Agents and Causes:
A Bayesian Error Attribution Model of Causal Reasoning

Ralf Mayrhofer (rmayrho@uni-goettingen.de)
York Hagmayer (york.hagmayer@bio.uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

One of the most fundamental assumptions underlying causal Bayes nets is the Markov constraint. According to this constraint, an inference between a cause and an effect should be invariant across conditions in which other effects of this cause are present or absent. Previous research has demonstrated that reasoners tend to violate this assumption systematically over a wide range of domains. We hypothesize that people are guided by abstract assumptions about the mechanisms underlying otherwise identical causal relations. In particular, we suspect that the distinction between agents and patients, which can be disentangled from the distinction between causes and effects, influences which causal variable people blame when an error occurs. We have developed a causal Bayes net model which captures different error attributions using a hidden common preventive noise source that provides a rational explanation of these apparent violations. Experiments will be presented which confirm predictions derived from the model.

Keywords: causal reasoning; Bayesian modeling; Bayes nets; Markov condition.

Introduction

Causal Bayes net theory is an increasingly popular approach to model causal reasoning in humans, especially in domains in which multiple variables are causally interrelated. Causal Bayes nets can be graphically represented as sets of (observable and hidden) variables that may represent present or absent events, and arrows that express the direction of the causal influences between the interconnected variables (for an example, see Fig. 1).

To make inferences in this network, additional assumptions need to be made about how the three arrows interrelate. A central assumption that turns probabilistic networks into Bayes nets is the Markov condition (see Pearl, 2000). The Markov condition states that for any variable *X* in a set of variables *S* not containing direct or indirect effects of *X*, *X* is jointly independent of all variables in *S* conditional on any set of values of the set of variables that are direct causes of *X*. An effect of *X* is a variable that is connected with a single arrow or a path of arrows pointing from *X* to it. The Markov condition implies in the common-cause model that each effect is independent of all the other effects conditional upon the presence or absence of its cause *C*.

The Markov condition provides Bayes nets with substantial computational power. Assuming conditional independence allows for learning and reasoning about subsets of variables while ignoring the states of other independent variables. For example, we can infer the presence or absence of an effect from the state of its cause without having to consider the states of the other conditionally independent effects. When using Bayes nets we are not forced to believe that in every situation effects of a common cause are conditionally independent. Whenever we have reasons to question this assumption, it is possible to model violations by adding hidden variables (again obeying the Markov constraint) representing unobserved causal influences. However, the validity of the Markov condition is typically assumed as a default unless we have domain knowledge that suggests hidden variables.

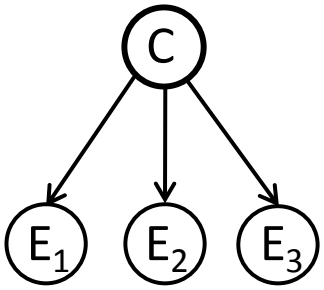

Figure 1: An example of a simple common-cause structure with a cause variable *C* and three effect variables E_1, E_2, E_3. The state of each effect variable depends statistically only on the state of the cause variable.

Rehder and Burnett (2005) developed a reasoning task which allowed for testing people's intuitions about the Markov condition. For example, subjects had to rate the conditional probability of an effect's presence given the state of its cause C. The crucial manipulation was whether other effects of C were present or absent. According to the Markov condition subject's ratings should be invariant across these conditions. Contrary to this prediction, the ratings were clearly sensitive to the states of other effects of C. The more collateral effects were present, the higher the rating of the conditional probability of the target effect given the presence of C. This Markov violation was extremely robust across many cover stories and domains.

