A Rational Model of Elemental Diagnostic Inference

Björn Meder (meder@mpib-berlin.mpg.de)
Ralf Mayrhofer (rmayrho@uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Whereas the traditional normative benchmark for diagnostic reasoning from effects to causes is provided by purely statistical norms, we here approach the task from the perspective of rational causal inference. The core feature of the presented model is the assumption that diagnostic inferences are constrained by hypotheses about the causal texture of the domain. As a consequence, the model's predictions systematically deviate from classical, purely statistical norms of diagnostic inference. In particular, the analysis reveals that diagnostic judgments should not only be influenced by the probability of the cause given the effect, but also be systematically affected by the predictive relation between cause and effect. This prediction is tested in three studies. The obtained pattern of diagnostic reasoning is at variance with the traditional statistical norm but consistent with a model of rational causal inference.

Keywords: Rational model; Causal learning; Causal reasoning; Bayesian inference; Computational Modeling

Introduction

In this paper we present a rational analysis of diagnostic reasoning – the process of reasoning from effects to causes. Diagnostic inferences are not only ubiquitous 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 or when we try to identify the causes of why our computer crashed once again. Whereas the traditional normative yardstick for such inferences is provided by purely statistical norms, we use the framework of *causal-model theory* (e.g., Pearl, 2000; Waldmann & Holoyak, 1992; Waldmann, Hagmayer, & Blaisdell, 2006) and *causal Bayesian inference* (Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008) to elucidate the relevant kinds of inputs, computations, and outputs involved in diagnostic reasoning.

We here focus on the most basic type of diagnostic inference, which involves a single cause-effect relation between two binary events. Based on a rational analysis of such diagnostic inferences we have developed a computational model that details the influence of competing hypotheses about causal structure and causal strength. Whereas it is usually assumed that diagnostic judgments should merely be a function of the empirical conditional probability $P(\text{Cause} \mid \text{Effect})$, our analysis reveals that diagnostic inferences should also be systematically affected by the predictive probability $P(\text{Effect} \mid \text{Cause})$ and by the causal power (Cheng, 1997) of the target cause. We tested the model's predictions in three studies. While the observed pattern of reasoning appears irrational from a purely statistical perspective, our analyses suggest that it may be viewed as resulting from a rational inference strategy that is well adapted to the goal of acquiring and using causal knowledge.

"Naïve Bayes" as a Norm of Diagnostic Inference

Let C denote a binary cause and E a binary effect, and let c^+, c^- and e^+, e^- indicate the presence and absence, respectively, of these events. Making a diagnostic judgment from effect to cause can then be expressed as estimating the conditional probability of the cause given the effect, $P(c^+|e^+)$. Given a joint frequency distribution over C and E the empirical conditional probability $P(c^+|e^+)$ can be directly estimated from the frequency of co-occurrences $N(\cdot)$. Alternatively, one can use Bayes' rule to derive this probability from the conditional probability of the effect given the cause, $P(e^+|c^+)$, the base rate of the target cause, $P(c^+)$, and the marginal probability of the effect, $P(e^+)$:

$$P(c^+|e^+) = \frac{P(e^+|c^+) \cdot P(c^+)}{P(e^+)} \qquad (1)$$

We refer to this approach as *naïve Bayes* because under this view the application of Bayes' rule is nothing but an elementary result of standard probability theory. In particular, no reference is made to the generative causal processes underlying the observed events, and no uncertainty about parameter estimates is assumed in these computations.

This use of Bayes' rule provides the classical statistical norm to which peoples' diagnostic judgments usually have been compared (e.g., Kahneman & Tversky, 1973). Several studies have shown that peoples' judgments often substantially deviate from this norm and have attempted to pinpoint factors which lead people to conform to this norm (e.g., Gigerenzer & Hoffrage, 1995). However, the prescriptive validity of this statistical norm has rarely been questioned (but see Krynski & Tenenbaum, 2007). We suggest that approaching diagnostic inferences from the perspective of causal reasoning may provide a more appropriate standard of rational diagnostic inference *and* a better descriptive model of peoples' diagnostic judgments.

A Rational Model of Diagnostic Inference

The core idea behind our model is the assumption that diagnostic inferences operate over causal representations that are estimated from data (cf. Krynski & Tenenbaum, 2007). Thus, the data we encounter are typically interpreted as arising from some unobserved causal processes, and our inference goal when making predictive and diagnostic inferences is to reason about causal relations, not about the noisy data we perceive.

