 2003; Cheng, 1997; Cheng & Novick, 2005; Meder & Mayrhofer, 2013; Waldmann, Cheng, Hagmayer, & Blaisdell, 2008), generalizations across different contexts (Liljeholm & Cheng, 2007), and inferences regarding interventions (Hagmayer & Meder, 2013; Hagmayer & Sloman, 2009; Meder, Hagmayer, & Waldmann, 2008, 2009; Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). These studies support the view that people have a natural tendency to go "beyond the data given" by inducing representations that mirror the causal structure of the environment.

### Simple Bayes: Empirical Probabilities as the Classical Norm of Diagnostic Reasoning

If a diagnostic judgment from effect to cause is required, such as estimating the probability of a disease given the presence of a symptom, it seems natural to assess the conditional probability of cause given effect. The diagnostic conditional probability, either obtained directly from the sample data or derived by Bayes' rule, has been endorsed by many researchers as the natural normative standard, although there is disagreement about whether people's reasoning conforms to this norm (e.g., Barbey & Sloman, 2007; Eddy, 1982; Gigerenzer & Hoffrage, 1995; Kahneman & Tversky, 1973; Koehler, 1996; Krynski & Tenenbaum, 2007).

Let $C$ denote a binary cause event and $E$ a binary effect, and let $\{c, \neg c\}$ and $\{e, \neg e\}$ indicate the presence and absence of these events. Given a joint frequency distribution over $C$ and $E$, the empirical conditional probability of a cause given its effect, $P(c|e)$, can be directly estimated from the observed relative frequencies (see left panel of Figure 1), or inferred by using Bayes' rule:

$$P(c|e) = \frac{P(e|c) \cdot P(c)}{P(e)} = \frac{P(e|c) \cdot P(c)}{P(e|c) \cdot P(c) + P(e|\neg c) \cdot P(\neg c)}. \quad (1)$$

Under the simple Bayes account, no reference is made to the generative causal processes underlying the observed events, and no uncertainty regarding parameter estimates is incorporated in the model. This model is strictly noncausal in that it can be applied to arbitrary statistically related events (see Waldmann & Hagmayer, 2013).

## Power PC Theory: Diagnostic Inferences Under Causal Power Assumptions

Cheng's (1997) power PC theory was the first theory in psychology to separate empirical indicators of causal strength (i.e., covariation) from estimates of unobservable causal power (see Waldmann & Hagmayer, 2013, for an overview). The theory assumes that people aim to infer causal power because one distal goal of cognitive systems is to acquire knowledge of stable causal relations rather than arbitrary statistical associations in noisy environments.

Power PC theory focuses on a default common-effect structure with an observable cause $C$ and an amalgam of unobservable background causes $A$ (graph $S_1$ in Figure 2). An estimate for the strength of the background cause(s), $w_a$, is given by $P(e|\neg c)$ (for mathematical convenience, $A$ is assumed to be constantly present; see Griffiths & Tenenbaum, 2005). The unobservable probability with which $C$ produces $E$ is called *generative causal power*, denoted $w_c$:

$$w_c = \frac{P(e|c) - P(e|\neg c)}{1 - P(e|\neg c)}. \quad (2)$$

This estimate of causal strength differs from the conditional probability of the effect given its cause, $P(e|c)$, because it "partials out" the influence of alternative causes that may also have generated the effect (see Cheng, 1997, for a detailed analysis).

According to power PC theory, people make the default assumptions that $C$ and $A$ independently influence $E$, that $A$ produces but does not prevent $E$, that causal powers are independent of the frequencies of $C$ and $A$, and that $E$ does not occur without being caused by either $C$ or $A$. These assumptions instantiate a particular generative causal structure known as a *noisy-OR* gate (Cheng, 1997; Glymour, 2003; Griffiths & Tenenbaum, 2005; Pearl, 1988), according to which the probability of effect given cause is given by

$$P(e|c;\, w_c, w_a) = w_c + w_a - w_c w_a, \quad (3)$$

where $w_c$ and $w_a$ denote the causal powers of target cause $C$ and background cause $A$, respectively.

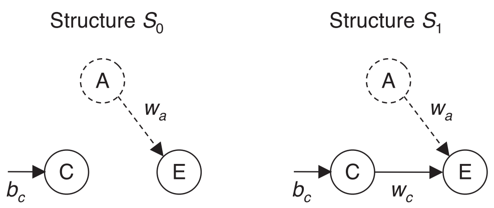


*Figure 2.* Alternative causal structures in the structure induction model. $C$ and $E$ denote a binary cause and effect event, respectively. According to structure $S_0$, there is no causal relation between candidate cause $C$ and candidate effect $E$, whereas structure $S_1$ states that there potentially exists a causal relation between $C$ and $E$. $S_1$ is the default structure in power PC theory. Parameter $b_c$ denotes the base rate of $C$, and parameters $w_c$ and $w_a$ represent the causal strengths of target cause $C$ and (unobservable) background cause $A$, respectively.

Although the primary focus of power PC theory has been on estimates of causal strength and predictive inferences, the account can also be applied to diagnostic inferences (Cheng & Novick, 2005; Waldmann et al., 2008). From the noisy-OR parameterization, it follows that the diagnostic probability of cause given effect is given by

$$P(c|e;\, b_c, w_c, w_a) = \frac{P(e|c) \cdot P(c)}{P(e)}$$

$$= \frac{(w_c + w_a - w_c w_a) \cdot b_c}{(w_c + w_a - w_c w_a) \cdot b_c + w_a(1 - b_c)} = \frac{w_c b_c + w_a b_c - w_c w_a b_c}{w_c b_c + w_a - w_c w_a b_c}, \quad (4)$$

where $w_c$ denotes the causal power of candidate cause $C$, $b_c$ is an estimate of the base rate of cause $C$, and $w_a$ corresponds to the causal power of the unobserved background $A$. Equation 4 specifies how the diagnostic probability of a cause given its effect can be derived under causal power assumptions (Waldmann et al., 2008).

Conceptually, the power PC model of diagnostic reasoning distinguishes between the (observable) data and the (unobservable) causal level and uses data to estimate causal parameters. However, because all parameters involved in these computations are maxi-

| | Effect present | Effect absent |
|---|---|---|
| Cause present | $N(c, e)$ | $N(c, \neg e)$ |
| Cause absent | $N(\neg c, e)$ | $N(\neg c, \neg e)$ |

| | Substance present | Substance absent |
|---|---|---|
| Disease present | ☹ + | ☹ − |
| Disease absent | ☺ + | ☺ − |

*Figure 1.* Contingency table (left) and experimental stimuli (right) used to represent the co-occurrences ($N$) of cause $C = \{c \text{ vs. } \neg c\}$ and effect $E = \{e \text{ vs. } \neg e\}$ in Experiments 1 and 2.

mum likelihood estimates directly derived from the sample data, the inferred diagnostic probability corresponds exactly to the empirical probability in the data. Thus, the power PC model effectively yields the same numeric predictions as the simple Bayes approach.

Fernbach et al. (2011), examining both predictive and diagnostic reasoning, tested a causal Bayes net that is formally equivalent to the power PC model described in this section. Thus, like the standard power PC model, their causal Bayes net model is not sensitive to the uncertainty of causal structures and their parameters. The results of their experiments seem consistent with the predictions of the standard causal Bayes net in diagnostic reasoning cases. However, their paradigm is not ideal to test the role of uncertainty because the studies were based on already acquired real-world knowledge rather than learning data, which makes it difficult to control levels of uncertainty.

## Structure Induction Model: Diagnostic Causal Reasoning With Structure Uncertainty

In this section, we present a new model of elemental diagnostic inference that goes beyond simple Bayes and the power PC framework (see also Meder, Mayrhofer, & Waldmann, 2009). Our structure induction model of diagnostic reasoning takes into account both the uncertainty regarding the underlying causal structure and the uncertainty regarding the parameters.

The characteristic feature of the structure induction model is that it does not operate on a single causal structure, as does power PC theory, but estimates the posterior probability of alternative causal structures given the observed data (Anderson, 1990; Griffiths & Tenenbaum, 2005, 2009; see also Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). According to this model, hypotheses about alternative causal structures and the existence of a causal relation between *C* and *E* constrain and guide diagnostic inferences. Importantly, taking into account structure uncertainty can lead to estimates of diagnostic probability that systematically deviate from the simple Bayes and power PC estimates.

We used the framework of Bayesian inference over graphical causal networks (i.e., causal Bayes nets) to implement the model. Analogous to the model of Griffiths and Tenenbaum (2005; see also Anderson, 1990), our model considers two alternative causal structure hypotheses that differ with respect to the presumed existence of a causal relation between *C* and *E*. The account uses the sample data to estimate the structures' parameters and posterior probabilities. To arrive at an estimate for a diagnostic inference that reflects the uncertainty with respect to the true underlying causal structure, the causal structure hypotheses are integrated out (Bayesian model averaging; Chickering & Heckerman, 1997). The resulting diagnostic estimate then reflects the uncertainty regarding the presence of a causal link between *C* and *E*. This is crucial for diagnostic inferences because an effect *E* only provides evidence for *C* within the underlying causal model if *C* and *E* are (directly or indirectly) linked, but not when these two events are unrelated.

In sum, the key mechanism behind the structure induction model is that the uncertainty about the underlying causal structure and the existence of a causal link between *C* and *E* are taken into account by averaging over possible causal structures when making a diagnostic inference from effect to cause. We next describe the computational steps of the model at the conceptual level; a detailed formal description is provided in Appendix A.

**Causal structure hypotheses.** Given a joint distribution over a cause *C* (e.g., virus) and an effect *E* (e.g., symptom) the structure induction model considers two alternative causal structures that might underlie the data. These graphs, denoted $S_0$ and $S_1$, are shown in Figure 2 (cf. Anderson, 1990; Griffiths & Tenenbaum, 2005, 2009). The (prior) probability distribution over the structures is denoted $P(S)$.

Each of the two graphs represents a hypothesis about a different generative causal structure potentially underlying the observed data. According to structure $S_0$, *C* and *E* are independent events; that is, there is *no* causal relation between candidate cause *C* and candidate effect *E*. Although the two events may sometimes coincidentally co-occur, the effect is exclusively generated by unobserved (independent) background causes *A*. The second structure, $S_1$, is the default causal structure in power PC theory (Cheng, 1997; Glymour, 2003; Griffiths & Tenenbaum, 2005). According to this structure, there exists a causal relation between *C* and *E*, but there are also alternative background causes *A* that can independently generate the effect (for evidence that such independence assumptions are the default in human causal reasoning, see, e.g., Hagmayer & Waldmann, 2007; Mayrhofer, Nagel, & Waldmann, 2010; but see Luhmann & Ahn, 2007).

Note that graph $S_0$ is not merely a special case of graph $S_1$ but constitutes a qualitatively different, less complex structure hypothesis that suggests an alternative explanation of the data. Importantly, despite its simpler form, causal structure $S_0$ can have a higher posterior probability than structure $S_1$. This is because using Bayesian inference over causal structure hypotheses allows it to capitalize on Bayesian Occam's razor (MacKay, 2003). Loosely speaking, Occam's razor states that if there are two models (theories, hypotheses) that explain the data equally well, then the simpler model should be preferred. Conversely, if we compare two models differing in complexity (e.g., in the number of parameters), the plausibility of the two models should not be evaluated only with respect to how well they predict the data, as more complex models can generate a greater variety of predictions. Bayesian inference embodies Occam's razor automatically since a model that makes more diverse predictions must spread its probability mass across many predictions, whereas the probability mass of simpler models will concentrate on the few predictions they are capable of making. As a consequence, simpler models can achieve a higher posterior probability when the data conform well with their (more limited) predictions, without assigning a simpler model a higher prior probability.

**Estimating causal structure parameters.** Associated with each of the two causal structures is a set of parameters (Figure 2), representing the base rate of the target cause ($b_c$), the causal strength of *C* with respect to *E* ($w_c$), and the strength of the background cause *A* ($w_a$). In the structure induction model, the parameters' posterior distributions are derived separately for each of the two causal structures, using Bayesian inference (see Appendix A). For structure $S_1$, parameters $b_c$, $w_c$, and $w_a$ are estimated. Structure $S_0$ has only two parameters: According to this structure there is no causal relation between *C* and *E*; therefore, only estimates for $b_c$ and $w_a$ are derived (i.e., the strength of *C*, $w_c$, is set to zero).

Given some data $D$, different posterior parameter distributions result under the two causal structures. Assuming a noisy-OR parameterization of the graphs and independent uniform (i.e., flat) Beta(1, 1) priors over the parameters, the mean posterior estimates under $S_1$ will approximate the maximum likelihood estimates of standard power PC theory. By contrast, under graph $S_0$ (which states that $C$ has no causal impact on $E$), the impact of the background cause may be overestimated, reflecting the base rate of the effect in the data, as the occurrence of the effect is attributed to $A$ alone.

**Posterior probabilities of causal structure hypotheses.** The posterior probability of the two causal structure hypotheses, $S_0$ and $S_1$, is proportional to the likelihood of the data given a structure, $P(D|S_i)$, weighted by the prior probability of the structure, $P(S_i)$ (see Appendix A). For our simulations, we assumed a uniform prior over the structures, that is, $P(S_0) = P(S_1) = 1/2$. Depending on the match between data and causal structure, different posteriors for the two causal structures result. In particular, the weaker the contingency between cause and effect, the more likely is $S_0$, which implies that there is no causal relation between $C$ and $E$.

**Integrating out the causal structures.** Given the parameterized causal structures and their posterior probability, one can derive different quantities of interest, such as diagnostic and predictive probabilities, using Bayesian model averaging (see Appendix A). For instance, to derive an estimate of the diagnostic probability of cause given effect, $P(c|e)$ is computed separately under each of the two (parameterized) causal structures, $S_0$ and $S_1$. To obtain a single estimate for the diagnostic probability, the structures are integrated out by summing over the estimates, with each estimate being weighted by the posterior probability of the respective graph. This diagnostic probability then reflects the uncertainty regarding the true underlying causal model, as well as the uncertainty of the parameter estimates. The same procedure can be applied to derive estimates of the predictive probability of effect given cause, $P(e|c)$ (see below and Appendix A).

## Differential Predictions of the Competing Models of Diagnostic Reasoning

One of the key differences between the models of diagnostic reasoning concerns the role of predictive probability and causal strength in diagnostic judgments. Both simple Bayes and power PC theory predict that diagnostic inferences should correspond to the diagnostic probability, $P(c|e)$, in the data sample. Therefore, the predictive probability, $P(e|c)$, and the causal strength of $C$ should not affect diagnostic inferences as long as $P(c|e)$ stays invariant. In contrast, the structure induction model predicts that estimates of causal strength and predictive probability should influence diagnostic judgments, as they influence the posterior probability of structures $S_0$ and $S_1$.

Figure 3 (left column) illustrates the diverging predictions of the structure induction model and the simple Bayes model for diagnostic inferences. Three data sets are considered in which the empirical diagnostic probability is invariant at $P(c|e) = .75$ and the base rate of the cause is $P(c) = .5$ (see Table 1 for numerical values). The diagnostic probability is identical in all three data sets, but the predictive probability of the effect given its cause, $P(e|c)$, takes the value .3, .6, or .9 (see Figure 3, top left).