Walsh and Sloman (2007) followed up on this research. They were interested in the boundary conditions of violations of the Markov condition. In one experiment they

presented subjects with a common-cause model in which loud music in an apartment building represented the common cause of the complaints of the neighbors on the left and the right side of the apartment in which the music was playing. Again the crucial test question referred to a case in which loud music was playing but the left neighbor was not complaining. According to the Markov condition this should not affect the rating of the likelihood that the right neighbor is complaining. However, Walsh and Sloman reasoned that the likelihood that complaints of the right neighbor are predicted should depend on the ad hoc explanations of why the left neighbor did not complain. If subjects were instructed that all neighbors were invited to the apartment in which music was playing, subjects should expect both neighbors not to complain (i.e., Markov violation). In contrast, when subjects were told that the left neighbor has left the building there is no reason to expect that the second neighbor will not complain (i.e., no Markov violation). The experiments confirmed these predictions although there was a fairly strong tendency to violate the Markov condition in all conditions. In this experiment the difference between the inferences is due to the fact that the initial causal model was differently augmented and changed in the contrasted conditions by adding further causal variables. In one condition an additional causal event, the invitation, was introduced, in the other condition one effect was effectively removed from the model, thus deleting its diagnostic relevance.

Agents and Causes

We are also interested in conditions moderating the degree of Markov violations. Whereas Walsh and Sloman (2007) have shown that different models containing different kinds of disabling events influence the inferences, our goal is to study the influence of assumptions about causal mechanisms while keeping the causal model on the surface level invariant. Causal Bayes nets combine assumptions about causal mechanisms with probabilistic covariations, but the assumed mechanisms are not elaborated. Tellingly, Pearl (2000) describes causal arrows as *mechanism placeholders*. Although recent empirical studies have casted doubt on the assumption that people have elaborate knowledge about mechanisms (Rozenblit & Keil, 2002), recent research on causal reasoning and language understanding has suggested that people may have abstract notions of basic properties of mechanisms (see Talmy, 1988; Wolff, 2007). Particularly relevant in the present context is the distinction between *agents* and *patients*, which is one of the important distinctions in our causal semantics introduced by Talmy. Agents are causal events that we represent as active in the generation of a causal relation. Patients are passive recipients of causal power. For example, in the familiar Michotte task the ball pushing the second ball is viewed as an agent endowed with force, whereas the ball that is being pushed is represented as a patient exerting resistance (White, 2009).

Agents and causes typically fall together but can be separated. Consider the example of tuners that receive music from a music station. Within a causal Bayes net the station would play the role of a common cause because sending out waves precedes the reception by tuners. However, depending on the focus, it is possible to view the sender as active senders and the tuners as passive receivers, or it is possible to highlight the active role of the tuners as receivers without whom no music can be heard. Thus, effects in a common cause model can be agents or patients depending on the framing. Our key prediction is that the agent role is associated with attributions of causal responsibility and blame. If something goes wrong in a causal transmission, then the agent will be the primary target of error attributions.

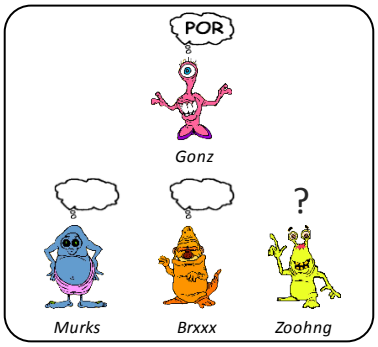

Figure 2: An example of a test item used in Waldmann et al. (2007).

Pilot Study

In an initial experiment, we tested this theory (Waldmann et al., 2007). Subjects were presented with instructions about four aliens, Gonz, Murks, Brxxx, and Zoohng, who mostly think of nothing and sometimes think of "POR" (food in alien language; material adapted from Steyvers et al., 2003). In one condition it was pointed out that Gonz is able to transmit its thoughts into the heads of the other alien (sending condition). In the contrasting condition it was pointed out that Murks, Brxxx, and Zoohng are able to read the thoughts of Gonz (reading condition). So, in both conditions the thoughts of Murks, Brxxx, and Zoohng are statistically and causally dependent on the thoughts of Gonz. Hence, both cases can be represented as a common-cause network (see Fig. 1; Gonz as the cause *C* and Murks, Brxxx, and Zoohng as the effects E_1, E_2, and E_3). However, the agent role was manipulated across conditions. Whereas in the sending condition cause and agent fall together, in the reading condition the effects were framed as agents. In the test phase subjects were requested to rate the conditional probability of a target alien, e.g., Zoohng, thinking of POR given the thoughts of the cause and the other effect aliens (for an example, see Fig. 2). Interestingly, the "Markov violation" was significantly stronger in the sending condition than in the reading condition (see interaction of upper two lines in Fig. 3), which confirms our prediction