Briefly, our model consists of the following five steps[1]: 1) Specify alternative causal structures that may underlie the data. 2) Use the data to estimate the parameter distributions associated with each causal structure. 3) Compute $P(c^+|e^+)$ for each (parameterized) causal structure. 4) Compute the posterior probability of each causal model. 5) Integrate out the causal models to obtain an overall diagnostic judgment $P(c^+|e^+)$. We next describe these steps in more detail.

(1) Specifying alternative causal structures Given information about the co-occurrences of a single candidate cause C and a single effect E there are three qualitatively different causal structures that might underlie the observed data. These three causal networks, in the following denoted as M_0, M_1, and M_2 are shown in Figure 1.

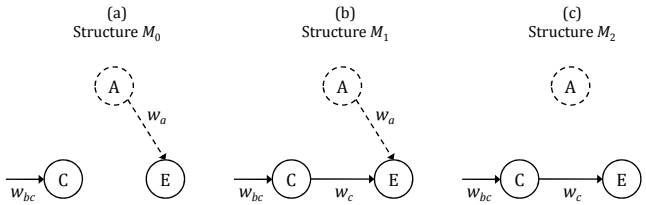

Figure 1: Alternative causal models which may underlie a co-occurrence of a cause event C and an effect event E.

Each causal structure consists of a set of nodes, which represent the domain variables, and directed edges ("causal arrows"), which represent hypotheses about the presence of causal influences connecting the variables. Thus, a core feature of such causal model representations is that they mirror a characteristic feature of our environment, namely the fact that some events, causes, have the power to generate or prevent other events, their effects.

Each structure expresses a different qualitative hypothesis about the generative causal processes assumed to underlie the observation of C and E. According to model M_0, there is *no* causal relation between C and E. Though the two events may sometimes co-occur, the effect is generated by some unobserved background cause (A). The second structure, M_1, states there exists a causal relation between C and E, that is, when C occurs it has the power to produce E. However, there are also alternative background causes that can generate the effect. Finally, according to structure M_2, event C is the only cause of E, as indicated by the missing arrow from A to E. Thus, C is necessary for the occurrence of E.

Note that structures M_0 and M_2, respectively, are not merely special cases of structure M_1, but constitute qualitatively different, less complex hypotheses suggesting different causal explanations for the observed data. For example, due to their simpler form causal structures M_0 or M_2 may have a higher posterior probability than structure M_1 (i.e., Bayesian Occam's Razor), which would not be possible if the former were merely special cases of the latter. Rather, the models form the background against which the observed data is evaluated. This, for example, allows the model to be sensitive to the question to what extent some data D provides evidence for or against the existence of particular causal relations potentially underlying the observations (see also Griffiths & Tenenbaum, 2005).

(2) Parameter estimation Connected with each causal structure is a set of parameters w: w_{bc} denotes the base rate of cause C, w_c denotes the causal strength of C, and w_a encodes the causal influence of an amalgam of further (unobserved) background causes of E.[2] By Bayes' rule, the posterior probability distributions of each model's parameters given the data, $P(w \mid D)$, is proportional to the likelihood of the data given the parameter set w:

$$P(w|D) \propto P(D|w) \cdot P(w) \tag{2}$$

$P(D \mid w)$ is the likelihood of the data given the parameter values for w_{bc}, w_c, and w_a, and $P(w)$ refers to the joint prior probability of the parameters. The prior distributions of the parameters w_{bc}, w_c, and w_a, are independently set to flat, uninformative Beta(1, 1) distributions (e.g., Anderson, 1990; Griffiths & Tenenbaum, 2005). Under a noisy-OR parameterization (e.g., Pearl, 2000), for which Cheng's (1997) causal power measure is the maximum likelihood estimate (MLE), the likelihood function $P(D \mid w)$ is given by

$$\begin{aligned}&[(1-w_{bc})(1-w_a)]^{N(c^-,e^-)} \cdot [(1-w_{bc})w_a]^{N(c^-,e^+)} \cdot \\ &[w_{bc}(1-w_c)(1-w_a]^{N(c^+,e^-)} \cdot [w_{bc}(w_c+w_a-w_cw_a)]^{N(c^+,e^+)}\end{aligned} \tag{3}$$

The posterior distributions of the parameters $P(w \mid D)$ are derived separately for each of the three causal structures. For structure M_1 the parameter set consists of w_{bc}, w_c, and w_a. By contrast, M_0 and M_2 have only two parameters whose probability distributions are updated in light of the available evidence. According to M_0, there is no causal relation between C and E, therefore only estimates for w_{bc} and w_a are derived. Conversely, M_2 represents the possibility that there are no alternative causes. Therefore, only estimates for w_{bc} and w_c are being computed. Note that the derived parameter distributions differ depending on the assumed causal structure. For example, given some data D the posterior distributions on w_c are not the same under structures M_1 and M_2.