Figure 3 shows how the structures' posterior probabilities (middle left) vary depending on the observed data (top left), and how these differences, in turn, influence the diagnostic probabilities (bottom left). The important feature here is that the posterior probabilities of $S_0$ and $S_1$ vary as a function of the sample data. In the data set in which the predictive probability $P(e|c) = .3$, causal structures $S_0$ and $S_1$ are about equally likely to have generated the data. This mirrors the intuition that the weak empirical contingency between $C$ and $E$ is not reliable enough to conclude that there is indeed a causal relation between $C$ and $E$. As a consequence, the diagnostic probability derived from the structure induction model is substantially lower than the empirical diagnostic probability (.61 vs. .75). More intuitively, the higher the posterior probability of $S_0$, the closer is the diagnostic probability to the base rate of the cause, which is .5 in all three data sets.

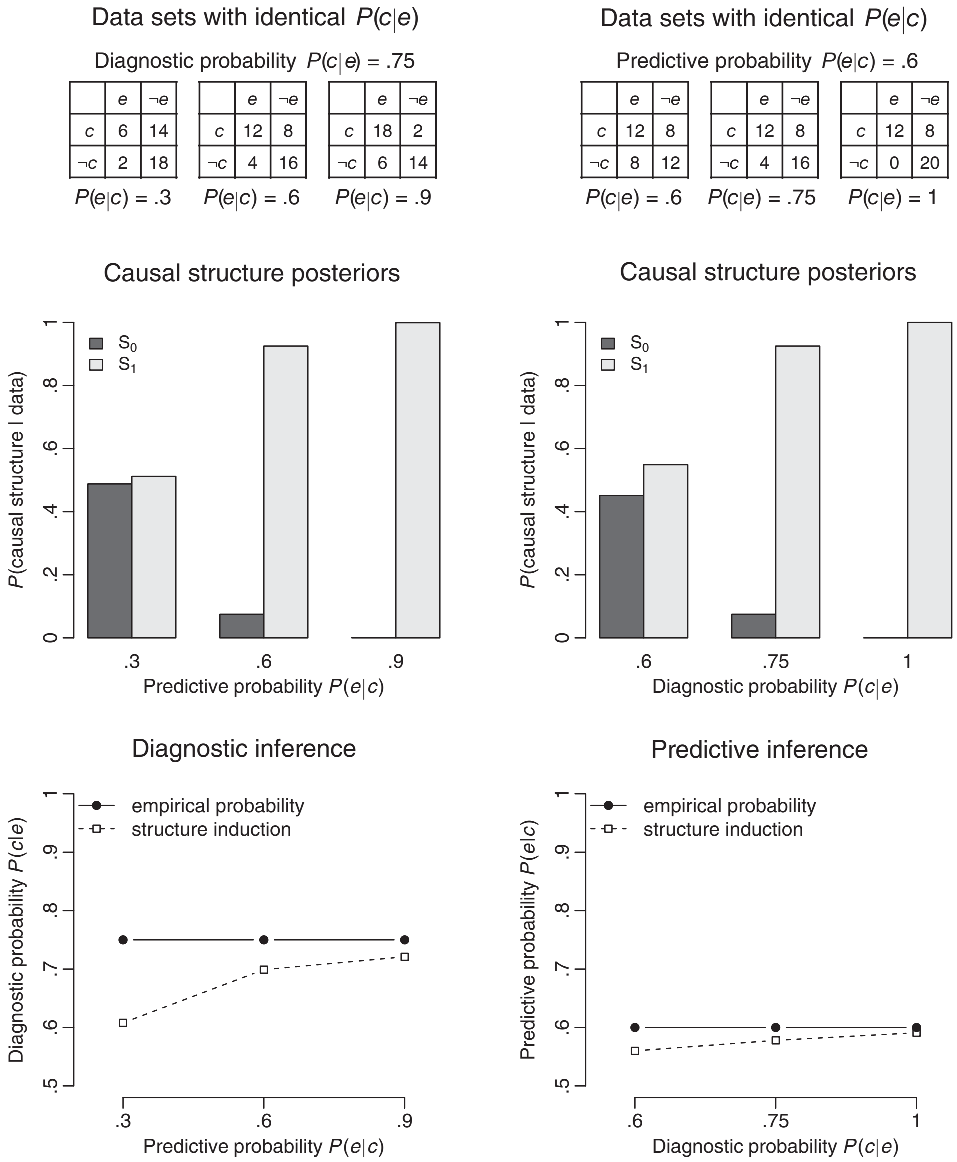
Figure 3 (bottom left) also shows how the discrepancy becomes weaker when the predictive probability increases to $P(e|c) = .6$ and $P(e|c) = .9$. For these data, graph $S_1$ is the most likely generating causal structure. As a consequence, the diagnostic probabilities derived from the structure induction model approach the empirical probability of the cause given its effect, generating an upward trend with increasing $P(e|c)$.

In sum, although the empirical diagnostic probability is identical in all three data sets, the diagnostic probabilities derived from the structure induction model systematically deviate from the probability of cause given effect in the sample. This discrepancy occurs because the alternative causal structures influence the overall estimate of the diagnostic probability in proportion to their posterior probabilities. Typically, the higher the posterior probability of $S_0$, the lower the resulting overall estimate of the diagnostic probability (for details, see Appendix A). According to $S_0$, $C$ and $E$ are independent events; therefore, observing the presence of $E$ does not increase the probability of $C$ (i.e., $P(c|e) = P(c)$). More intuitively, if the observed data provide only weak evidence for a causal link from $C$ to $E$, one is less sure about the diagnostic evidence provided by $E$ and adjusts the diagnostic judgment regarding $C$ accordingly toward the base rate of the target cause.

## Asymmetries Between Diagnostic and Predictive Causal Inferences

Although the primary focus of this article is on diagnostic reasoning, predictive inferences from cause to effect can also be modeled within the structure induction model (see also the General Discussion). In this case, an estimate of the predictive probability, $P(e|c)$, is derived under each of the two structures, which are then integrated out to obtain a single estimate (see Appendix A). Interestingly, whereas the structure induction model predicts that diagnostic judgments should be affected by the predictive probability and the strength of the target cause, in the situations considered here predictive judgments should predominantly be a function of the empirical predictive probability $P(e|c)$, irrespective of the diagnostic probability $P(c|e)$.

The right column of Figure 3 provides an example of the asymmetry between diagnostic and predictive inferences. In all three data sets, the predictive probability of effect given cause is



|  | e | ¬e |
|---|---|---|
| c | 6 | 14 |
| ¬c | 2 | 18 |

$P(e|c) = .3$

|  | e | ¬e |
|---|---|---|
| c | 12 | 8 |
| ¬c | 4 | 16 |

$P(e|c) = .6$

|  | e | ¬e |
|---|---|---|
| c | 18 | 2 |
| ¬c | 6 | 14 |

$P(e|c) = .9$

|  | e | ¬e |
|---|---|---|
| c | 12 | 8 |
| ¬c | 8 | 12 |

$P(c|e) = .6$

|  | e | ¬e |
|---|---|---|
| c | 12 | 8 |
| ¬c | 4 | 16 |

$P(c|e) = .75$

|  | e | ¬e |
|---|---|---|
| c | 12 | 8 |
| ¬c | 0 | 20 |

$P(c|e) = 1$

*Figure 3.* Diagnostic and predictive reasoning in the structure induction model. The left column shows an example of diagnostic inferences for three data sets with identical probability of cause given effect, $P(c|e)$ = .75, but different predictive probabilities, $P(e|c) = \{.3, .6, .9\}$ (top left). The data sets entail different posterior probabilities of structures $S_0$ and $S_1$ (middle left); therefore, the diagnostic probabilities derived from the structure induction model differ from the empirical probabilities when integrating out the causal structures (bottom left). The right column gives an example of predictive inferences for three data sets with an identical predictive probability, $P(e|c)$ = .6, but different diagnostic probabilities, $P(c|e) = \{.6, .75, 1\}$ (top right). Although the structures' posteriors also vary across these data sets (middle right), the predictive probabilities derived from the structure induction model only weakly differ from the predictions of the simple Bayes model. All estimates were derived using uniform priors over the causal structures and their parameters.

$P(e|c)$ = .6, whereas the diagnostic probability, $P(c|e)$, takes the values .6, .75, and 1, respectively (see Figure 3, top right). Although the posterior probability of the causal structures $S_0$ and $S_1$ also varies across the three data sets (see Figure 3, middle right), this variation has little influence on the aggregate estimate of $P(e|c)$ (bottom right). The reason for the asymmetry between predictive and diagnostic inferences is that under structure $S_0$ (which drives the deviation in the case of diagnostic inferences), the estimated value of $w_a$ is larger than under $S_1$ because all occurrences of the effect must be necessarily attributed to the influence of the background cause. Thus, whereas a higher posterior probability of $S_0$ entails a lower diagnostic probability, only a weak effect is implied for estimates of $P(e|c)$ when integrating out the causal structures (see Table 1).

Table 1
*Data Sets and Model Predictions for Experiments 1 and 2*

| Data | | | | Conditional probabilities | | Empirical power PC parameters (MLE) | | | Model predictions for diagnostic probability $P(c \mid e)$ | | | | Model predictions for causal responsibility $P(c \rightarrow e \mid e)$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N(c, e)$ | $N(c, \neg e)$ | $N(\neg c, e)$ | $N(\neg c, \neg e)$ | Diagnostic $P(c \mid e)$ | Predictive $P(e \mid c)$ | Base rate ($b_c$) | Causal strength $C$ ($w_c$) | Background cause $A$ ($w_a$) | Simple Bayes/ Power PC (MLE) | Structure induction (uniform priors) | Bayesian power PC (uniform priors) | Bayesian power PC (SS priors) | Causal attribution (MLE) | Bayesian causal attribution (uniform priors) |
| 6 | 14 | 4 | 16 | .60 | .30 | .50 | .125 | .20 | .60 | .536 | .627 | .626 | .250 | .315 |
| 12 | 8 | 8 | 12 | .60 | .60 | .50 | .333 | .40 | .60 | .554 | .599 | .583 | .333 | .322 |
| 18 | 2 | 12 | 8 | .60 | .90 | .50 | .750 | .60 | .60 | .576 | .585 | .578 | .500 | .428 |
| 6 | 14 | 2 | 18 | .75 | .30 | .50 | .222 | .10 | .75 | .608 | .712 | .740 | .556 | .481 |
| 12 | 8 | 4 | 16 | .75 | .60 | .50 | .500 | .20 | .75 | .699 | .715 | .763 | .625 | .554 |
| 18 | 2 | 6 | 14 | .75 | .90 | .50 | .857 | .30 | .75 | .722 | .722 | .761 | .714 | .661 |
| 6 | 14 | 0 | 20 | 1.00 | .30 | .50 | .300 | .00 | 1.00 | .837 | .872 | .903 | 1.000 | .776 |
| 12 | 8 | 0 | 20 | 1.00 | .60 | .50 | .600 | .00 | 1.00 | .925 | .925 | .946 | 1.000 | .892 |
| 18 | 2 | 0 | 20 | 1.00 | .90 | .50 | .900 | .00 | 1.00 | .948 | .948 | .959 | 1.000 | .940 |

*Note*. The four left-most columns show the data sets, with $N$ denoting the number of co-occurences of cause and effect. These are followed by the corresponding diagnostic and predictive conditional probabilities and the empirical power PC parameters. The parameters $w_c$ and $w_a$ denote the strength of target cause $C$ and background cause $A$, and $b_c$ denotes the base rate of $C$. The six right-most columns show the predictions of the different models for the diagnostic probability of cause given effect, $P(c \mid e)$, and estimates of causal responsibility, $P(c \rightarrow e \mid e)$, respectively. Predictions for the structure induction model were derived using a noisy-OR parameterization with the parameters $b_c$, $w_c$, and $w_a$ being independently set to a uniform Beta(1, 1) prior distribution and a uniform prior over structures $S_0$ and $S_1$, $P(S_0) = P(S_1) = .5$. Predictions were derived using Monte Carlo simulations with $m = 1{,}000{,}000$ samples for each parameter. Predictions for the power PC model with uniform priors and the Bayesian causal attribution model were derived analogously; predictions for the power PC model with sparse and strong (SS) priors over $w_c$ and $w_a$ were derived with $\alpha = 5$ (Lu et al., 2008). MLE = maximum likelihood estimate. See text and appendices for details.

This predicted asymmetry in inferences about predictive and diagnostic probabilities allows us to test the structure induction model against older accounts of diagnostic reasoning, such as the *conversion fallacy* (Dawes, Mirels, Gold, & Donahue, 1993), according to which people tend to confuse predictive and diagnostic probabilities. If people tend to report predictive probabilities when asked about diagnostic probabilities the asymmetries predicted by the structure induction model should not be observed in the data.

Further relevant research was presented by Fernbach and colleagues (2010, 2011; see also Fernbach & Rehder, 2013), who compared predictive and diagnostic reasoning using real-world stories. No learning data were presented in their studies. As the normative standard, they used a causal Bayes net model that for the simple case of elemental diagnostic and predictive causal reasoning corresponds to the power PC model. In their paradigm, people's diagnostic judgments showed sensitivity to alternative causes of the target effect (as predicted by the Bayes net model), but their predictive inferences revealed a neglect of alternative causes. Predictive inferences were better predicted by causal power than by predictive probability. However, these studies did not control the learning input and are therefore ill-suited for assessing the role of uncertainty in causal judgments so that we will revisit these theoretical claims in our Experiment 1, in which we collected both predictive and diagnostic judgments after having presented learning data.

## Summary

Our theoretical analyses have shown that modeling diagnostic reasoning from the perspective of causal inference can lead to identical predictions to those of a purely statistical approach, such as in the case of simple Bayes and power PC theory (and its isomorphic causal Bayes net representation). By contrast, we have developed a computational model of diagnostic reasoning that is sensitive to the uncertainty of predictive and diagnostic inferences. By considering alternative causal structures that may underlie the data, the diagnostic probabilities derived from this model systematically deviate from the empirical probability of cause given effect observed in a data sample.

The structure induction model provides a formalization of the intuition that diagnostic inferences should take into account the uncertainty about the generating causal structure and the existence of a causal link between $C$ and $E$. Are people's diagnostic judgments sensitive to this uncertainty? The goal of the following experiment was to test whether people's diagnostic inferences indeed take into account uncertainty, or whether diagnostic inferences simply reflect the objective diagnostic conditional probabilities in the learning data (as predicted by simple Bayes, power PC theory, and basic versions of causal Bayes net theory). The experiment also allows us to test whether people confuse diagnostic and predictive queries (i.e., conversion fallacy).