that errors are associated with the agent. If there is only one agent (sending condition) then the failure of one of the receiving aliens to read his thoughts becomes diagnostic for a failure of the sending agent that also should affect the other aliens. In contrast, if the effects are represented as agents, then the error attributions should be locally attributed to the respective effect. The failure of one reader to read the thoughts of the cause alien should not predict whether the other readers will also fail or not.

Another important finding of the pilot study which we will follow up in Experiment 1 is that Markov violations in the sending condition were only observed when the cause was present but not when the cause was absent (see lower two lines in Fig. 3). Intuitively this can be interpreted as evidence for the assumption that sending errors can only occur when the cause alien is trying to send.

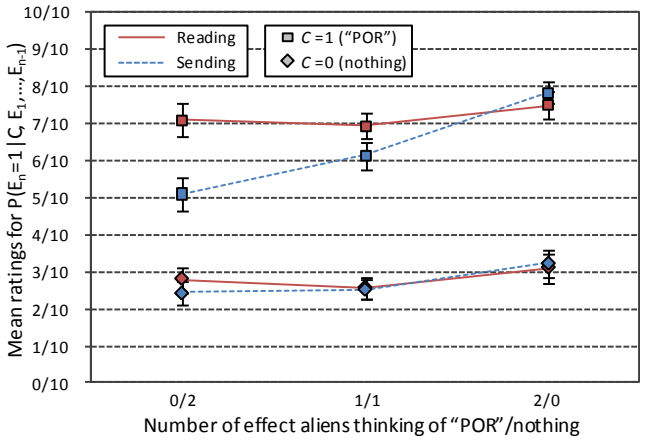

Figure 3: Mean ratings (and standard error) representing the estimates of the relative number of times the target alien thinks of "POR" in ten fictitious situations. The X axis represents the number of collateral effect aliens thinking of "POR". The upper two lines correspond to the cause alien thinking also of "POR", the lower two lines to the cause alien thinking of nothing. The dashed lines indicate the sending condition, whereas the solid lines indicate the reading condition.

In the next section we report a model that captures our intuitions about the role of agents in causal models. Subsequently we will report experiments testing the model.

A Bayes Net Model of Error Attributions

In Bayes nets, errors which are due to hidden mechanisms can be represented by hidden nodes in the network. We propose that *each cause* contains a hidden *common preventive noise* (PN) node which is connected to all effects, and can therefore alter the influence of the causes on their effects. Hence, in common-cause model there is one *PN* attached to its effects. This common noise source summarized *all* influences which potentially decrease the ability of the cause to bring about its effects (e.g., common preventer; missing enabling conditions, etc.)[1] (see Fig. 4). The strength of this noise source (w_{PN}) and its a priori base rate are domain dependent. In the sending condition, we assume that w_{PN} is pre-set to high values, thus increasing the influence of common preventive noise. In the reading condition, people should primarily attribute errors to the error links that are attached to each effect node and that are in Bayes nets assumed to be independent of each other. Thus, different parameterizations of w_{PN} explain the different degrees of Markov violations in the sending versus reading conditions.