(3) Deriving $P(c^+|e^+)$ The next step is to derive an estimate of the diagnostic probability $P(c^+|e^+; M_i, w)$ under the different models given their parameters. According to structure M_0, there is no causal link between C and E, therefore observing E provides no diagnostic evidence for C. Thus, $P(c^+|e^+; M_0, w) = P(c^+; w) = w_{bc}$. In structure M_1, $P(c^+|e^+)$ is derived under the noisy-OR assumption:

$$P(c^+|e^+; M_1, w) = \frac{w_{bc}w_c + w_{bc}w_a - w_{bc}w_cw_a}{w_{bc}w_c + w_a - w_{bc}w_cw_a} \tag{4}$$

Depending on the parameters' prior distributions this estimate may or may not coincide with the empirical conditional probability.

[1] The stepwise description is for illustrative purposes only.

[2] Each parameter is defined over the interval [0,1].

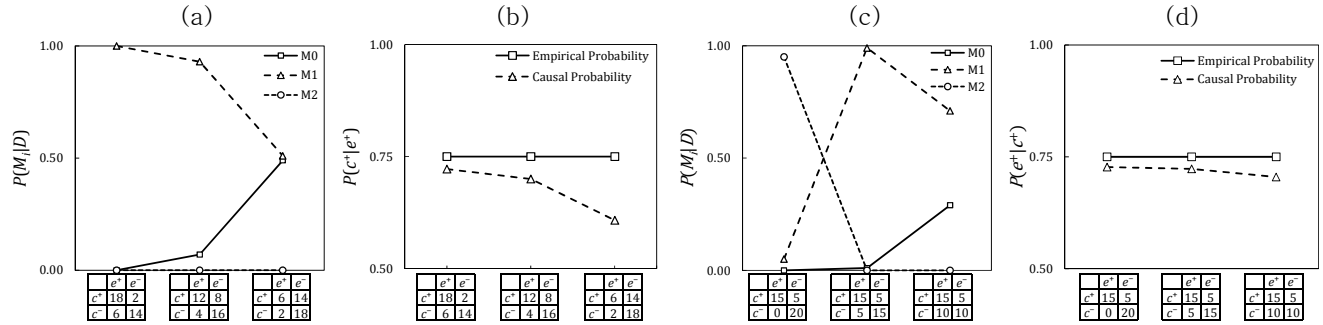

Figure 2. a) Posterior probabilities of models M_0, M_1, and M_2 for different data sets with fixed $P(c^+|e^+)$ and varying $P(c^+|e^+)$. b) Estimates of $P(c^+|e^+)$ after model averaging. c) Posterior probabilities of models M_0, M_1, and M_2 for three data sets with fixed $P(e^+|c^+)$ and varying $P(c^+|e^+)$. d) Estimates of $P(e^+|c^+)$ after model averaging.

Finally, structure M_2 expresses the possibility that C is the only existing cause of E. Therefore, whenever effect event E is present, it is certain that C is also present. Thus, $P(c^+|e^+; M_2, w) = 1$. The target inference for each structure can be computed by integrating over the parameters' values.[3]

(4) Computing the models' posterior probabilities The posterior probability of the models is proportional to the likelihood of the data given the models, weighted by the prior probability of the model (Bayes' rule):

$$P(M_i|D) \propto P(D|M_i) \cdot P(M_i) \tag{5}$$

$P(D \mid M_i)$ is the likelihood of the data given structure M_i, which is simply the integral over the likelihood function of the parameters under structure M_i.[4] $P(M_i)$ denotes the prior probability of structure M_i. We assume that prior to observing any data all three models are equally likely, thus, $P(M_0) = P(M_1) = P(M_2) = 1/3$.