## Experiment 1

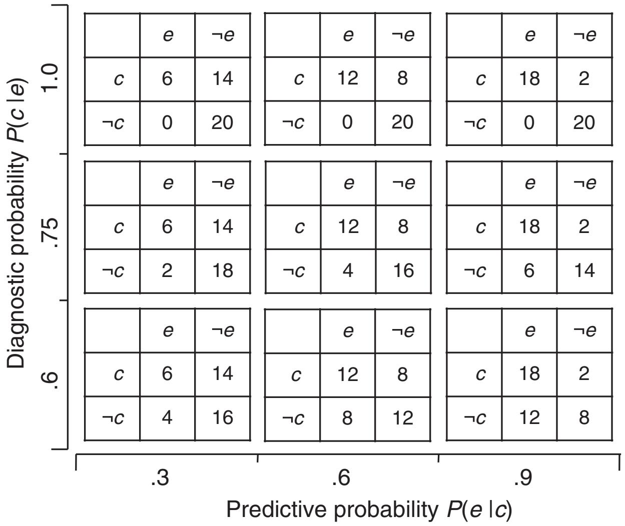
The main goal of Experiment 1 was to examine people's diagnostic judgments by systematically varying the predictive and diagnostic probability in the learning data. We generated nine data sets of sample size $N = 40$ by factorially combining three levels of the diagnostic probability $P(c|e) = \{.6, .75, 1\}$ with three levels of the predictive probability $P(e|c) = \{.3, .6, .9\}$ to cover the relevant parameter space. The resulting nine data sets are shown in Figure 4.

The second row of Figure 5 illustrates the diverging predictions of the simple Bayes and power PC account (left) and the structure induction model (right). (See Table 1 for a detailed overview of the parameter estimates and model predictions for the nine data sets.) Both accounts agree that the overall size of the diagnostic estimates should vary as a function of the level of the diagnostic probability, $P(c|e) = \{.6, .75, 1\}$ in the data. The crucial difference is that the structure induction model—but not the simple Bayes account—also predicts specific trends *within* each level of the diagnostic probability. In particular, for each level of the empirical diagnostic probability, the structure induction model entails an influence of predictive probability, with higher diagnostic judgments for higher levels of $P(e|c)$. This influence results in systematic upward trends across different data sets with the same empirical probability of cause given effect.

We asked participants to make a diagnostic judgment from effect to cause as well as a predictive judgment from cause to effect. The predictive judgments were elicited to control for the possibility that people confuse diagnostic and predictive judgments. Moreover, analyzing the predictive judgments allowed us to test if the findings of Fernbach et al. (2011) on the neglect of alternative causes in predictive inferences were replicated in our learning tasks.

## Method

**Participants and design.** Thirty-six students from the University of Göttingen (32 women; $M_{\text{age}} = 22.3$ years) participated


Diagnostic probability P(c | e)
1.0
.75
.6
.3
.6
.9
Predictive probability P(e | c)

| | e | ¬e |
|---|---|---|
| c | 6 | 14 |
| ¬c | 0 | 20 |

| | e | ¬e |
|---|---|---|
| c | 12 | 8 |
| ¬c | 0 | 20 |

| | e | ¬e |
|---|---|---|
| c | 18 | 2 |
| ¬c | 0 | 20 |

| | e | ¬e |
|---|---|---|
| c | 6 | 14 |
| ¬c | 2 | 18 |

| | e | ¬e |
|---|---|---|
| c | 12 | 8 |
| ¬c | 4 | 16 |

| | e | ¬e |
|---|---|---|
| c | 18 | 2 |
| ¬c | 6 | 14 |

| | e | ¬e |
|---|---|---|
| c | 6 | 14 |
| ¬c | 4 | 16 |

| | e | ¬e |
|---|---|---|
| c | 12 | 8 |
| ¬c | 8 | 12 |

| | e | ¬e |
|---|---|---|
| c | 18 | 2 |
| ¬c | 12 | 8 |

*Figure 4.* Learning data in Experiments 1 and 2. Nine data sets of sample size $N = 40$ were created by factorially combining three levels of the diagnostic probability $P(c|e) = \{.6 \text{ vs. } .75 \text{ vs. } 1.0\}$ with three levels of the predictive probability $P(e|c) = \{.3 \text{ vs. } .6 \text{ vs. } .9\}$.

for course credit or were paid €5. The factors level of diagnostic probability $P(c|e)$ (.6 vs. .75 vs. 1), level of predictive probability $P(e|c)$ (.3 vs. .6 vs. .9), and type of causal judgment (predictive vs. diagnostic) were varied within subject, yielding a 3 × 3 × 2 within-subject design.

**Materials and procedure.** We used a medical scenario in which physicians investigated how novel diseases causally relate to the presence of certain substances found in the blood of patients. Participants were informed that they would receive information regarding the cause event (i.e., disease present or absent) and the effect event (i.e., substance present or absent) for a sample of patients. They were also informed that every outcome would be possible: It could be that the disease always leads to the presence of the substance, that the disease never generates the substance, or that the disease is probabilistically related to the presence of the substance.

Then participants were familiarized with the stimuli (see right panel of Figure 1) and told that they would be asked to make two judgments after being presented with the data. One question would require them to make a *diagnostic inference* from the presence of the effect (substance) to its cause (disease). The other question would require them to make a *predictive inference* from the presence of the cause (disease) to the effect (substance).

Subsequent to reading the instructions, participants received the learning data. Each data set was presented on a sheet of paper showing 40 individual cases organized in a table with five columns and eight rows. Each case referred to a patient who had been tested for the presence of the disease and the presence of the substance and was depicted by the corresponding symbol combination (see right panel of Figure 1). For each disease–substance combination, nine random arrangements of the cases were created and used for the experiment. Participants could inspect the data set for as long as they wanted and were then presented with the two test questions. The data were removed before the test questions were presented and could not be inspected again. The order of questions was counterbalanced across participants.

The *diagnostic question* asked for an inference from effect to cause: "How certain are you that a novel patient who has the substance [Rothan] in his blood has been infected with [Midosis]?" The rating scale ranged from 0 (*I am absolutely certain that the patient does not have the disease*) to 7 (*I am absolutely certain that the patient does have the disease*). The *predictive question* asked for an inference from cause to effect: "How certain are you that a novel patient who has been infected with [Midosis] has the substance [Rothan] in his blood?" Estimates were given on a rating scale ranging from 0 (*I am absolutely certain that the patient does not have the substance in his blood*) to 7 (*I am absolutely certain that the patient does have the substance in his blood*). After they answered the two questions, participants proceeded to the next disease–substance combination, with each scenario using different fictitious labels for cause (disease) and effect (substance) (e.g., Midosis → Rothan). The order of the nine disease–substance combinations was counterbalanced across subjects.

## Results

Figure 5 (top left) shows participants' mean diagnostic judgments for the nine data sets. Not surprisingly, participants' diagnostic judgments varied overall as a function of the level of the diagnostic probability in the learning data (i.e., higher judgments for higher levels of $P(c|e)$). The crucial finding is that *within* a given level of $P(c|e)$ participants' estimates increased when the *predictive* relation between cause and effect became stronger. For instance, for the data sets in which the diagnostic probability $P(c|e)$ was fixed to .75 but the strength of the predictive relation $P(e|c)$ varied (.3 vs. .6 vs. .9), the diagnostic judgments show an upward trend when the predictive relation became stronger, increasing from 3.7 to 4.3 to 5. A similar trend was obtained for data sets in which $P(c|e)$ was fixed to .6. The influence of the predictive relation on participants' diagnostic judgments when $P(c|e) = 1$ is not fully clear; here, only a weak influence of the predictive probability was observed (see Experiment 2).

To validate the observed trends, we conducted an analysis of variance (ANOVA) with level of diagnostic probability and level of predictive probability as within-subject factors (see below for a quantitative model comparison). Participants' responses to the diagnostic inference questions were influenced not only by the diagnostic probability, $F(2, 70) = 74.0$, $p < .001$, $\eta^2 = .68$, but also by the strength of the *predictive* relation between cause and effect, $F(2, 70) = 12.8$, $p < .001$, $\eta^2 = .27$. The higher the predictive probability, the higher was the diagnostic judgment, even when the diagnostic probability was held constant. The analysis also revealed a weak interaction between the two factors, $F(4, 140) = 2.7$, $p = .04$, $\eta^2 = .07$, resulting from the weaker trend for the data sets in which $P(c|e) = 1$. Taken together, these findings are at variance with the simple Bayes and power PC accounts, which predict no influence of predictive probability on diagnostic judgments.[1]

Could the observed trends result from a confusion of predictive and diagnostic probabilities (conversion fallacy)? If that was the case, a similar pattern should be observed for participants' predictive inferences, namely, an influence of diagnostic probability on predictive judgments when $P(e|c)$ is fixed to a specific level. This

[1] An inspection of Figure 5 (top left) shows that the diagnostic judgments for $P(c|e) = .6$ and $P(c|e) = .75$ are fairly close to each other, in terms of their absolute values. To make sure that the influence of level of diagnostic probability on diagnostic judgments is not solely driven by the scenarios in which $P(c|e) = 1$, we conducted an ANOVA comparing $P(c|e) = .6$ with $P(c|e) = .75$. In line with the overall ANOVA, diagnostic judgments were influenced not only by the level of the diagnostic probability, $F(1, 70) = 13.2$, $p < .001$, $\eta^2 = .27$, but also by the level of the predictive probability, $F(2, 70) = 15.0$, $p < .0001$, $\eta^2 = .30$ (the interaction was not significant, $p = .18$). An ANOVA comparing $P(c|e) = .75$ with $P(c|e) = 1$ yielded a similar result: Participants' diagnostic judgments varied as a function of $P(c|e)$, $F(1, 70) = 67.0$, $p < .0001$, $\eta^2 = .66$, as well as of the predictive probability $P(e|c)$, $F(2, 70) = 8.6$, $p < .0001$, $\eta^2 = .20$. Consistent with the overall effect, there was also a weak significant interaction, $F(2, 70) = 8.6$, $p < .05$, $\eta^2 = .10$.

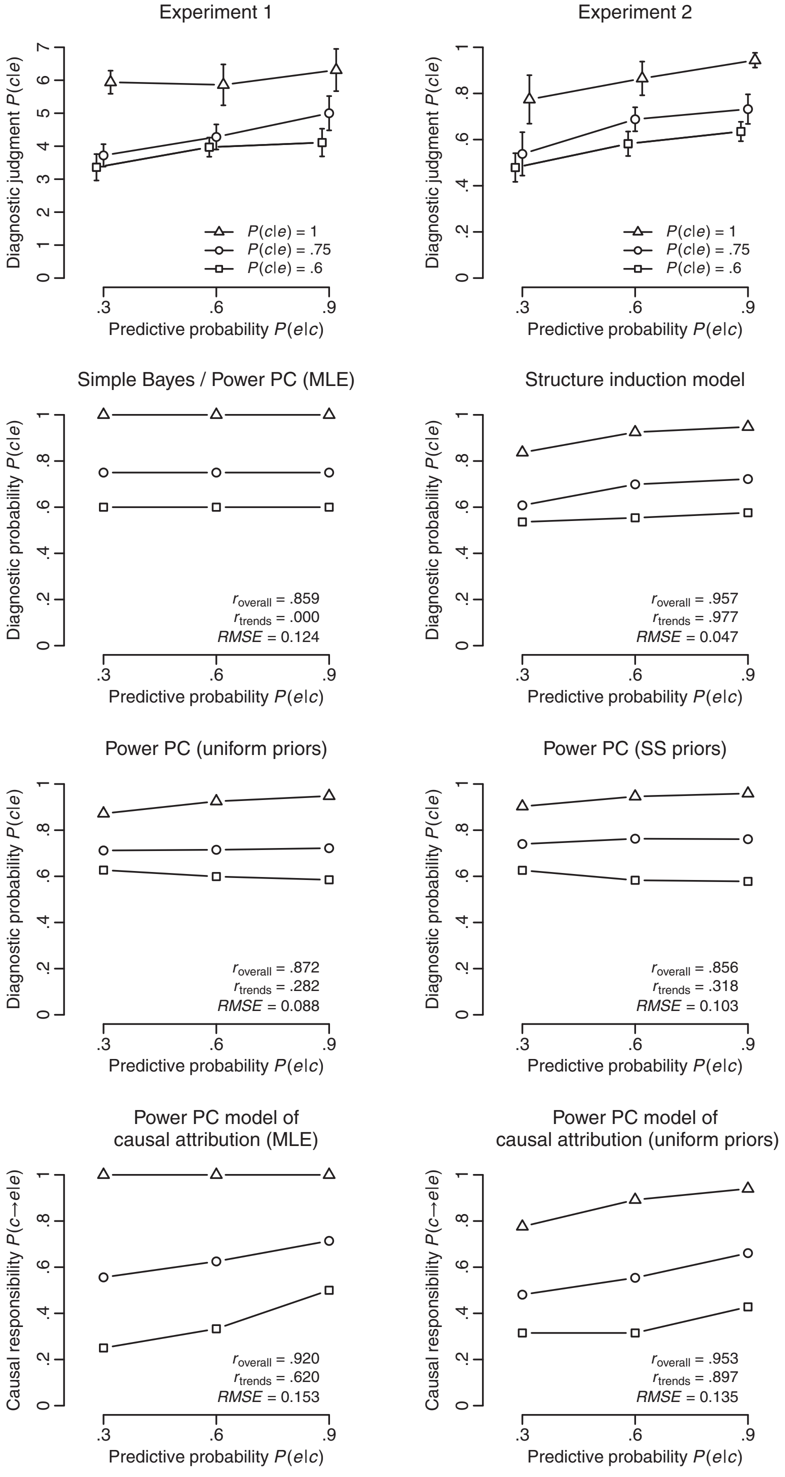


*Figure 5.* Diagnostic judgments ($M$ ± 95% CI) from effect to cause in Experiments 1 and 2 and model predictions (see also Table 1). In Experiment 1, judgments were given on a scale from 0 to 7; in Experiment 2, judgments were given on a scale from 0 to 100. The model fits refer to diagnostic judgments of Experiment 2 (see Table 2 for model fits for Experiment 1). The notation $r_{\text{overall}}$ denotes the overall correlation between a model's predictions and the mean human judgments; $r_{\text{trends}}$ denotes the mean of the correlations computed separately for each of the three levels of the diagnostic probability $P(c|e)$, a measure indicating how well the observed upward trends within each level are accounted for. MLE = maximum likelihood estimate; RMSE = root-mean-square error of a model's predictions and the mean human judgments; SS = sparse and strong.

test is also interesting as the structure induction model predicts no such influence; that is, participants' predictive judgments should only vary as a function of the observed $P(e|c)$ in the data, irrespective of the level of the diagnostic probability $P(c|e)$.

The results indicate that the predictive judgments were sensitive to the observed probability of effect given cause but unaffected by the diagnostic probability (see Figure 6). This observation is supported by an ANOVA with level of predictive probability and level of diagnostic probability as within-subject factors, which revealed a main effect of predictive probability, $F(2, 70) = 100.7$, $p < .001$, $\eta^2 = .74$, but no effect of diagnostic probability, $F(2, 70) = 1.3$, $p = .28$, $\eta^2 = .04$, and no interaction ($F < 1$). These results refute the possible hypothesis that the observed pattern of diagnostic judgments merely resulted from a confusion of diagnostic and predictive probabilities.