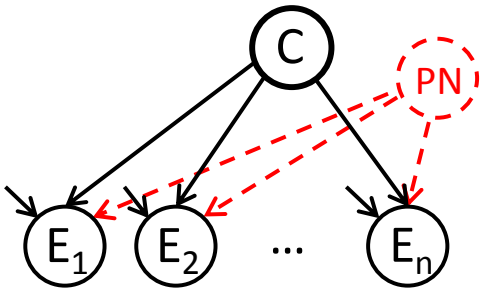

Figure 4: A simple common-cause structure extended by an unobserved common preventive noise node *PN*. The preventive noise interacts with the causal influence of *C*. If *PN* is present the power of *C* is lowered for all its effects. Thus, if $E_1, \ldots, E_{n-1}$ are observed as absent, even if the cause *C* is present, the presence of *PN* is likely. This lowers the predicted probability of E_n being present.

Asking people to judge the probability of a target effect alien (E_n) thinking of POR given the thoughts of the other aliens (C, E_1, …, E_{n-1}) is formally equivalent with asking the conditional probability of E_n: $P(E_n = 1|C, E_1, \ldots, E_{n-1})$. In a regular common cause structure (without a common noise source) this question simplifies to $P(E_n = 1|C)$ due to the Markov condition: The presence of the target effect only depends on the state of the cause, not on the states of the collateral effects. Introducing an unobserved common preventive noise node and integrating it out leads to the following derivation[2]:

$$P(E_n = 1|C, E_1, \ldots, E_{n-1}) = \sum_{PN} P(E_n = 1|C, \boldsymbol{PN}, E_1, \ldots, E_{n-1}) \cdot P(\boldsymbol{PN}|C, E_1, \ldots, E_{n-1}) = \sum_{PN} P(E_n = 1|C, \boldsymbol{PN}) \cdot P(\boldsymbol{PN}|C, E_1, \ldots, E_{n-1})$$

The second simplifying step in this derivation is possible because in the network with the common preventive noise

[1] Note that this is a specific preventive cause which does not affect the probability of the effect when the cause is absent.

[2] Actually, also the prior assumptions of the parameter values given by a set of Beta distributions are integrated out. To simplify the discussion we left this out in the description. The complete derivation includes a multiple integral over the parameter vector: $P(E_n = 1|C, E_1, \ldots, E_{n-1}) = \int P(E_n = 1|C, E_1, \ldots, E_{n-1}, \mathbf{w}) p(w) dw$

node the Markov condition holds: Given C *and* PN the target effect E_n is independent of the collateral effects. Thus, reasoning in this simple model can be thought as a two-step process: First the state of the noise is inferred, and then given that state (and given the a priori state of C) the state of the unobserved target effect is inferred.

The model predicts that inferences about the presence of an unobserved target effect in the presence of the cause should be influenced by the number of collateral effects that are present or absent. Absent effects in the presence of the cause should via the PN lower the ratings for the target effect. This influence should increase with increasing numbers of absent effects when the cause is present. When the cause is absent, however, no such pattern should be observed.

Experiment 1

When the cause varies between present (i.e., active) and absent (i.e., inactive), the model predicts an asymmetric influence of *PN* since in the cause's absence the *PN* cannot prevent *C* to bring about the target effect. Thus, the Markov violation in the sender condition should only be observed when the cause is present. In our pilot experiment we have indeed confirmed this prediction. In contrast, our model predicts a symmetric influence of the *PN* when the cause has two distinct but causally active states (i.e. A/B instead of 0/1). This prediction is tested in Experiment 1.

Method

Participants 56 students from the University of Göttingen participated in exchange for candy.

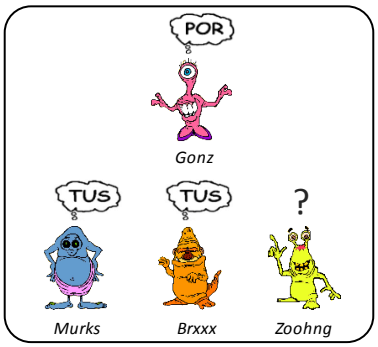

Figure 5: An example of a test item used in Experiment 1.