One way to analyze how the predictions of naïve Bayes differ from our model is to look at data sets for which naïve Bayes predicts identical judgments, whereas our model predicts differences. Figure 2a shows the causal models' posterior probabilities for three different data sets with a fixed conditional probability of the cause given the effect (i.e., $P(c^+|e^+) = 0.75$) and varying probability of the effect given the cause (0.9, 0.6, and 0.3, respectively). Since the value of $P(e^+|c^+)$ provides the upper boundary for the causal strength estimate of C this variation strongly affects the likelihood of the three causal models (M_0, M_1, M_2). In particular, the probability of M_1 and M_0 varies systematically with the value of $P(e^+|c^+)$: the weaker this relation is, the more likely M_0 becomes, since a weak observed contingency may merely be a coincidence. When $P(e^+|c^+) = 0.9$, M_1 is the most likely model, but when $P(e^+|c^+) = 0.3$ both structures M_1 and M_0 are equally likely. The probability of M_2 remains at zero since in all data sets the effect sometimes occurs in the absence of the candidate cause.

(5) Integrating out the causal models The final step is to integrate out the alternative causal structures to obtain a single value for the diagnostic probability $P(c^+|e^+)$. This is done by summing over the values of $P(c^+|e^+; M_i)$ derived under each causal model weighted by the posterior probability of the respective model:

$$P(c^+|e^+; D) = \sum_i P(c^+|e^+; M_i) \cdot P(M_i|D) \tag{6}$$

The result of this Bayesian model averaging is a single value for $P(c^+|e^+)$, that takes into account uncertainty about causal structures and parameter estimates. Fig. 2b shows how the resulting estimate for $P(c^+|e^+)$ deviates from the empirical probability. A comparison with Fig. 2a shows that this downward trend is due to the increased likelihood of structure M_0 across the three data sets: the more likely M_0 becomes, the stronger the derived estimate of $P(c^+|e^+)$ deviates from the empirical conditional probability.

Asymmetries between diagnostic and predictive inferences The analyses also reveal an interesting asymmetry between predictive and diagnostic inferences.[5] Whereas our model predicts that diagnostic judgments should be affected by the predictive probability and by the causal strength of the target cause, the converse is not necessarily true. Thus, predictive judgments should usually only be affected by the predictive probability $P(e^+|c^+)$, but not by the diagnostic probability $P(c^+|e^+)$. The reason for this asymmetry is that under structure M_0 the estimated value of w_a is larger than under M_1 since all occurrences of the effect must necessarily be attributed to the influence of the background causes. As a consequence, an increase in the likelihood of M_0 usually entails only a small decrease for estimates of $P(e^+|c^+)$ when integrating out the causal models. Figures 2c and 2d illustrate this prediction for three data sets with a fixed level of $P(e^+|c^+)$ but different values of $P(c^+|e^+)$ (1.0, 0.75, and 0.6, respectively). The figure also shows that M_2 is the most likely model when the effect never occurs in the absence of the target cause, which entails that $P(c^+|e^+) = 1.0$.

Summary The presented model provides a rational account of diagnostic reasoning. Because the computations involve alternative hypotheses about the existence and strength of causal dependencies, the model's predictions substantially

[3] $P(c^+|e^+; M_i) = \iiint_0^1 P(c^+|e^+; M_i, w)P(w)dw$

[4] $P(D|M_i) = \iiint_0^1 P(D|w, M_i)P(w)dw$

[5] Predictive inferences from cause to effect are modeled analogously to the diagnostic inferences. Thus, an estimate of $P(e^+|c^+)$ is derived under the three causal structures, which are then integrated out to obtain a single estimate.

deviate from the empirical conditional probability in the data ("naïve Bayes"). These analyses demonstrate that causal norms and statistical norms do not necessarily coincide and illustrates how rational Bayesian causal inference can lead to very different predictions than a purely statistical account.

Experiment 1

The main goal of Experiment 1 was to investigate whether people's diagnostic judgments are indeed not only affected by the diagnostic probability $P(c^+|e^+)$, but also by the predictive probability $P(e^+|c^+)$ and the causal strength of the candidate cause. We therefore factorially combined three levels of the diagnostic probability $P(c^+|e^+)$ (1.0, 0.75, and 0.6, respectively) with three levels of the predictive probability $P(c^+|e^+)$ (0.9, 0.6, and 0.3, respectively). The resulting nine conditions are shown in Table 1.

Participants and Design Thirty-six University of Göttingen undergraduates participated for course credit. The factors 'learning data' and 'type of causal judgment' (predictive vs. diagnostic) were varied within subjects.