### Discussion

The observed diagnostic judgments are at variance with the predictions of simple Bayes and power PC (and related causal Bayes net) theories. These accounts predict that diagnostic inferences should solely reflect the size of the diagnostic probability, which in our study was kept constant across different levels of predictive probability. However, the upward trends in the diagnostic conditions are consistent with the predictions of the structure induction model, which takes into account uncertainty about the presence of a causal relation as a mitigating factor. The analogous analysis of the predictive judgments also refutes other theories, such as the conversion fallacy. Participants' diagnostic inferences were overall sensitive to the diagnostic probability across constant levels of predictive probability. Moreover, we generally obtained clear asymmetries between predictive and diagnostic judgments, which speaks against the idea that people do not distinguish appropriately between these two inferences. Similar results have been obtained in other experimental studies, demonstrating the robustness of this finding under a variety of conditions (see Experiments 2 and 3 in Meder, Mayrhofer, & Waldmann, 2009).

These results are also interesting in light of the findings of Fernbach and colleagues (2011), who reported that people's predictive inferences resembled estimates of causal power rather than conditional probability. In the data sets used here, the causal power of the target cause, $w_c$, increases within a given level of the predictive probability. For instance, for the three data sets in which $P(e|c) = .6$, the power PC estimates of causal strength, $w_c$, are .33, .5, and .6 (see Table 1). Despite this strong variation, participants' predictive inferences did not vary across these data sets. As the predictive probability was constant across the different causal strength levels due to the variation of the strength of the alternative background cause, $w_a$, these findings indicate that participants were able to appropriately take into account alternative causes and distinguish between conditional probability and causal power, possibly because learning data were available.

## Alternative Causal Inference Models of Diagnostic Reasoning

The results of Experiment 1 are at variance with the simple Bayes account, power PC theory, and basic causal Bayes net models, but are consistent with our structure induction model. Are there alternative approaches predicting similar trends to those of the structure induction model? We examined four further computational models: two Bayesian variants of the power PC model (Lu et al., 2008) and two variants of a causal attribution model (Cheng & Novick, 2005; Holyoak et al., 2010). The first two models are Bayesian interpretations of power PC theory that incorporate the notion of parameter uncertainty. The latter type of model assumes that in diagnostic reasoning people do not aim to estimate the probability of a cause given an effect but attempt to estimate a conceptually different quantity, causal responsibility. Although these models have not yet been directly tested empirically as models of diagnostic reasoning (but see Holyoak et al., 2010, for using such a model within a theory of analogical inference), theoretically they could provide an account of how people generally make diagnostic inferences when asked to infer the probability of cause given effect.

In the following sections, we will first describe these alternative models in the context of the key question of whether they predict qualitatively similar trends to those of the structure induction model. A quantitative model comparison (including the data of both Experiments 1 and 2) will be presented after Experiment 2.

### Bayesian Variants of the Power PC Model of Diagnostic Reasoning: Uniform Versus Sparse and Strong Priors

In its original formulation (Cheng, 1997), the power PC model uses maximum likelihood point estimates to parameterize causal structure $S_1$ (see Figure 2). An extension of the power PC model incorporates parameter uncertainty by using distributions over parameters (Holyoak et al., 2010; Lu et al., 2008).

To test the influence of parameter uncertainty with respect to diagnostic inferences, we implemented two Bayesian versions of the power PC model, differing in the prior distributions of the causal structure's parameters. Both models operate on a single causal structure, $S_1$, which is the default structure of power PC theory. For the first variant, we derived model predictions using uniform priors over parameters $b_c$, $w_c$, and $w_a$. This choice facilitates the comparison with the structure induction model, which also uses uniform priors. Using flat priors also conforms to the instructions, which informed participants that the causal relation between cause and effect could range from zero to perfect.

We also derived diagnostic probabilities using the *sparse and strong* (SS) prior suggested by Lu and colleagues (2008). Applied to structure $S_1$, this constitutes the *SS power model*. This account is based on the idea that causal learning and inference are guided by general systematic assumptions about the structure of the (causal) environment, which entails a preference for fewer ("sparse") and stronger causes.[2] The SS prior is defined as a joint prior distribution over parameters $w_a$ and $w_c$ of structure $S_1$, using exponential functions (see Appendix B for details). Following Lu et al. (2008), we derived model predictions for the strength-estimate version of the model by setting the free parameter $\alpha = 5$

[2] Note that the term "sparse" does not refer to the assumption that causes are rare in the sense that they have a low base rate, which would refer to a nonuniform prior over the base rate parameter, $b_c$. Rather, the claim is that people prefer causal parsimony in the sense of the assumption that $E$ is caused either by $C$ or by the alternative background cause $A$.

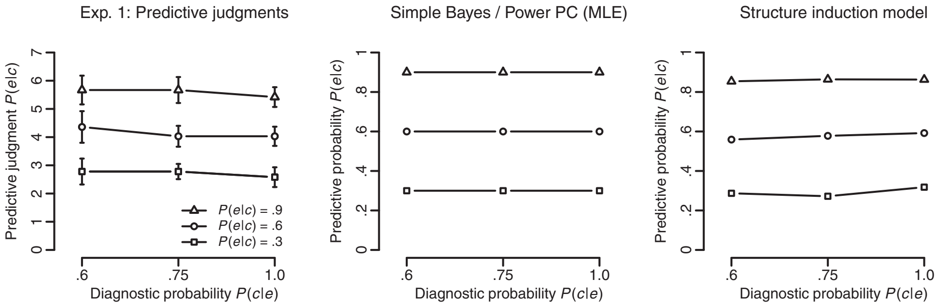


*Figure 6.* Predictive judgments ($M$ ± 95% CI) from cause to effect in Experiment 1 and model predictions. Judgments were given on a scale from 0 to 7. Exp. = experiment; MLE = maximum likelihood estimate.

(see Lu et al., 2008, for details).[3] The two peaks of this joint prior are at $w_c = 1$ and $w_a = 0$ and, conversely, $w_c = 0$ and $w_a = 1$. The key question here is whether such a priori assumptions about the causal strength parameters of structure $S_1$ would also account for our empirical findings.

The third row of Figure 5 shows the predictions of the two Bayesian variants of power PC theory for the nine data sets used in Experiment 1. The left panel shows the diagnostic probabilities derived from the power PC model with uniform priors (see also Table 1). For the three data sets in which $P(c|e) = .6$, this account entails a weak *downward* trend when the predictive probability, $P(e|c)$, increases, contrary to what we observed in Experiment 1. For the intermediate level of $P(c|e) = .75$, no influence of predictive probability is predicted, which is also inconsistent with our empirical findings. The only pattern that is qualitatively accounted for are the data sets in which $P(c|e) = 1$; here, the model entails an upward trend. Taken together, the predictions of power PC theory using uniform priors are inconsistent for two of the three levels of the diagnostic probability considered here (see below for quantitative model fits).

The diagnostic probabilities derived from the SS power model are similarly inconsistent with the results of Experiment 1. Again, when $P(c|e) = .6$, this account predicts a *downward* trend, and no influence of predictive probability is entailed when $P(c|e) = .75$. For both scenarios, participants' judgments showed an upward trend when $P(e|c)$ increases. Similar to the power model with uniform priors, the SS power model does predict an upward trend when $P(c|e) = 1$, which is consistent with the empirical findings.

Taken together, the predictions of both Bayesian variants of the power PC model are to a large extent inconsistent with subjects' diagnostic inferences, regardless of whether uniform priors are used or an SS prior that incorporates generic assumptions about causal strength.[4]

## Diagnostic Reasoning as Causal Attribution

A different theoretical approach that might be applied to our task are models of *causal attribution*. Typically, diagnostic inferences are assumed to provide an estimate of the probability of cause given effect, $P(c|e)$. However, a causal inference framework also allows for modeling other types of diagnostic queries, such as estimates of how likely it is that the observed effect was indeed produced by candidate cause $C$. Reinterpreting the test question this way means that responses should be modeled by measures of *causal responsibility*, rather than diagnostic conditional probability (Cheng & Novick, 2005; Holyoak et al., 2010). As with the Bayesian variants of power PC theory, we first present the alternative models of causal attribution in the context of our task. A quantitative test of the appropriateness of the models for our findings is presented after Experiment 2.

Let $c \rightarrow e$ denote that effect $E$ is produced by cause $C$. Then, the question of whether the occurrence of effect $E$ can be attributed to the occurrence of $C$ translates into determining the conditional probability $P(c \rightarrow e|e)$, which is different from the diagnostic probability $P(c|e)$. When $E$ can be independently produced by $C$, by $A$, or by both $C$ and $A$, these possibilities entail that when $C$ is a probabilistic cause of $E$ (i.e., $w_c < 1$), there are always some instances in which $C$ and $E$ have co-occurred, but $E$ is in fact produced by $A$. By "partialing out" the influence of $A$, we can derive $P(c \rightarrow e|e)$, the probability that $C$ caused $E$ given the occurrence of the effect (for details, see Appendix C and Cheng & Novick, 2005).

We consider two variants of a causal attribution model. In the model's original form, inferences were modeled within the standard maximum likelihood power PC theory framework, which uses the default common-effect structure $S_1$ and maximum likelihood estimates (Cheng & Novick, 2005). Holyoak and colleagues (2010; see also Lu et al., 2008) extended this approach to include parameter uncertainty, using distributions over parameter estimates (i.e., using the power PC model with uniform priors, discussed above).

In general, like the structure induction account, models of causal attribution also entail different diagnostic judgments within a given level of $P(c|e)$, as the causal strength of target cause $C$ ($w_c$) and background cause $A$ ($w_a$) are not invariant across these data sets. The relative size of these parameters determines the estimate

[3] We also examined the structure version of the SS power model (with $\alpha = 5$ and $\beta = 20$, as proposed by Lu et al., 2008). Since this model achieved a worse fit than all other models in the model comparisons, we do not discuss this account further.

[4] We also explored the influence of a "sufficiency" prior over the causal strength parameter $w_c$ (Mayrhofer & Waldmann, 2011; Yeung & Griffiths, 2011). This prior expresses that people may have a tendency to assume that causal relations are (quasi-)deterministic, even when the observed learning data are probabilistic (Goldvarg & Johnson-Laird, 2001; Griffiths & Tenenbaum, 2009). Regarding our task, this prior yields similar predictions to those of the power PC model with either uniform or SS priors.

of $P(c \rightarrow e \mid e)$ (see Equation C1 in Appendix C). The last row of Figure 5 shows the predictions of the two causal attribution models for the nine data sets used in Experiment 1 (see Table 1 for details). An inspection of the models' predictions shows that for the data sets in which $P(c \mid e)$ was fixed to .6 and .75, respectively, both accounts entail an upward trend when the predictive probability increases, although quantitatively the strength and absolute level of the trends vary across the models. The two models make diverging predictions for the data sets in which the cause is necessary and, therefore, $P(c \mid e) = 1$. Whereas the Bayesian variant using uniform prior distributions predicts an upward trend for these data sets, the basic maximum likelihood power PC does not, with its predictions corresponding to the simple Bayes account. There is no trend for the latter model because in these conditions the effect only occurs in the presence of the cause, and since maximum likelihood point estimates are used (i.e., $w_a = 0$), the model infers that the target cause $C$ necessarily generated $E$.

While these models of causal attribution often generate qualitatively similar trends to those of our structure induction model, a crucial difference concerns the absolute values of the derived estimates of $P(c \rightarrow e \mid e)$, which are usually lower than the estimates of $P(c \mid e)$ derived from the structure induction model. In particular, estimates of causal responsibility can be lower than the base rate of the cause, as the instances in which $C$ occurred but did not cause $E$ are partialed out (see Equation C1 in Appendix C). By contrast, the base rate of the cause, $P(c)$, provides the lower boundary for estimates derived from the structure induction model (which, for instance, happens when structure $S_0$ gains all the posterior probability mass; in this case the account predicts that $P(c \mid e) = P(c)$).

With respect to our empirical findings, the attribution models need to make the assumption that our participants generally interpreted the diagnostic test question as asking for the probability that the occurrence of $E$ was caused by $C$. Instead of trying to come up with an estimate of the probability of cause given effect, as requested, participants needed to interpret the diagnostic test question as referring to an estimate of causal responsibility.[5] While this is certainly a possibility, this explanation raises the question of how our participants interpreted the predictive query. The equivalent here would be that participants gave judgments of causal responsibility in the predictive direction as well, which simply means providing an estimate of the causal power of $C$ (i.e., $w_c$; as proposed by Fernbach et al., 2010, 2011). The finding that participants' predictive judgments were invariant against variations of causal power within a given level of the predictive probability speaks against the claim that participants conceptually misinterpreted the highly parallelized diagnostic inference questions. In sum, the causal attribution account requires the additional assumption that our diagnostic test questions were interpreted differently from intended, namely, as referring to causal responsibility rather than conditional probability, while the similar predictive questions were actually understood as intended, namely, as referring to the conditional probability of effect given cause. Nevertheless, it cannot be ruled out at this point that our participants specifically misinterpreted the diagnostic inference question and provided an estimate of causal responsibility, rather than conditional probability. We therefore addressed this possibility in Experiment 2, which was designed to allow for a quantitative comparison of the alternative computational models.

# Experiment 2

Experiment 1 demonstrated that our participants' diagnostic inferences were at variance with the simple Bayes and the power PC model using maximum likelihood point estimates (cf. Meder, Mayrhofer, & Waldmann, 2009). However, the Bayesian variants of the power PC model and the two attribution models also predict an influence of the predictive probability on diagnostic judgments. The attribution models in particular predict similar trends to those of our structure induction model and may, with some additional assumptions, therefore also account for the results of Experiment 1.

To provide stronger evidence for the structure induction model, the main goal of Experiment 2 was to conduct a quantitative model comparison to see which model explains human diagnostic judgments best. Although the structure induction model and the attribution models differ in terms of the predictions of the absolute values, the choice of a confidence scale in Experiment 1 did not allow us to test rigorously for these predicted differences. In Experiment 2, we therefore used a probability scale that allowed us to also interpret the absolute values and assess deviations from the predicted absolute values in our model fit analyses (using the root-mean-square error, RMSE).

One other difference between the structure induction model and the attribution models is the different assumptions they make about the interpretation of the test question. Since the main goal of our research is to explore diagnostic judgments (as opposed to causal responsibility assessments), we attempted to make sure in Experiment 2 that participants correctly interpreted the test question. We therefore added an instruction test phase prior to the actual experiment to clarify the requested diagnostic inference (see below).

Finally, an argument that could be raised against the findings of Experiment 1 is that we may only have found influences of predictive probability on diagnostic judgments because our procedure of requesting both predictive and diagnostic judgments suggested to participants that they should keep track of both diagnostic and predictive probability, which may have influenced their diagnostic judgments. Thus, to provide stronger evidence for our model, in Experiment 2 we requested only diagnostic judgments.