Procedure and Material In the instruction phase we presented subjects with instruction about four aliens: Gonz, Brxxx, and Zoohng, who usually think of "TUS" and sometimes think of "POR" (indicated by a bubble containing "TUS" or "POR"; see Fig. 5) (POR and TUS were counterbalanced). In two conditions it was either stated that the upper alien can transmit both thoughts to the lower three (sending condition), or that the lower three aliens can read the thoughts of the upper one (reading condition). It was pointed out that the effect aliens frequently think of "POR" or "TUS" when the cause alien thinks of "POR" or "TUS".

In the test phase, subjects were presented with six test panels with all the non-target aliens thinking of either "POR" or "TUS" (for an example, see Fig. 5). The order of test panels was randomized. For each panel, subjects were asked to imagine ten situations with the given configuration, and then to judge in how many of these situations the target alien (indicated by a question mark above its head) would probably think of "POR". This way we obtained probability assessments from the subjects.

Design The predictions were tested in a 2×2×3 ANOVA design with "sending" vs. "reading" as a between-subjects factor. The state of the cause alien ("POR" or "TUS" thoughts) and the number of collateral effect aliens thinking of "POR" (0, 1, or 2) were manipulated within subjects.

Results and Discussion

Figure 6 displays the results for Experiment 1. In general, the ratings for the target effect alien thinking of "POR" were higher when the cause alien thinks of "POR" ($F_{1,54}$=146.05, p<.001, η_p^2=.73).

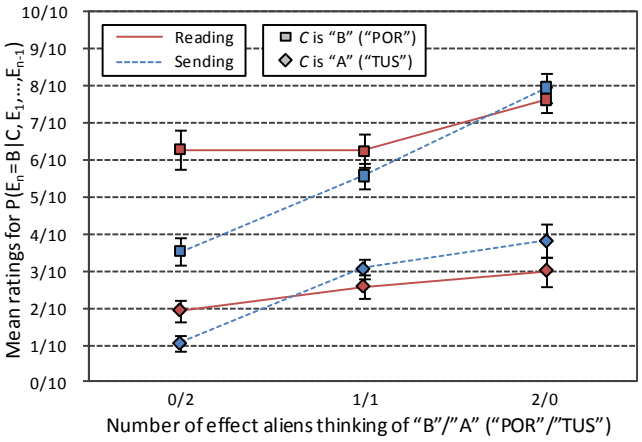

Figure 6: Mean ratings (and standard error) representing the estimates of the relative number of times the target alien thinks of "POR" in ten fictitious situations. The X axis represents the number of collateral effect aliens thinking of "POR". The upper two lines correspond to the cause alien thinking of "POR", the lower two lines to the cause alien thinking of "TUS". The dashed lines indicate the sending condition, whereas the solid lines indicate the reading condition.

As predicted by the model, people's judgments were symmetrically influenced by the states of the other effects for both states of the cause: In case of *C* representing "POR" (the upper two lines in Fig. 6) the ratings substantially increased with the number of effect aliens thinking of "POR" ($F_{2,108}$=31.47, p<.001, η_p^2=.37). As in our pilot study this influence was stronger in the sending condition than in

the reading condition yielding a significant interaction ($F_{2,108}$=8.94, p<.001, η_p^2=.14). In case of *C* representing "TUS" (the lower two lines in Fig. 6) the ratings also increased the more effect aliens thought of "POR" ($F_{2,108}$=20.25, p<.001, η_p^2=.27). As predicted by the model and in contrast to what we observed for the absent state of the cause in our pilot study the influence of the collateral effects was also stronger in the sending condition than in the reading condition when the cause alien thought of "TUS" ($F_{2,108}$=4.20, p<.05, η_p^2=.07). The descriptively weaker two-way interaction in the "TUS" case is predicted by the model as a consequence of the low base rate of "TUS". No three-way interaction was obtained, as predicted ($F_{2,108}$=1.37, p=.26).