Instructions We used a medical scenario according to which physicians are investigating how certain diseases causally relate to the presence of certain substances found in the blood of people. Participants were told that they would be requested to make two judgments after being presented with some data. They were informed that one question would require them to make an inference from the cause event (disease) to the effect event (substance), whereas the second question would refer to a diagnostic inference question from effect (substance) to its potential cause (disease).

Table 1. Learning data in Experiment 1.

	$P(c^+\|e^+)$		
	1.0	0.75	0.6
$P(e^+\|c^+)$	18/20	18/20	18/20
$P(e^+\|c^-)$	0/20	6/20	12/20
$P(e^+\|c^+)$	12/20	12/20	12/20
$P(e^+\|c^-)$	0/20	4/20	8/20
$P(e^+\|c^+)$	6/20	6/20	6/20
$P(e^+\|c^-)$	0/20	2/20	4/20

Learning Data Subsequent to reading the instructions participants received a sheet of paper presenting 40 (randomized) individual cases referring to patients who had been tested for the presence of the disease and substance, respectively (Table 1). Each disease-substance combination was denoted by different (fictitious) labels (e.g., Midosis/Rothan). The order of the nine disease-substance combinations was counterbalanced across subjects.

Test Phase After examining the data sheet, participants were presented with the two test questions, with the order of questions being counterbalanced across participants. The predictive question reads like this (translated from German): "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". 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". Subsequent to answering the two questions participants proceeded to the next disease-substance combination.

Results and Discussion Table 2 shows participants' responses to the predictive and diagnostic inference questions. A first inspection of the data indicates that the predictive judgments were not affected by the diagnostic probability $P(c^+|e^+)$, but that the diagnostic causal judgments seem to decrease proportionally to the size of the predictive probability $P(e^+|c^+)$.

Table 2. Mean estimates (±SEM) for predictive and diagnostic inference questions in Experiment 1. All judgments were made on a scale from 0 to 7.

		$P(c^+ \mid e^+)$					
		1.0		0.75		0.6	
		Pred.	Diag.	Pred.	Diag.	Pred.	Diag.
$P(e^+ \mid c^+)$	0.9	5.67 (.23)	6.31 (.17)	5.67 (.13)	5.00 (.17)	5.42 (.17)	4.11 (.20)
	0.6	4.36 (.27)	5.86 (.31)	4.03 (.18)	4.28 (.19)	4.03 (.17)	3.97 (.14)
	0.3	2.78 (.25)	5.94 (.32)	2.78 (.23)	3.72 (.26)	2.58 (.17)	3.36 (.21)

For the predictive inference questions, an analysis of variance with level of predictive probability $P(e^+|c^+)$ and level of diagnostic probability $P(c^+|e^+)$ as within-subject variables revealed a main effect of predictive probability, $F(2, 70) = 100.7$, $p < .001$, but no effect of diagnostic probability, $F(2, 70) = 1.29$, $p = .28$. Thus, participants' responses to the predictive causal inference questions were only determined by the predictive probability $P(e^+|c^+)$. The same analysis was conducted for the diagnostic questions. Not surprisingly, participants' responses to these questions were strongly influenced by the diagnostic probability $P(c^+|e^+)$, $F(2, 70) = 74.02$, $p < .001$. However, the analysis also revealed a strong influence of the *predictive* probability $P(e^+|c^+)$, $F(2, 70) = 12.83$, $p < .001$. The lower the predictive probability $P(e^+|c^+)$, the lower the diagnostic judgments turned out to be, despite identical empirical probabilities $P(c^+|e^+)$. An exception seemed to be when the cause is a necessary event for the effect (i.e., when $P(c^+|e^+) = 1.0$); here the response pattern indicates only a weak influence of the predictive probability $P(e^+|c^+)$. As a consequence, the analysis also revealed a significant interaction, $F(4, 140) = 2.51$, $p < .05$.

Taken together, these results show that participants' predictive judgments were only affected by the probability of the effect given the cause, while the diagnostic inferences were affected by both the predictive and diagnostic proba-

bility. In particular, the diagnostic judgments systematically decreased with the predictive strength of the cause event. These patterns clearly support our rational model and indicate that participants attempted to make assessments on the causal level rather than on the data level.