## Method

**Participants and design.** Forty-nine students from the University of Göttingen ($M_{\text{age}} = 23.1$ years; 71% female) participated as part of a series of various unrelated computer-based experiments. They either received course credit or were paid €8 per hour. We used the same nine data sets as in Experiment 1. The factors level of diagnostic probability $P(c \mid e) = \{.6 \text{ vs. } .75 \text{ vs. } 1\}$ and level of predictive probability $P(e \mid c) = \{.3 \text{ vs. } .6 \text{ vs. } .9\}$ were varied

[5] Note that estimates of causal responsibility can also be derived from the structure induction model. Thus, even if participants aimed to provide an estimate of causal attribution, this does not exclude that they are sensitive to causal structure uncertainty. The basic rationale of such a model is the same as with deriving conditional probabilities; that is, estimates of $P(c \rightarrow e \mid e)$ are derived separately under structures $S_0$ and $S_1$, which are then integrated out using Bayesian model averaging (see Appendix C). Since this model yields similar predictions to those of the two other attribution models, we do not discuss the model in detail here.

within subject, yielding a 3 × 3 within-subject design with diagnostic judgments as dependent measure.

**Materials and procedure.** We used the same materials as in Experiment 1 in a computer-based experiment. After reading the instructions, participants were informed that they would be asked to make a diagnostic judgment after being presented with each data set. The judgment would require them to estimate the probability that a novel patient has the disease given that she has the substance in her blood.

Before proceeding to the actual experiment, participants were requested to answer various multiple-choice questions to ensure that they correctly understood the symbols used, the experimental procedure, and the task. One of the multiple-choice questions was specifically designed to minimize possible confusion regarding the requested diagnostic judgment; the goal was to ensure that participants would attempt to estimate the diagnostic probability of cause given effect. The diagnostic test question requested subjects to estimate how probable it is that a novel patient (from the same population from which the data sample was obtained) has the disease given the substance: "Imagine that you examine another person and notice that the person has the substance in her blood. How probable is it that this person has the disease?"

To make sure that subjects understood the test question, the corresponding multiple-choice question had four answer options: (a) estimate how probable it is that the disease generated the substance, (b) estimate how probable it is that causes other than the target disease generated the substance, (c) estimate how probable it is that only the disease and no other causes generated the substance, and (d) estimate how probable it is that the person has the disease. The first three options correspond to different types of attribution questions (see Cheng & Novick, 2005), whereas the last option corresponds to assessing the conditional probability of cause given effect. This option, which we counted as the correct response, corresponded to the instructions participants had read previously.

If any of the questions in the multiple-choice test were answered incorrectly, participants were asked to re-read the instructions and take the test again until they had committed zero errors or had gone through the instructions three times.[6] After the comprehension test, participants were presented with the 40 cases of each data set, randomly arranged on the computer screen in an eight-columns-by-five-rows grid. They could inspect each data set for as long as they wanted (for a minimum of 30 s). The data were removed before participants made the diagnostic judgment. For each data set, participants estimated the probability that a novel patient with the substance in his blood has the disease. The 11-point probability rating scale ranged from 0 (*The patient definitely does not have the disease*) to 100 (*The patient definitely does have the disease*). After making the diagnostic judgment, participants proceeded to the next disease–substance combination. The order of the data sets was randomized.

### Results

Of the 49 participants, 15 failed to answer all the multiple-choice questions of the instruction test correctly after three iterations. These participants were excluded from the following analyses, leaving 34 valid participants ($M_{\text{age}}$ = 22.9 years, 76% female).

Figure 5 (top right) shows the results of Experiment 2. As in Experiment 1, within each level of the diagnostic probability, participants' judgments varied systematically as a function of the strength of the predictive probability, resulting in higher diagnostic judgments when $P(e|c)$ increased. The results also disambiguate one finding from Experiment 1 in which only a weak upward trend was observed when $P(c|e) = 1$; this time, a strong upward trend was observed for this level of the diagnostic probability, too.

An ANOVA with level of diagnostic probability and level of predictive probability as within-subject factors revealed a main effect of level of diagnostic probability, $F(2, 66) = 234.4$, $p < .0001$, $\eta^2 = .64$, and a main effect of predictive probability, $F(2, 66) = 16.7$, $p < .0001$, $\eta^2 = .34$ (the interaction was not significant, $F < 1$).[7] Thus, as in Experiment 1, we observed that people's diagnostic judgments varied as a function of the predictive probability of effect given cause across data sets in which the diagnostic probability of cause given effect was held constant.[8]

## Model Comparison

How well can the different computational models of diagnostic inference account for the empirical data? A successful model should account for three aspects of the data: the overall influence of the three different levels of $P(c|e)$, the upward trends within each level of the diagnostic probability as a function of $P(e|c)$, and the absolute values of participants' judgments.

We therefore report three types of fit measures, each of which specifically addresses one aspect of the data. First, we evaluated the alternative models with respect to how well they capture the overall trends in the data, based on the overall correlation between the mean human judgments and the models' predictions (henceforth denoted $r_{\text{overall}}$). The informativeness of this measure, however, is somewhat limited as it reflects both the models' capacity to account for the variation in diagnostic judgments resulting from manipulating the level of the diagnostic probability across the data sets (i.e., fixing $P(c|e)$ to values of .6, .75, and 1, respectively) and the variation within each level of the diagnostic probability (i.e.,

[6] Participants who failed to answer all questions correctly in the third iteration were allowed to proceed with the experiment but were excluded from our analyses.

[7] Given that the diagnostic judgments for $P(c|e) = .6$ and $P(c|e) = .75$ are fairly close to each other (see Figure 5, top right) we conducted analogous ANOVAs including only these two levels of $P(c|e)$. This analysis yielded a main effect of diagnostic probability, $F(1, 70) = 14.7$, $p < .001$, $\eta^2 = .31$, and a main effect of level of $P(e|c)$, $F(2, 70) = 15.7$, $p < .0001$, $\eta^2 = .32$ (the interaction was not significant, $p = .66$). The comparison of $P(c|e) = .75$ with $P(c|e) = 1$ gave a similar result: Diagnostic judgments were influenced by the level of $P(c|e)$, $F(1, 70) = 33.8$, $p < .0001$, $\eta^2 = .51$, but also by the predictive probability $P(e|c)$, $F(2, 70) = 14.9$, $p < .0001$, $\eta^2 = .31$; again, the interaction was not significant ($p = .65$).

[8] One reviewer raised the question of interindividual differences or possible aggregation artifacts. To address this concern, we ran cluster analyses for both Experiments 1 and 2. If, for instance, only a certain proportion of subjects generated upward trends but others gave identical judgments for a given level of the diagnostic probability, this analysis should yield two clusters of subjects. However, these analyses did not reveal particular clusters among participants. As a further check, we generated P-P plots to examine the distribution of errors (deviations from the means) on a subject-wise basis. If there were subgroups with specific inference patterns, these plots should reveal visual clusters of participants. The plots also did not indicate the existence of specific subgroups.

the upward trends resulting from varying the predictive probability, $P(e|c)$, when the diagnostic probability, $P(c|e)$, is fixed). For instance, although the simple Bayes model entails that there should be no influence of predictive probability when the diagnostic probability is fixed, it will nevertheless account for the variation resulting from the three different levels of the diagnostic probability across the data sets.

To specifically test how well the different models capture the observed upward trends within each level of $P(c|e)$, we computed the correlation of the models' predictions separately for each of the three levels of the diagnostic probability and report the mean of these three correlations (henceforth denoted $r_{\text{trends}}$).[9]

Finally, and most importantly, we evaluated the models with respect to their capacity to account for the absolute judgments. We therefore report the RMSE for each model, which provides a measure of the absolute deviation between the models' predictions and participants' diagnostic judgments (see Lu et al.'s, 2008, model comparison for an analogous analysis). This analysis is particularly informative regarding the comparison of the structure induction model and the attribution models, as the latter predict similar upward trends but very different absolute judgments. For instance, for the data sets in which the diagnostic probability was fixed to .6, the two attribution models predict diagnostic judgments below or equal to the base rate of the cause, $P(c) = .5$, whereas the diagnostic probabilities derived from the structure induction can never be smaller than the base rate of the cause (see Figure 4 and Table 1).

## Model Comparison: Experiment 2

We first report the model comparison for Experiment 2 because in this study we used a probability scale that allows for the comparison of participants' absolute judgments with the models' predictions. The results of the model comparison for Experiment 2 are shown in Figure 5, separately for each model (in the bottom right corner of each graph; see also Table 2). The structure induction model produced a better fit of the human data than the other models, on each of the three measures: It had the highest overall correlation (.957), the highest trend correlation (.977), and the lowest RMSE (.047). The two Bayesian variants of power PC theory achieved a considerable overall correlation (.872 for the model with the uniform priors and .856 for the SS power model), but a comparison with the simple Bayes account shows that this measure is not particularly informative: Although simple Bayes does not predict the observed upward trends when $P(c|e)$ is fixed, it achieved a correlation of .859, as it does account for the overall variation resulting from manipulating the diagnostic probability across the data sets. The more informative $r_{\text{trends}}$ of the Bayesian variants of power PC theory were .282 and .318, respectively, which are very low compared to the structure induction model. Both models also had a poorer fit than the structure induction model in terms of RMSE, indicating that the models' predictions substantially deviated from the actual responses of our participants. These analyses refute the two Bayesian variants of Power PC theory and suggest that incorporating parameter uncertainty or generic priors is not sufficient to account for the obtained pattern of judgments.[10]

The attribution models performed better than the Bayesian variants of power PC theory but could also not reach the fits of the structure induction model. The attribution model with maximum likelihood point estimates (MLE attribution model) and its Bayesian variant with uniform priors achieved an overall correlation of .920 and .953, respectively. The MLE attribution model performed poorly in terms of its capacity to account for the specific trends, with $r_{\text{trends}} = .620$. This relatively low correlation results from the model predicting no trend when $P(c|e) = 1$, which is inconsistent with the empirical results. The Bayesian attribution model was better able to account for the specific trends with $r_{\text{trends}} = .897$ but still had a lower fit than the structure induction model.

While the correlations suggest that the attribution models can account qualitatively for the trends, their RMSE values reveal the strong discrepancy between the models' predictions and the absolute responses: Both models had a much higher RMSE than the structure induction model (0.153 and 0.135, respectively, vs. 0.047). In fact, of all the models the two attribution models had the highest RMSE. Both models predict judgments that are too low for the data sets in which $P(c|e) = .3$. Moreover, the attribution model that uses point estimates additionally fails to capture the observed trend when $P(c|e) = 1$.

In sum, the model comparison shows that the structure induction model was most successful in accounting for participants' diagnostic judgments. The model predicts the overall influence of the different levels of $P(c|e)$ across the different data sets, the upward trends within each level of the diagnostic probability as a function of $P(e|c)$, and the absolute size of participants' judgments. The Bayesian variants of power PC theory and the attribution models can each account for aspects of the data, but none was able to fully capture participants' diagnostic judgments.

## Model Comparison: Reanalysis of Experiment 1

We conducted the same model comparisons for Experiment 1. Recall that in this study diagnostic judgments had been expressed on a confidence scale from 0 to 7, which makes it difficult to directly interpret the absolute values and compare them to the models' predictions. To compute the RMSE, we therefore transformed the human judgments by dividing them by 7, the maximum of the confidence scale used. (Note that the correlations are invariant against this transformation, as they involve a linear transformation of the data.) As we did for Experiment 2, we computed $r_{\text{overall}}$, $r_{\text{trends}}$, and RMSE for participants' diagnostic judgments.

The results of this analysis were consistent with the results of the model comparison for Experiment 2, with the structure induction model achieving the highest fit on all three measures (see Table 2). As in Experiment 2, the other models captured aspects of the data but did not reach the fit of the structure induction model.

[9] If the model predicts no variation within a given level of $P(c|e)$ (i.e., simple Bayes), the correlation is not defined; in this case we set it to zero.

[10] We also examined the fits of the SS power model for a wider range of values for the α parameter. Lu et al. (2008) set $\alpha = 5$; when $\alpha = 0$, the SS power model reduces to the power PC model with uniform priors. We varied the α parameter between 1 and 10 in steps of 1. No model variant achieved the fit of the structure induction model. For instance, the highest $r_{\text{trends}}$ in Experiment 1 was .19 (as opposed to .788 for the structure induction model), which was obtained with $\alpha = 1$. In Experiment 2, the highest fit was obtained with $\alpha = 3$ ($r_{\text{trends}} = .325$, as opposed to .977 for the structure induction model). Overall, the model fits tend to get worse when α increases (which expresses growing strength of the "strong and sparse" prior).

Table 2
*Comparison of Computational Models of Diagnostic Reasoning for Experiments 1 and 2*

| Experiment | Fit measure | Simple Bayes / power PC (MLE) | Structure induction (uniform priors) | Bayesian power PC (uniform priors) | Bayesian power PC (SS priors) | Causal attribution (MLE) | Bayesian causal attribution (uniform priors) |
|---|---|---|---|---|---|---|---|
| 1 | $r_{\text{overall}}$ | .919 | **.962** | .910 | .890 | .954 | .958 |
| | $r_{\text{trends}}$ | .000 | **.788** | .203 | .111 | .619 | .762 |
| | RMSE | 0.126 | **0.055** | 0.092 | 0.110 | 0.140 | 0.121 |
| 2 | $r_{\text{overall}}$ | .859 | **.957** | .872 | .856 | .920 | .953 |
| | $r_{\text{trends}}$ | .000 | **.977** | .282 | .318 | .620 | .897 |
| | RMSE | 0.124 | **0.047** | 0.088 | 0.103 | 0.153 | 0.135 |

*Note*. Boldface indicates highest correlation and lowest root-mean-square error (RMSE), respectively, in each row. In Experiment 1, diagnostic judgments were given on a scale ranging from 0 to 7; these were transformed to a probability scale by dividing by 7. Experiment 2 used a probability scale ranging from 0 to 100 (transformed to a 0–1 scale). MLE = maximum likelihood estimate; SS priors = sparse and strong priors (Lu et al., 2008); $r_{\text{overall}}$ = overall correlation between a model's predictions and the mean human judgments; $r_{\text{trends}}$ = mean of the correlations computed separately for each of the three levels of the diagnostic probability $P(c|e)$.

For instance, the Bayesian variants of power PC theory did well in terms of the overall correlation, as they also entail that participants' diagnostic judgments should vary across the different levels of $P(c|e)$. However, the low mean trend correlations (.203 and .111, respectively) weaken these accounts, and their RMSE was also much higher than the RMSE of the structure induction model.

Conversely, the attribution models achieved a good fit in terms of $r_{\text{trends}}$, as they also predict qualitative upward trends within each level of $P(c|e)$. However, the very high RMSE compared to the structure induction model (.140 and .121, respectively, vs. .055) shows that the attribution models fail to capture the absolute levels of judgments. In summary, the model comparisons for the diagnostic judgments obtained in Experiment 1 are consistent with the model comparisons for Experiment 2, which together provide strong support for the structure induction model.