The results confirm our model. Subjects' inferences were influenced by the location of the agent (sending vs. reading) in a fashion predicted by the error attribution model. Moreover, the model's predictions about the type of states of binary causal variables were confirmed. Our patterns in the sending condition correspond to the findings of Rehder and Burnett (2005), who also found symmetric influences of the states of other effect variables for both states of the cause. Although Rehder and Burnett described these states as present and absent, the two states in their experiments actually also represented two active states on a continuous dimension (typical vs. atypical).

Experiment 2

In our model, the common preventive noise node *PN* is attached to the specific cause it regulates. Therefore, in a causal chain structure each causal link should have its own *PN* node (see Fig. 7). This entails that the strength of each *PN* in the chain should not bias people's assumptions about the states of other variables. Consequently, our model predicts that in causal chain structures no Markov violation should be observed and that manipulations of people's assumptions about the location of the agent (i.e., sending vs. reading) should not have any effect. This prediction is tested in Experiment 2.

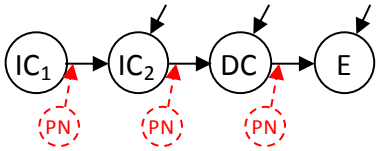

Figure 7: An extended causal chain model with two indirect causes (IC_1, IC_2), a direct cause (*DC*) and a final effect (*E*). Since the preventive noise (*PN*) is part of the causal process and therefore attached to each direct cause-effect relation, in each cause variable has its own preventive noise source.

Method

Participants 50 students from the University of Göttingen participated in exchange for candy.

Procedure and Material As in Experiment 1, we presented subjects with instruction about four aliens: Gonz, Brxxx, and Zoohng, who—as in the basic experiment—usually think of nothing and sometimes think of "POR" (indicated by an empty bubble or a bubble containing "POR", respectively; see Fig. 8). It was pointed out that—in the sending condition—an alien can transmit its "POR"-thoughts to its right neighbor or—in the reading condition—an alien can read the "POR"-thoughts of its left neighbor. Again it was stated that effect aliens frequently think of "POR" when the corresponding cause alien (the left neighbor) also thinks of "POR".

In the test phase subjects were presented with six test panels with the non-target aliens thinking of "POR" or nothing (for an example, see Fig. 8). The order of test panels was randomized. The target alien was generally the right most alien in the chain. As in Experiment 1, subjects were asked to judge in how many of ten situations the target alien would probably think of "POR".

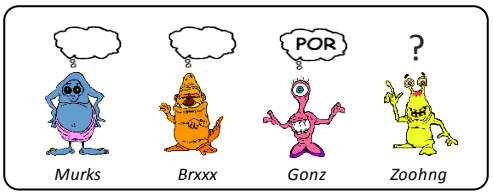

Figure 8: An example of a test item used in Experiment 2.

Design The predictions were tested in a 2×2×3 ANOVA design with "sending" vs. "reading" constituting a between-subjects factor and the state of the direct-cause alien ("POR" or nothing) as well as the number of indirect-cause aliens thinking of "POR" as within-subjects factors (0, 1, or 2).

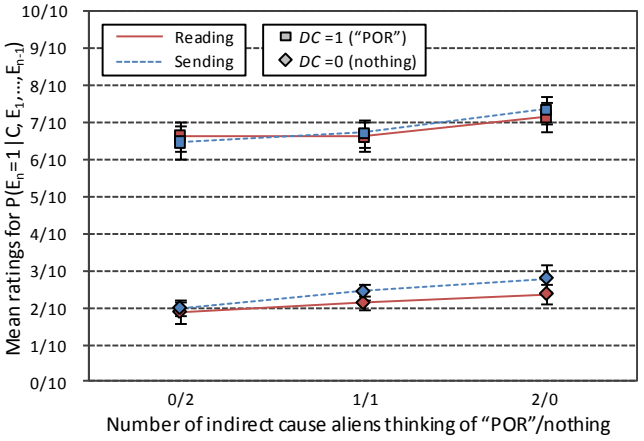

Figure 9: Mean rating (and standard error) of number of times the target aliens thinks of "POR" in ten fictitious situations plotted against the number of indirect-cause aliens thinking of "POR" (columns). The upper two lines correspond to the direct cause (*DC*) alien thinking also of "POR", the lower two lines to the direct cause alien thinking of nothing. The dashed lines represent the sending

condition, whereas the solid lines represent the reading condition.