Experiment 2

The goal of Experiment 2 was to examine more closely conditions in which the candidate cause is necessary for the effect, which implies that observations of the effect are perfectly diagnostic for the candidate cause. The diagnostic probability $P(c^+|e^+)$ was fixed to values of 1.0 and 0.8, respectively, whereas the predictive probability $P(e^+|c^+)$ could take values of 0.8 and 0.4. The conditions in which $P(c^+|e^+) = 1.0$ provide an interesting test case as our model predicts *no* influence of the predictive probability in this case. The reason is that in these conditions structure M_2 (which entails that $P(c^+|e^+) = 1.0$)) has the highest posterior probability of all three models (cf. Fig. 2c). By contrast, for the two conditions in which $P(c^+|e^+) = 0.8$ an influence of the predictive probability is predicted since structure M_0 is more likely when $P(e^+|c^+) = 0.4$ than when $P(e^+|c^+) = 0.8$.

Participants and Design Ninety-six University of Göttingen undergraduates participated for course credit or were paid 5€. The factor 'learning data' was varied between subjects, the factor 'type of causal judgment' (predictive vs. diagnostic) was varied within subjects. We used the same procedure and materials as in Experiment 1.

Table 3. Learning data in Experiments 2 and 3.

	Experiment 2		*Experiment 3*	
	$P(c^+\|e^+)$		$P(c^+\|e^+)$	
	1.0	0.8	0.8	0.4
$P(e^+\|c^+)$	16/20	16/20	8/10	8/10
$P(e^+\|c^-)$	0/20	4/20	2/10	12/60
$P(e^+\|c^+)$	8/20	8/20	8/20	8/20
$P(e^+\|c^-)$	0/20	2/20	2/10	12/60

Results and Discussion Table 4 shows the results for the predictive and diagnostic inference questions. As in Experiment 1, participants' responses to the predictive inference questions were only determined by the value of the predictive probability $P(e^+|c^+)$, $F(1, 92) = 53.45$, $p < .001$, but not by the diagnostic probability. Similarly, the obtained diagnostic judgments systematically varied with the level of the diagnostic probability, $F(1, 92) = 12.18$, $p < .001$. A more detailed analysis of the diagnostic responses revealed an interesting asymmetry between the conditions in which the target cause was necessary and those in which the effect also occurred in the absence of the target cause. For the two conditions in which the target cause was not necessary for the occurrence of the effect (i.e., $P(c^+|e^+) = 0.8$), participants' judgments again systematically varied in accordance with the predictive strength of the target cause. Thus, people were more certain that the target cause would be present when the cause had a high predictive power (i.e., $P(e^+|c^+) = 0.8$) than when it had a low predictive power (i.e., $P(e^+|c^+) = 0.4$), $t(46) = 1.89$, $p < .05$ (one-tailed). By contrast, the predictive probability had no influence on the diagnostic judgments when the cause was necessary for the occurrence of the effect (i.e., when $P(c^+|e^+) = 1.0$). In these conditions equal estimates were obtained regardless of the predictive strength of the candidate cause.

Table 4. Mean estimates (±SEM) for predictive and diagnostic inference questions in Experiments 2 and 3.

		Experiment 2				*Experiment 3*			
		$P(c^+\|e^+)$				$P(c^+\|e^+)$			
		1.0		0.8		0.8		0.4	
		Pred.	Diag.	Pred.	Diag.	Pred.	Diag.	Pred.	Diag.
$P(e^+\|c^+)$	0.8	5.42 (.20)	5.75 (.38)	5.00 (.26)	5.00 (.31)	5.06 (.22)	5.31 (.19)	4.78 (.29)	3.66 (.30)
	0.4	3.46 (.20)	5.75 (.33)	3.58 (.26)	4.13 (.35)	3.13 (.22)	4.25 (.31)	3.13 (.22)	3.28 (.24)

In summary, in Experiment 2 two major findings were obtained. First, in the conditions in which the target cause was not a necessary event for the occurrence of the effect (i.e., $P(c^+|e^+) = 0.8$), participants' diagnostic judgments again declined with the decrease of the predictive probability $P(e^+|c^+)$. These results replicate the previous findings in a between-subjects design. The second result concerns the conditions in which $P(c^+|e^+) = 1.0$, for which our model predicts that the value of $P(c^+|e^+)$ should not affect the diagnostic judgments. Consistent with this prediction, no influence of the strength of the predictive relation was obtained between these conditions.