## Model Comparison: Bootstrap Analysis of Experiments 1 and 2

The model comparisons showed that the structure induction model is superior to all other models on all three fit measures. However, the fact that a model entails a higher correlation or a lower RMSE than other models in a specific experiment might be just a coincidence. To provide an idea of the reliability of the relative model fits, we ran a bootstrap analysis for both experiments. For this analysis, we drew 10,000 random bootstrap samples from the empirical data (subject-wise, with replacement) of size $n = 36$ (for Experiment 1) and $n = 34$ (for Experiment 2), respectively. For each sample we computed the models' $r_{\text{overall}}$, $r_{\text{trends}}$, and RMSE and kept a count of the winning model, that is, the model that achieved the best fit on each of the three criteria for each of the 10,000 bootstrap replications. This simulation provides a measure of replicability, that is, the expected probability that a model turns out to be the best-fitting model when replicating the experiment under the exact same conditions. (Note that none of the models have free parameters; therefore, we did not have to deal with the problem of overfitting.)

Table 3 shows the results of this analysis. Consistent with the results of the fit analyses, the structure induction model best accounted for the results on all three measures, for both experiments across the majority of the bootstrap replications. Particularly striking is the good fit in terms of RMSE: For Experiment 1, the structure induction model had a lower RMSE than all other models in 99.9% of the replications, and in Experiment 2 it had a lower RMSE in 100% of the replications. These findings provide further evidence for the descriptive validity of the structure induction model and the theoretical claim that people take into account uncertainty about the existence of a causal link in the generating causal structure.

## General Discussion

The long-standing normative benchmark for diagnostic reasoning has been the simple Bayes model, according to which diagnostic judgments are determined by the empirical probability of cause given effect in the data sample. We have argued that this norm is myopic, as it does not distinguish between the data and the causal level; it lacks the representational power to take into account alternative causal structures that may underlie the data. At the psychological level, sensitivity to the generating causal models is important because the distal goal of cognitive systems is to represent the world in terms of stable causal relations, rather than arbitrary statistical associations that may be distorted by noise (Krynski & Tenenbaum, 2007; Waldmann & Hagmayer, 2013; Waldmann et al., 2006, 2008).

Modeling diagnostic reasoning from the perspective of causal inference does not necessarily lead to different predictions from those obtained by a purely statistical approach. The standard power PC model (Cheng, 1997) and basic variants of causal Bayes nets (Fernbach et al., 2011) predict identical diagnostic probabilities to those of the simple Bayes model, which does not distinguish between observable contingencies and unobservable causal relations. However, a causal inference approach can also be constructed that is sensitive to the inherent uncertainty of data as evidence for an underlying causal structure. Our structure induction model formalizes the intuition that diagnostic reasoning should be sensitive to the question of whether the sample data warrant the existence of a causal relation between the candidate cause and the effect (Anderson, 1990; Griffiths & Tenenbaum, 2005, 2009). The key prediction of this model is that diagnostic judgments not only should vary

Table 3
*Comparison of Computational Models of Diagnostic Reasoning Based on Bootstrap Analysis of Diagnostic Judgments in Experiments 1 and 2*

| Experiment | Fit measure | Simple Bayes / power PC (MLE) | Structure induction | Bayesian power PC (uniform priors) | Bayesian power PC (SS priors) | Causal attribution (MLE) | Bayesian causal attribution (uniform priors) |
|---|---|---|---|---|---|---|---|
| 1 | $r_{\text{overall}}$ | 0.4% | **52.1%** | 0.0% | 0.0% | 29.9% | 17.6% |
| | $r_{\text{trends}}$ | 0.0% | **62.0%** | 0.2% | 0.0% | 17.9% | 19.8% |
| | RMSE | 0.0% | **99.9%** | 0.0% | 0.0% | 0.0% | 0.1% |
| 2 | $r_{\text{overall}}$ | 0.0% | **68.2%** | 0.0% | 0.0% | 0.5% | 31.3% |
| | $r_{\text{trends}}$ | 0.0% | **97.3%** | 0.0% | 0.0% | 0.0% | 2.6% |
| | RMSE | 0.0% | **100.0%** | 0.0% | 0.0% | 0.0% | 0.0% |

*Note*. The percentages indicate the proportion of times a model had the best fit on the respective criterion out of 10,000 bootstrap replications. Boldface indicates highest percentage in each row. MLE = maximum likelihood estimate; SS priors = sparse and strong priors (Lu et al., 2008); $r_{\text{overall}}$ = overall correlation between a model's predictions and the mean human judgments; $r_{\text{trends}}$ = mean of the correlations computed separately for each of the three levels of the diagnostic probability $P(c|e)$; RMSE = root-mean-square error.

as a function of the observed probability of cause given effect, but should depend on the plausibility of a causal relation from $C$ to $E$.

We tested this prediction in two experiments and observed that participants' diagnostic judgments systematically varied within a given level of $P(c|e)$, a finding that is at variance with simple Bayes and power PC theory (including corresponding causal Bayes net models) but consistent with the structure induction model. The observed asymmetry between diagnostic and predictive inferences (Experiment 1) also refutes the idea that people do not appropriately distinguish between predictive and diagnostic inferences (i.e., conversion fallacy).

## Diagnostic Probability Versus Causal Attribution

The focus of our analyses and experiments was on modeling estimation of the conditional probability of cause given effect. In this context, we also discussed models of causal attribution (Cheng & Novick, 2005), which provide an estimate of how likely it is that an observed effect was indeed produced by the target cause, $P(c \rightarrow e|e)$. We carefully instructed our subjects about the intended meaning of the test question (especially in Experiment 2), which led to results that can be better explained by the structure induction model than by causal attribution theories. However, we believe that in other contexts causal attribution theories may better account for inferences from effects to causes (e.g., Holyoak et al., 2010).

For instance, in many real-world situations of diagnostic reasoning, such as in medical settings, estimates of causal responsibility are what we are interested in (see also Gerstenberg & Lagnado, 2010). Imagine a doctor examining a patient whose symptoms indicate that she is likely to suffer from inflammation of the peritoneum, which may require medication and/or immediate surgery. The problem faced by the doctor is that there are several possible causes, including perforation of a hollow organ through previous surgery, inflammation of the appendix, and tumors. Since each of these causes requires a different treatment, in this case a diagnostic judgment should primarily be concerned with causal responsibility. For example, the mere presence of a tumor may be irrelevant at this point in time if it is not the cause of the inflammation, since immediate treatment requires determining which event is causally responsible for the production of the effect. Key questions for future research include to what extent people are indeed sensitive to the notion of causal responsibility, and whether they are capable of distinguishing between estimates of conditional probability and estimates of causal responsibility.

## Beyond the Sample

The most striking result of our experiments is that subjects' diagnostic inferences deviated from the conditional probability of the cause given effect in the observed data samples, even though the data were presented in a frequency format, and extensive instructions (in Experiment 2) were provided to make sure that subjects correctly understood the diagnostic test question. This finding seems to be at odds with the many demonstrations of accurate diagnostic judgments when given natural frequency data and test questions requesting estimates of conditional frequencies (see Gigerenzer & Hoffrage, 1995; Barbey & Sloman, 2007). Conditional frequency judgments have often been deemed superior on the grounds of the frequentist philosophical view of probability, which argues that it does not make sense to request a probability estimate for a single case (see Hacking, 2001). We have no doubt that with the right instructions subjects can be nudged to use a counting strategy that leads to a correct conditional frequency judgment. What we were interested in finding out, however, was how subjects process learning data to arrive at an inductive diagnostic inference about a novel case. The results of the experiments clearly show that our subjects had no difficulties answering questions about the probability of a disease in a single fictitious patient. According to a Bayesian perspective, probability judgments reflect degrees of beliefs, which can, as our experiments show, be influenced by observed conditional frequencies. Thus, in our view, the deviations of the diagnostic inferences from the conditional frequencies in the sample do not result from a misrepresentation of the test question. Rather, they reflect an attempt to make a diagnostic conditional probability judgment on the level of the underlying causal model. Given that the generating causal structure is unobservable, the judgment must take into account the inherent uncertainty in making inferences from the observed sample to the causal model.

Our finding that subjects are sensitive to uncertainty when making causal inferences adds to a growing body of related findings (e.g., Griffiths & Tenenbaum, 2005; Lu et al., 2008). Interestingly, there is a tension between these demonstrations of sensitivity to uncertainty and studies showing that judgmental biases can often be explained by statistical distortions in the observed sample (see Fiedler & Juslin, 2006, for an overview). According to the information sampling view, people are myopic and to a large extent unable to look beyond the sample. They often lack the necessary metacognitive knowledge to rectify biases caused by statistical distortions in the data sample (Fiedler, 2012). These results are not inconsistent with our findings, given that our task did not require any awareness of possible distortions in the sample. However, what we have shown is that, at least in the context of causal induction, people are sensitive to the fact that samples carry some degree of uncertainty. Thus, subjects are not entirely myopic; they are capable of looking beyond the sample.

## Structure Induction and Sample Size

How do variations in sample size affect the predictions of the structure induction model? In the limit when $N \rightarrow \infty$, the diagnostic probability derived from the structure induction model will converge toward the maximum likelihood estimate, yielding the same predictions as the simple Bayes account. This happens because in the long run the posterior probability of $S_1$ approximates 1 and the posterior probability of $S_0$ converges toward zero (as long as causal strength is stable and greater than zero).

However, the convergence of the structure induction model on the maximum likelihood value of $P(c|e)$ can be fairly slow, as $S_0$ can receive substantial posterior probability even for relatively large sample sizes. Figure 7 illustrates this fact with two of the data sets used in our experiments. In one data set, the diagnostic probability $P(c|e) = .75$ (see Figure 7, left column); in the other data set, $P(c|e) = .6$ (see Figure 7, right column). To show the differential influence of sample size on the predictions of the structure induction model, the contingency table (i.e., $N = 40$ cases) is multiplied by 5 and 10, respectively, yielding sample sizes of $N = 200$ and $N = 400$. This increase in sample size does not affect the maximum likelihood estimates derived from the simple Bayes and standard power PC model, which are insensitive to the uncertainty of parameter estimates (i.e., these models make identical predictions regardless of sample size). In contrast, the structure induction model makes differential predictions for the two scenarios.

When the objective diagnostic probability is relatively strong, $P(c|e) = .75$, the diagnostic probability derived from the structure induction model converges relatively fast toward the maximum likelihood estimate when sample size increases (see Figure 7, left column). The right-hand side of Figure 7 shows that this is not generally true, however. When the objective diagnostic probability is slightly weaker, $P(c|e) = .6$, the derived diagnostic probability is much less affected by the increase in sample size. Even with a sample size of $N = 400$ cases, the posterior probability of structure $S_0$ is still .33; accordingly, the diagnostic probability derived from the structure induction model differs from the maximum likelihood estimate. The reason for this lower sensitivity to sample size is that in this data set the cause is fairly weak (the maximum likelihood estimate of causal strength is $w_c = .12$). Therefore a large sample size is needed to refute the possibility that the co-occurrence between $C$ and $E$ is merely coincidental. Thus, the exact influence of sample size on the predictions of the structure induction model crucially depends on an interaction of various factors, such as the size of the sample and the strength of the causal relation (cf. Griffiths & Tenenbaum, 2005).

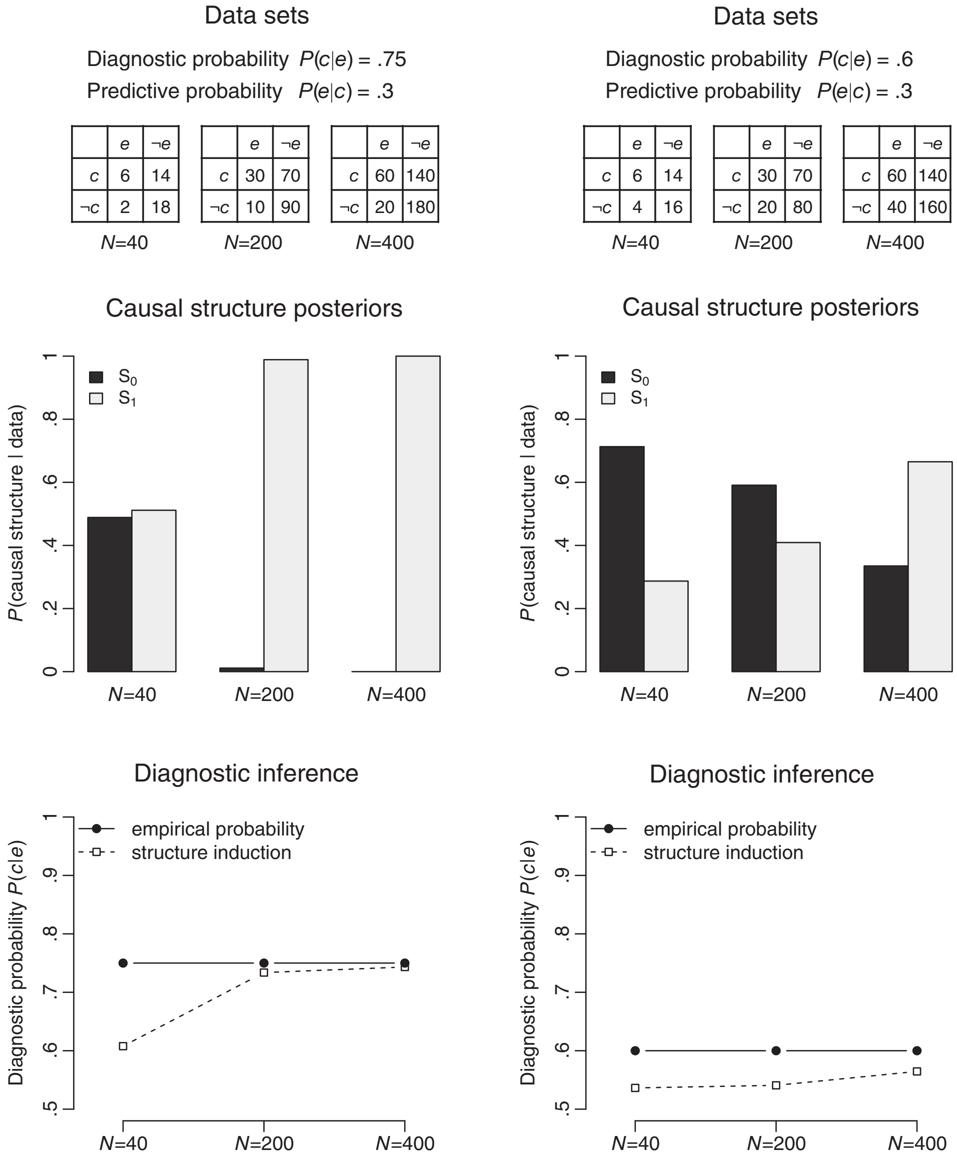
In sum, whereas models operating with maximum likelihood point estimates inferred directly from the sample data (e.g., the simple Bayes and standard power PC model) are insensitive to variations in sample size, the structure induction model is sensitive to sample size (as are Bayesian models in general), as this influences the parameter estimates and posterior probability of structure $S_0$ and $S_1$. This is consistent with research showing that people are, at least to some extent, sensitive to the size of the data set on which causal inferences are based (Griffiths & Tenenbaum, 2005; Lu et al., 2008; see also Sedlmeier & Gigerenzer, 1997).