Results

The results of Experiment 2 are shown in Figure 9. As in Experiment 1, the ratings for the target effect alien thinking of "POR" were higher when the direct-cause alien also thought of POR ($F_{1,48}$=191.99, p<.001, η_p^2=.80).

The prediction that different assumptions about the agents in the chain (sending vs. reading) should not matter was clearly supported. As predicted by the model, the sending vs. reading manipulation revealed no interaction with the states of the non-direct causes, neither in the presence of the direct cause ($F_{2,96}$<1, p=.5) nor in its absence ($F_{2,96}$<1, p=.66). However, in contrast to the predictions, significant, although very weak violations of the Markov condition in both the presence (the upper two lines in Fig. 9; $F_{2,96}$=11.77, p<.001, η_p^2=.20) as well as the absence of the direct causes (the lower two lines in Fig. 9; $F_{2,96}$=6.47, p<.01, η_p^2=.12) could be seen (see also Rehder & Burnett, 2005). The three-way interaction was clearly not significant ($F_{2,96}$<1, p=.99).

Discussion

The results of Experiment 2 show sensitivity to the instructed causal model and support the assumption inherent in our Bayesian model that preventive noise sources are attached to specific causes. Hence, whether preventive noise predicts error correlations is dependent on the underlying causal structure in which these nodes are an intrinsic property of each cause-effect relations.

However, our model cannot account for the small but still significant Markov violations in the data. Possibly subjects doubt that chain variables fully screen off previous influences or there are additional assumptions underlying causal chain representations.

General Discussion

Traditional causal theories view causes as endowed with the power to generate effects. However, little is known about how the mechanisms relating causes and effects are represented, and what influence assumptions about the mechanisms have on causal inferences. We have pinpointed one relevant factor, the distinction between agents and patients which can be separated from the distinction between causes and effects. We have used the example of sending versus reading to disentangle the location of the agent from the location of the cause. Our main hypothesis is that people tend to attribute potential errors to agents rather than patients. This intuition was formalized in a Bayesian model of error attribution which adds hidden preventive noise nodes to capture our intuitions about sources of error. Interestingly, this model explains violations of the Markov condition using a model that honors the Markov condition. Two experiments were conducted which tested and largely confirmed specific predictions of the model.

Traditionally there has been a conflict between covariation and mechanism (or force) theories. The present research shows that it is fruitful to combine the two approaches. Causal models are needed to guide processing of statistical covariations in data. However, the simple assumptions typically underlying these models are insufficient because additional knowledge about the mechanism seems to influence both the assumed hidden and observed structure of the model and the parameterization (see Mayrhofer et al., 2008). Future research will have to further elaborate the intricate relation between mechanism assumptions and causal models.

Acknowledgments

We wish to thank Marie-Theres Kater and Mira Holzer for assistance in data collection, and Noah Goodman, Josh Tenenbaum, and Tom Griffiths for helpful comments on the project. This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (DFG Wa 621/20).

References

Mayrhofer, R., Goodman, N. D., Waldmann, M. R., & Tenenbaum, J. B. (2008). Structured correlation from the causal background. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society* (pp. 303-308).

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

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

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521-562.

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

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*, 49-100.

Waldmann, M. R., Mayrhofer, R., & Hagmayer, Y. (2007). *Mind reading aliens: Causal capacities and the Markov condition*. Unpublished manuscript.

Walsh, C. R., & Sloman, S. A. (2007). Updating beliefs with causal models: Violations of screening off. In M. A. Gluck, J. R. Anderson & S. M. Kosslyn, *A Festschrift for Gordon H. Bower* (pp. 345-358). New York: Erlbaum.

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, *116*, 580-601.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*, 82-111.