Experiment 3

The previous two experiments demonstrated how people's diagnostic inferences are systematically influenced by the probability with which the effect occurs in the presence of the candidate cause. In these studies, the diagnostic probability $P(c^+|e^+)$ was fixed at constant levels by decreasing $P(e^+|c^-)$ along with $P(e^+|c^+)$ (i.e., by varying the strength of the (unobserved) alternative cause). In Experiment 3 we fixed the strength of the unobserved background cause to $P(c^+|e^-) = 0.2$. To fix the diagnostic probability to two different levels (0.8 and 0.4, respectively) we varied the base rate $P(c^+)$ accordingly (Table 3). The rational model again predicts that the diagnostic judgments would decrease with a decline of the predictive probability $P(e^+|c^+)$.

Participants and Design Thirty-two University of Göttingen undergraduates participated for course credit or were paid 5€. The factors 'learning data' and 'type of causal judgment' (predictive vs. diagnostic) were varied within subjects. We used the same procedure and materials as before.

Results and Discussion Table 4 shows the results of Experiment 3. For the predictive inference questions, an analysis of variance with predictive probability $P(e^+|c^+)$ (0.8 vs. 0.4)

and diagnostic probability $P(c^+|e^+)$ (0.8 vs. 0.4) as within-subject variables revealed only a main effect of predictive probability, $F(1, 31) = 68.78$, $p < .001$, but no effect of the diagnostic probability and no interaction. By contrast, participants' responses to the diagnostic inference questions were not only influenced by the value of $P(c^+|e^+)$, $F(1, 31) = 16.53$, $p < .001$, but also by the predictive probability $P(e^+|c^+)$, $F(1, 31) = 14.25$, $p < .001$.

Thus, as in Experiments 1 and 2 participants' estimates for the predictive inference questions were only influenced by the predictive power of the candidate cause, $P(e^+|c^+)$, their diagnostic causal judgments were not only determined by the diagnostic probability $P(c^+|e^+)$, but also by the causal strength of the candidate cause. This pattern again supports our rational model in a situation in which the diagnostic probabilities in the data were fixed by varying the base rate of the target cause.

General Discussion

We have presented a rational model of diagnostic reasoning based on the framework of causal-model theory and causal Bayesian inference. The key feature of this model is that it assumes that diagnostic inferences operate on the causal level, and therefore are sensitive to the noise inherent in the data that are used to infer the underlying causal relations. As a consequence, the model's predictions strongly deviate from a purely data oriented statistical account, which is often considered the normative benchmark for diagnostic inferences. Our studies revealed that participants' diagnostic judgments were systematically affected by the predictive relation between cause and effect and therefore by the causal power of the target cause. This inference pattern is at variance with a purely statistical model ("naïve Bayes") but consistent with a rational causal inference strategy that goes "beyond the information given."

Another weakness of using a purely statistical framework is that it lacks the representational power to express different types of diagnostic queries. A causal inference framework, by contrast, does not only allow us to compute how likely the candidate cause is present given the occurrence of the effect. It can also be used to provide an answer to other queries, for example concerning the degree of "causal responsibility," which refers to the probability that the observed effect was indeed produced by the candidate cause (cf. Cheng & Novick, 2005). In related research we have extended the presented model to inferences regarding this quantity, too, and we have also been able to disentangle different types of diagnostic queries empirically. In addition, we have also applied the model to more complex situations involving multiple observed causes.

In summary, we suggest that acquiring causal knowledge about a domain may be viewed as a fundamental distal goal of an agent with data providing the proximal evidence for achieving this goal. Causal model representations which take into account the generative nature of the causal processes in the environment enable the agent to reach her distal goals, such as prediction, diagnosis, and planning of actions (Krynski & Tenenbaum, 2007; Waldmann, Hagmayer, & Blaisdell, 2006; Waldmann, Cheng, Hagmayer, & Blaisdell, 2008). Analyzing human judgment and decision making from the perspective of causal reasoning may allow us to reach a deeper understanding of the mind than cognitive models restricted to standard probability calculus or traditional logic.

Acknowledgements

We wish to thank Jana Samland, Dana Barthels, and Mira Holzer for collecting the data, and Hongjing Lu, Keith Holyoak, Jonathan Nelson, and Tobias Gerstenberg for helpful comments on the project. This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (DFG, Wa 621/20-1). The first author (B.M.) is now at the Max Planck Institute for Human Development, Berlin, Germany.

References

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

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

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal reasoning: Reply to White (2005) and to Luhmann & Ahn (2005). *Psychological Review, 112,* 694-707.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334-384.

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

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

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955-982.

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

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

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