The exact influence of sample size on people's diagnostic inference is an issue for future research. One hypothesis is that large samples substantially reduce people's uncertainty about parameter estimates and the existence of a causal relation between $C$ and $E$. In this case, people's diagnostic inferences may approximate the observed conditional probability of cause given effect (at least in situations with relatively strong causes or weak background causes). On the other hand, memory limitations may place limits on sample size sensitivity. For instance, working memory capacity may constrain the number of observations that are actually considered when making an inference (e.g., Kareev, 2000) or result in a temporal weighing of information in diagnostic reasoning (Meder & Mayrhofer, 2013). In this case, uncertainty in diagnostic reasoning would not result from limited sample data but from the bounded rationality of the reasoner.

## Computational Versus Process Models of Diagnostic Reasoning

The present work is concerned with the normative and descriptive adequacy of alternative models of elemental diagnostic reasoning. We focused on computational-level models that help us to understand *why* specific behaviors are observed by specifying the cognitive task being solved, the information involved in solving it, and the logic by which it can be solved (Anderson, 1990; Chater & Oaksford, 1999, 2008; Marr, 1982; for critical reviews, see Brighton & Gigerenzer, 2012; Jones & Love, 2011). These accounts are less interested in pinpointing the actual cognitive processes underlying the behavior and more concerned with providing an explanation of observed behavior in relation to the goals of the reasoner and the structure of the environment. In the case of diagnostic reasoning, we have observed a pattern of judgments that looks irrational from the perspective of the classical norm of diagnostic reasoning, the simple Bayes model, but can be explained by the structure induction model, which formalizes an inference strategy that takes into account uncertainty regarding the causal structure of the environment.

While a computational-level description does not necessarily preclude the possibility that the mind somehow implicitly carries out the involved computations, in the spirit of "man as intuitive statistician"



| | e | ¬e |
|---|---|---|
| c | 6 | 14 |
| ¬c | 2 | 18 |

N=40

| | e | ¬e |
|---|---|---|
| c | 30 | 70 |
| ¬c | 10 | 90 |

N=200

| | e | ¬e |
|---|---|---|
| c | 60 | 140 |
| ¬c | 20 | 180 |

N=400

| | e | ¬e |
|---|---|---|
| c | 6 | 14 |
| ¬c | 4 | 16 |

N=40

| | e | ¬e |
|---|---|---|
| c | 30 | 70 |
| ¬c | 20 | 80 |

N=200

| | e | ¬e |
|---|---|---|
| c | 60 | 140 |
| ¬c | 40 | 160 |

N=400

*Figure 7.* Differential influence of sample size on diagnostic probabilities derived from the structure induction model. Left: Three data sets with identical predictive and diagnostic probability ($P(e|c) = .3$ and $P(c|e) = .75$) but different sample size $N$ (40, 200, 400). Increasing the sample size strongly influences the posterior probability of structure $S_0$ (middle left). As a consequence, the diagnostic probability derived from the structure induction model approximates the empirical diagnostic probability when $N$ increases (bottom left). Right: Three data sets with identical predictive and diagnostic probability ($P(e|c) = .3$ and $P(c|e) = .6$) but different sample size $N$ (40, 200, 400). Increasing the sample size has a moderate influence on the posterior probability of $S_0$, and the difference between the empirical probability and the diagnostic probability derived from the structure induction model remains largely unaffected.

(Peterson & Beach, 1967), we agree with the view that this kind of model is primarily a "rational description" rather than a "rational calculation" (Chater, Oaksford, Nakisa, & Redington, 2003). Therefore, an interesting follow-up to the present research would be to explore algorithmic-level models that specify the actual processes and inferential steps by which people arrive at their diagnostic judgments. Such an account should be constrained by the predictions of the structure induction model and the obtained empirical findings, but at the same time it will provide a computationally simpler approach.

The crucial task for such a model would be to specify how to come up with an estimate of the probability that there exists a causal relation between $C$ and $E$ without using the full machinery of Bayesian inference. One possibility would be to resort to some proxy that tends to correlate with the presence of a causal link but is easier to compute, for instance the contingency $\Delta P$ (Ward & Jenkins, 1965), some measure of the covariation of $C$ and $E$ (Hattori & Oaksford, 2007) or, even simpler, the observed predictive probability of effect given cause, $P(e|c)$ (cf. Meder, Gerstenberg, Hagmayer, & Waldmann, 2010). Using the observed conditional diagnostic probability as an anchor that is adjusted in the directions provided by the proxy might be one possible implementation of a heuristic that could be tested in future research.

## Extensions and Limitations of the Structure Induction Model

While the focus of the present work was on diagnostic inferences from effect to cause, the structure induction model can also be used to model other types of causal inferences. The basic rationale is identical, namely, deriving the quantity of interest separately under each causal structure and then integrating them out to obtain a single estimate that reflects the associated structure uncertainty. One example was already given above: modeling predictive inferences (see Figure 6 and Appendix A). Another example is inferring estimates of causal responsibility, $P(c \rightarrow e | e)$, which then would yield a structure induction model of causal attribution (see Appendix C). Thus, the structure induction model is not restricted to diagnostic inferences but provides a principled method for taking into account causal structure uncertainty, offering a link to the existing literature on the consideration of structure in causal learning, reasoning, and decision making (Anderson, 1990; Griffiths & Tenenbaum, 2005; Hagmayer & Meder, 2013; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Meder et al., 2008; Waldmann & Hagmayer, 2013; Waldmann & Holyoak, 1992).

Another issue for future research concerns the generation of the space of possible causal structures. We focused here on the two elementary structures $S_0$ and $S_1$, which are particularly relevant in the context of elemental causal induction, as $S_1$ is the default structure in power PC theory and $S_0$ formalizes the hypothesis that there is no causal relation between $C$ and $E$. In line with other researchers (Anderson, 1990; Griffiths & Tenenbaum, 2005), we consider the question of whether the data provide sufficient evidence for the existence of a causal link as the most elemental type of structure uncertainty that a rational agent should take into account. However, depending on the computational complexity of the task, domain-specific knowledge and other cues to causal structure, the hypothesis space may differ.

Meder, Mayrhofer, and Waldmann (2009) used a variant of the structure induction model that included a third possible structure, one in which the effect is exclusively generated by the cause (i.e., $w_a$ is fixed to 0). In many scenarios, this structure will receive zero posterior probability, because a single case in which the effect occurs in the absence of the cause will render the posterior probability of this structure minimal. Therefore, including this structure will not affect the model's predictions in such cases. However, an extended model might be necessary when prior knowledge or the observed data strongly suggest necessary causal relations. The studies in Meder, Mayrhofer, and Waldmann (2009) also explored a wider range of experimental conditions, such as a between-subjects design and scenarios in which the base rate of the cause was varied. The results of these studies provide further evidence for the robustness of our findings.

More generally, an issue for further research will be to explore a greater variety of tasks and domains in which there might be different prior assumptions about the relative plausibility of the different causal structures. Formally, such prior knowledge can be incorporated by using a non-uniform prior over the space of causal structure hypotheses. For instance, consider a scenario in which participants are presented with the same data sets but the candidate cause is not a virus but the zodiac sign of the patients. Such a context would probably induce a higher prior on $S_0$ (i.e., absence of a causal link) and not a uniform prior over the causal structures.

This issue leads to the more general question of how the structure induction model could be scaled up to situations involving multiple variables and more complex causal networks. One possibility is that the hypothesis space could consist of all possible causal structures containing the considered variables (e.g., Steyvers et al., 2003). This approach, however, does not scale, as the number of possible causal structures grows exponentially with the number of variables. One way to address this problem is to consider the role of additional information that constrains the space of possible causal models, such as prior knowledge or temporal information (Buehner & May, 2003; Gopnik & Meltzoff, 1997; Hagmayer & Waldmann, 2002; Keil, 1989, 2003; Lagnado & Sloman, 2006; Lagnado et al., 2007; Murphy & Medin, 1985; Waldmann, 1996).

Griffiths and Tenenbaum (2009) have developed a general computational framework that uses prior knowledge as a constraint on the causal models under consideration. This knowledge is more general and at a higher level of abstraction than a specific causal hypothesis (Griffiths & Tenenbaum, 2007; Tenenbaum, Griffiths, & Niyogi, 2007). Their "causal grammar," which is implemented as a hierarchical Bayesian model, specifies the variables that form the causal structure, the considered relations between them, and the functional form of these relations. Importantly, it generates and constrains the space of plausible causal models, thereby addressing the problem of combinatorial explosion. Integrating our model with this framework would allow for scaling the approach to more complex scenarios and would provide a generic way to include prior knowledge.

## Concluding Remarks

The goal of the present work was to examine the normative and descriptive validity of different computational models of elemental causal diagnostic reasoning. The intuition behind our theoretical analyses was that it is important to distinguish between the (observable) data level and the (unobservable) causal level and to take into account alternative causal structures that may have generated the observed contingencies. We have argued that a purely statistical model of diagnostic reasoning is inadequate from a normative perspective. Moreover, our empirical studies reveal the descriptive inadequacy of such an account. In fact, although participants' behavior in our studies looks flawed and biased from the perspective of the simple Bayes model, our analyses show that the judgment patterns should instead be considered as resulting from a causal inference strategy that is well adapted to the uncertainties of the world.

## References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35,* 303–314. doi:10.1037/0022-3514.35.5.303

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual process. *Behavioral and Brain Sciences, 30,* 241–254. doi:10.1017/S0140525X07001653

Brighton, H., & Gigerenzer, G. (2012). Are rational actor models "rational" outside small worlds? In S. Okasha, & K. Binmore (Eds.), *Evolution and rationality: Decisions, co-operation, and strategic behaviour* (pp. 84–109). doi:10.1017/CBO9780511792601.006

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,* 1119–1140. doi:10.1037/0278-7393.29.6.1119

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 56,* 865–890. doi:10.1080/02724980244000675

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences, 3,* 57–65. doi:10.1016/S1364-6613(98)01273-X

Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind*. New York, NY: Oxford University Press.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes, 90,* 63–86. doi:10.1016/S0749-5978(02)00508-3

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. doi:10.1037/0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal reasoning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review, 112,* 694–706. doi:10.1037/0033-295X.112.3.694

Chickering, D. M., & Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning, 29,* 181–212. doi:10.1023/A:1007469629108

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 928–935. doi:10.1037/0096-1523.7.4.928

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions of the literature on judgment under uncertainty. *Cognition, 58,* 1–73. doi:10.1016/0010-0277(95)00664-8

Dawes, R. M., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science, 4,* 396–400. doi:10.1111/j.1467-9280.1993.tb00588.x

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). doi:10.1017/CBO9780511809477.019

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York, NY: Wiley.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science, 21,* 329–336. doi:10.1177/0956797610361430

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General, 140,* 168–185. doi:10.1037/a0022100

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation, 4,* 64–88. doi:10.1080/19462166.2012.682655

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. *The Psychology of Learning and Motivation, 57,* 1–55. doi:10.1016/B978-0-12-394293-7.00001-7

Fiedler, K., & Juslin, P. (Eds.). (2006). *Information sampling and adaptive cognition*. New York, NY: Cambridge University Press.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition, 115,* 166–171. doi:10.1016/j.cognition.2009.12.011

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704. doi:10.1037/0033-295X.102.4.684

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences, 7,* 43–48. doi:10.1016/S1364-6613(02)00009-8

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naïve causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25,* 565–610. doi:10.1207/s15516709cog2504_3

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384. doi:10.1016/j.cogpsych.2005.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 323–345). doi:10.1093/acprof:oso/9780195176803.003.0021

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661–716. doi:10.1037/a0017201

Hacking, I. (2001). *An introduction to probability and inductive logic*. doi:10.1017/CBO9780511801297

Hagmayer, Y., & Meder, B. (2013). Repeated causal decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 33–50. doi:10.1037/a0028643

Hagmayer, Y., & Sloman, S. A. (2009). People conceive of their choices as intervention. *Journal of Experimental Psychology: General, 138,* 22–38. doi:10.1037/a0014585

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition, 30,* 1128–1137. doi:10.3758/BF03194330

Hagmayer, Y., & Waldmann, M. R. (2007). Inferences about unobserved causes in human contingency learning. *The Quarterly Journal of Experimental Psychology, 60,* 330–355. doi:10.1080/17470210601002470

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science, 31,* 765–814. doi:10.1080/03640210701530755

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General, 139,* 702–727. doi:10.1037/a0020488

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34,* 169–188. doi:10.1017/S0140525X10003134

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251. doi:10.1037/h0034747

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review, 107,* 397–402. doi:10.1037/0033-295X.107.2.397

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7,* 368–373. doi:10.1016/S1364-6613(03)00158-X

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19,* 1–53. doi:10.1017/S0140525X00041157

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136,* 430–450. doi:10.1037/0096-3445.136.3.430

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 451–460. doi:10.1037/0278-7393.32.3.451

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). doi:10.1093/acprof:oso/9780195176803.003.0011

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science, 18,* 1014–1021. doi:10.1111/j.1467-9280.2007.02017.x

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–984. doi:10.1037/a0013256

Luhmann, C. C., & Ahn, W. (2007). BUCKLE: A model of unobserved cause learning. *Psychological Review, 114,* 657–677. doi:10.1037/0033-295X.114.3.657

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge, England: Cambridge University Press.

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

Mayrhofer, R., Nagel, J., & Waldmann, M. R. (2010). The role of causal schemas in inductive reasoning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1082–1087). Austin, TX: Cognitive Science Society.

Mayrhofer, R., & Waldmann, M. R. (2011). Heuristics in covariation-based induction of causal models: Sufficiency and necessity priors. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3110–3115). Austin, TX: Cognitive Science Society.

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and intervening: Rational and heuristic models of causal decision making. *The Open Psychology Journal, 3,* 119–135. doi:10.2174/1874350101003020119

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review, 15,* 75–80. doi:10.3758/PBR.15.1.75

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition, 37,* 249–264. doi:10.3758/MC.37.3.249

Meder, B., & Mayrhofer, R. (2013). Sequential diagnostic reasoning with verbal information. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1014–1019). Austin, TX: Cognitive Science Society.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2009). A rational model of elemental diagnostic inference. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2176–2181). Austin, TX: Cognitive Science Society.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316. doi:10.1037/0033-295X.92.3.289

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan-Kaufmann.

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

Peterson, C. R., & Beach, L. R. (1967). Man as intuitive statistician. *Psychological Bulletin, 68,* 29–46. doi:10.1037/h0024722

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10,* 33–51. doi:10.1002/(SICI)1099-0771(199703)10:1<33::AID-BDM244>3.0.CO;2-6

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130,* 380–400. doi:10.1037/0096-3445.130.3.380

Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29,* 5–39. doi:10.1207/s15516709cog2901_2

Steyvers, M., Tenenbaum, J., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453–489. doi:10.1207/s15516709cog2703_6

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 301–322). doi:10.1093/acprof:oso/9780195176803.003.0020

Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. doi:10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). doi:10.1017/CBO9780511809477.011

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

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 453–484). doi:10.1093/acprof:oso/9780199216093.003.0020

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 216–227. doi:10.1037/0278-7393.31.2.216

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed., )*Oxford handbook of cognitive psychology* (pp. 733–752). doi:10.1093/oxfordhb/9780195376746.013.0046

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science, 15,* 307–311. doi:10.1111/j.1467-8721.2006.00458.x

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121,* 222–236. doi:10.1037/0096-3445.121.2.222

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology/Revue canadienne de psychologie, 19,* 231–241. doi:10.1037/h0082908

Yeung, S., & Griffiths, T. (2011). Estimating human priors on causal strength. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedingsof the 33rd Annual Conference of the Cognitive Science Society* (pp. 1709–1714). Austin, TX: Cognitive Science Society.

# Appendix A

## Structure Induction Model

The structure induction model considers two causal structures that may underlie the observed joint distribution of target cause $C$ and effect $E$, structures $S_0$ and $S_1$ (Figure 2). Associated with each structure is a set of parameters $\theta$, which for structure $S_1$ consists of the base rate ($b_c$) and causal strength ($w_c$) of target cause $C$, and the influence of the background cause $A$ ($w_a$). Structure $S_0$ has only two parameters, $b_c$ and $w_a$ (i.e., $w_c$ is fixed to 0). The posterior probability distributions of a structure's parameters conditional on the data, $P(\theta|D; S_i)$, is given by

$$P(\theta|D; S_i) = \frac{P(D|\theta; S_i)P(\theta|S_i)}{P(D|S_i)}, \tag{A1}$$

where $P(D|\theta; S_i)$ is the likelihood of the data given the parameter values (see Equations A2 and A3, respectively), $P(\theta|S_i)$ refers to the prior probabilities of the parameters, which we set to independent Beta(1, 1) distributions (i.e., flat priors), and $P(D|S_i)$ is a normalizing constant.

Under a noisy-OR parameterization (Pearl, 1988), for which Cheng's (1997) causal power measure is the maximum likelihood estimate (Griffiths & Tenenbaum, 2005), the likelihood functions $P(D|\theta; S_i)$ are given by

$$P(D|\theta; S_0) = [(1 - b_c)(1 - w_a)]^{N(\neg c,\neg e)} \cdot [(1 - b_c)w_a]^{N(\neg c,e)} \cdot [b_c(1 - w_a)]^{N(c,\neg e)} \cdot [b_c w_a]^{N(c,e)} \tag{A2}$$

for structure $S_0$ and

$$P(D|\theta; S_1) = [(1 - b_c)(1 - w_a)]^{N(\neg c,\neg e)} \cdot [(1 - b_c)w_a]^{N(\neg c,e)} \cdot [b_c(1 - w_c)(1 - w_a)]^{N(c,\neg e)} \cdot [b_c(w_c + w_a - w_c w_a)]^{N(c,e)} \tag{A3}$$

for structure $S_1$.

The parameters' posteriors are derived separately for each of the two causal structures $S_0$ and $S_1$ (see Figure 2). Under structure $S_0$, the parameter $w_c = 0$; thus, the likelihood function simplifies accordingly (see Equation A2 vs. A3).

### Causal Structure Posteriors

The posterior probability of the causal structures $S_0$ and $S_1$ is given by

$$P(S_i|D) = \frac{P(D|S_i)P(S_i)}{P(D)} \text{ with } P(D) = \sum_{i\in\{0,1\}} P(D|S_i)P(S_i), \tag{A4}$$

where $P(D|S_i)$ denotes the likelihood of the data given the structure (see Equation A5), $P(S_i)$ denotes the prior probability of structure $S_i$ (which was set to .5 for both structures, i.e., a uniform prior), and $P(D)$ is a normalizing constant.

The likelihood of the data given structure $S_i$, $P(D|S_i)$, is the integral over the likelihood functions of the parameters (see Equations A2 and A3) under structure $S_i$:

$$P(D|S_i) = \iiint P(D|\theta; S_i)P(\theta|S_i)d\theta \text{ with } \theta = (b_c, w_c, w_a), \tag{A5}$$

where $P(\theta|S_i)$ denotes the joint prior probability over the structures' parameters, which we set to independent Beta(1, 1) distributions (i.e., flat priors) for all parameters within each structure.

### Modeling Diagnostic Inferences

The diagnostic probability of cause given effect is derived separately for each parameterized causal structure ($S_0$ and $S_1$) by integrating over the parameters' values weighted by their posterior probability:

$$\begin{aligned} P(c|e; D, S_i) &= \iiint P(c|e; \theta, S_i)P(\theta|D; S_i)d\theta \\ &= \iiint P(c|e; \theta, S_i)\frac{P(D|\theta; S_i)}{P(D|S_i)}P(\theta|S_i)d\theta \end{aligned} \tag{A6}$$

with

$$P(c|e;\theta, S_0) = b_c \tag{A7}$$

and

$$P(c|e;\ \theta, S_1) = \frac{(w_c + w_a - w_c w_a)b_c}{(w_c + w_a - w_c w_a)b_c + w_a(1 - b_c)}. \tag{A8}$$

The same procedure can be used to model predictive inferences (i.e., for deriving an estimate of the probability of effect given cause, $P(e|c)$; see Equation 3). The result is an estimate of the diagnostic (or predictive) probability under each of the two (parameterized) causal structures.

### Integrating Out the Causal Structures

The final step is to integrate out the two alternative causal structures to obtain a single value for the diagnostic probability $P(c|e)$ that takes into account uncertainty about the underlying generative causal structure. This is done by summing over the values of $P(c|e)$ derived under structures $S_0$ and $S_1$ (see Equation A6), with each value being weighted by the posterior probability of the respective structure (see Equation A4):

(*Appendices continue*)

$$P(c|e;\, D) = \sum_{i\in\{0,1\}} P(c|e;\, D, S_i) \cdot P(S_i|D). \tag{A9}$$

The result of this Bayesian model averaging is a single value for the diagnostic probability $P(c|e)$ that takes into account uncertainty about the causal structures and about the associated parameter values. The same procedure can be used to obtain an estimate of the predictive probability of effect given cause, $P(e|c)$, or an estimate of causal responsibility (see Appendix C).

### Implementation of the Structure Induction Model

R code that implements the structure induction model is available in the online supplemental materials (http://dx.doi.org/10.1037/a0035944.supp). The R code uses Monte-Carlo simulations to evaluate the integrals (see Equations A5 and A6) by drawing $m$ = 1,000,000 independent samples of parameters $b_c$, $w_c$, and $w_a$ from a uniform Beta(1, 1) distribution on the interval [0, 1] (i.e., flat prior; cf. Griffiths & Tenenbaum, 2005).

Accordingly, Equation A5 is approximated by

$$P(D|S_i) = \frac{1}{m}\sum_{k=1}^{m} P(D|\theta^k;\, S_i) \text{ with } \theta^k = \left(b_c^k, w_c^k, w_a^k\right) \tag{A10}$$

and Equation A6 by

$$P(c|e;\, S_i) = \frac{1}{m}\sum_{k=1}^{m} P(c|e;\, \theta^k, S_i)\frac{P(D|\theta^k;\, S_i)}{P(D|S_i)} \text{ with } \theta^k = \left(b_c^k, w_c^k, w_a^k\right) \tag{A11}$$

with $\theta^k$ being one sample from three independent Beta(1, 1) distributions. Note that for structure $S_0$ the parameter set $\theta^k$ consists of only two parameters, $b_c$ and $w_a$.

## Appendix B

## Bayesian Power PC Model of Diagnostic Reasoning

The standard power PC model (Cheng, 1997) uses maximum likelihood point estimates to parameterize the default causal structure $S_1$ (Figure 2). Here we consider two Bayesian variants of the model, one with uniform priors over the parameters $b_c$, $w_c$, and $w_a$, and one using the sparse and strong (SS) prior suggested by Lu et al. (2008).

### Power PC Model With Uniform Priors

For the power PC model with uniform priors, the parameters of structure $S_1$ ($b_c$, $w_c$, and $w_a$) were set to independent Beta(1, 1) distributions. The parameters' posterior distributions were derived using Equations A1 and A3 (Appendix A). As in the structure induction model, the parameters are then integrated out (see Equation A5), and the diagnostic probability of cause given effect is computed (see Equations A6 and A8). The implementation is analogous to that of the structure induction model, that is, the likelihood function and the diagnostic probability of cause given effect are approximated through Monte-Carlo simulations (see Equations A10 and A11). The derived diagnostic probability takes into account uncertainty regarding parameter estimates, but does not consider uncertainty regarding alternative causal structures that may have generated the observed data.

### Power PC Model With Sparse and Strong (SS) Priors

The SS power model was originally developed to account for structure and strength judgments based on contingency data (Lu et al., 2008).[B1] However, the idea that people bring generic assumptions about causal systems to the task can also be applied to diagnostic inferences. Lu et al. (2008) implemented the SS power model as a Bayesian inference over structure $S_1$ with specific (i.e., sparse and strong) priors over the structure's parameters $w_c$ and $w_a$ (see also Equation 10 in Lu et al., 2008):

$$P(w_c, w_a) \propto e^{-\alpha w_a - \alpha(1-w_c)} + e^{-\alpha(1-w_a) - \alpha w_c}, \tag{B1}$$

where $\alpha$ is a free parameter controlling a reasoner's preference for sparse and strong causes. Following Lu et al. (2008), we set $\alpha$ = 5.

Since the SS power model does not make assumptions about the base rate of the cause (as it does not matter for strength judgments), we assumed a flat prior for $b_c$. With these prior assumptions, the same logic as for structure $S_1$ in the structure induction model can be applied to compute the diagnostic probability within the SS power model:

$$P(c|e;\, D) = \iiint P(c|e;\, \theta)P(\theta|D)d\theta \text{ with } \theta = (b_c, w_c, w_a) \tag{B2}$$

with $P(c|e;\, \theta)$ as described in Equation A8 and $P(\theta|D)$ as described in Equation A1 for structure $S_1$ within the structure induction model (see Appendix A).

[B1] We use the strength-estimate version of the model's prior here; the structure version of the SS power model employs additional assumptions about $w_c$. Since the structure version of the SS power model achieved worse fits than the strength version, we do not discuss this model variant further.

(*Appendices continue*)

### Implementation of the SS Power Model of Diagnostic Reasoning

As for the structure induction model, we evaluated the integral given in Equation B2 with a Monte-Carlo simulation by drawing $m$ = 1,000,000 independent samples from the respective prior distributions:

$$P(c|e) = \frac{1}{m}\sum_{k=1}^{m} P(c|e;\, \theta^k)\frac{P(D|\theta^k)}{P(D)} \quad \text{with} \quad \theta^k = \left(b_c^k, w_c^k, w_a^k\right) \tag{B3}$$

with ($w_c$, $w_a$) drawn from the joint SS prior distribution (see Equation B1) and $b_c$ drawn from a Beta(1, 1) distribution.

## Appendix C

## Causal Attribution Models

A causal inference framework allows us to model different types of diagnostic quantities beyond the conditional probability of cause given effect, such as estimates of how likely it is that the observed effect was indeed produced by candidate cause $C$. This constitutes a measure of *causal responsibility*, rather than diagnostic conditional probability (Cheng & Novick, 2005; Holyoak et al., 2010). Different implementations of a causal attribution model are possible that differ in the degree to which they take into account uncertainty regarding the parameter estimates and the underlying causal structure.

### Power PC Model of Causal Attribution

Let $c \rightarrow e$ denote that effect $E$ is produced by cause $C$ (Cheng & Novick, 2005). Whether the occurrence of effect $E$ can be attributed to the occurrence of $C$ translates to determining the conditional probability $P(c \rightarrow e|e)$. If "$C$ caused $E$" holds, it is necessarily the case that $E$ occurred, therefore $P(e|c \rightarrow e) = 1$. According to Bayes' rule,

$$\begin{aligned} P(c \rightarrow e|e) &= \frac{P(e|c \rightarrow e) \cdot P(c \rightarrow e)}{P(e)} = \frac{P(c \rightarrow e)}{P(e)} \\ &= \frac{b_c w_c}{b_c w_c + w_a - b_c w_c w_a}. \end{aligned} \tag{C1}$$

Equation C1 allows for deriving estimates of causal responsibility under causal power assumptions. Note that all parameters involved in these computations can be inferred from observable contingency data. In the standard power PC model, the parameters are maximum likelihood point estimates directly derived from the data.

These derivations also reveal that typically $P(c \rightarrow e|e) < P(c|e)$. The reason for the smaller values is that estimates of causal responsibility partial out cases in which the effect was caused by the background event $A$. Only when there are no alternative causes (i.e., $P(e|\neg c) = w_a = 0$), it holds that $P(c \rightarrow e|e) = P(c|e)$. Conversely, the maximal difference is obtained when $C$ and $E$ are independent, that is, when $w_c = 0$. In this case, $P(c \rightarrow e|e) = 0$, as the lack of a causal relation between $C$ and $E$ entails that $C$ cannot be responsible for the occurrence of $E$, whereas $P(c|e) = P(c)$.

### Bayesian Power PC Model of Causal Attribution

Holyoak et al. (2010) used a Bayesian variant of power PC theory to infer estimates of causal responsibility that take parameter uncertainty into account (Holyoak et al., 2010; see also Lu et al., 2008). The parameters of structure $S_1$ are represented by uniform prior probability distributions, and the posteriors are estimated using Bayesian inference (see Equation A1). The key difference from the standard power PC model with uniform priors (see Appendix B) is that not the conditional probability of cause given effect (see Equation A8) but an estimate of causal responsibility, $P(c \rightarrow e|e)$ (i.e., Equation C1), is derived. As with the other Bayesian models, Monte-Carlo simulations were used to implement the model (see Equations A10 and A11).

### Structure Induction Model of Causal Attribution

The derivation of an estimate of causal responsibility in the structure induction model is analogous to deriving diagnostic or predictive probabilities (see Appendix A). As before, the posterior parameter distributions of the two structures $S_0$ and $S_1$ are derived, and a posterior over the structure space is computed. The conceptual difference is that instead of deriving a conditional probability, an estimate of causal responsibility under each structure is obtained. Structure $S_0$ states that there is no causal relation between $C$ and $E$, and therefore this structure entails that $P(c \rightarrow e|e) = 0$. Note the difference from the diagnostic probability of cause given effect under $S_0$, which can never be lower than the base rate of the cause (see the corresponding Equation A7). Under structure $S_1$, Equation C1 is used to derive an estimate of causal responsibility instead of Equation A8.

As before, when the structures are integrated out, the resulting estimate of $P(c \rightarrow e|e)$ depends on the relative posterior probabilities of the structures. For instance, since structure $S_0$ entails that $P(c \rightarrow e|e) = 0$, the higher the posterior of $S_0$, the more closely the final estimate approximates zero. We used Monte-Carlo simulations to implement the model (see Appendix A for details).

Received April 6, 2013
Revision received November 18, 2013
Accepted November 19, 2013 ■