

Inferring interventional predictions from observational learning data

BJÖRN MEDER, YORK HAGMAYER, AND MICHAEL R. WALDMANN
University of Göttingen, Göttingen, Germany

Previous research has shown that people are capable of deriving correct predictions for previously unseen actions from passive observations of causal systems (Waldmann & Hagmayer, 2005). However, these studies were limited, since learning data were presented as tabulated data only, which may have turned the task more into a reasoning rather than a learning task. In two experiments, we therefore presented learners with trial-by-trial observational learning input referring to a complex causal model consisting of four events. To test the robustness of the capacity to derive correct observational and interventional inferences, we pitted causal order against the temporal order of learning events. The results show that people are, in principle, capable of deriving correct predictions after purely observational trial-by-trial learning, even with relatively complex causal models. However, conflicting temporal information can impair performance, particularly when the inferences require taking alternative causal pathways into account.

How do people infer the consequences of active interventions in events when only knowledge from passive observations is available? Associative accounts of causal cognition distinguish between two types of causal learning, observational learning which consists of associating observed causal events, and instrumental learning which consists of associating actions with outcomes (e.g., Dickinson, 2001). Thus, a tempting solution as to how people go from passive observations to knowledge of the consequences of active interventions would be to equate observational knowledge with instrumental knowledge and proceed from there. Unfortunately, this strategy will often lead to erroneous inferences and ineffective actions. For example, observing (“seeing”) the status of a barometer enables us to predict the approaching weather, but this does not license the inference that manipulating (“doing”) the barometer will affect the weather. Thus, although associative accounts distinguish between observational and interventional learning, they are prone to errors when predictions for instrumental actions are derived from observational knowledge.

Although both interventional and observational learning have been investigated in a number of studies (e.g., Gopnik et al., 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) very few studies have pursued the question whether people are capable of deriving correct interventional predictions after purely observational learning. Sloman and Lagnado (2005) addressed this question using logical reasoning tasks and demonstrated that participants correctly distinguished between observational and interventional predictions. Waldmann and Hagmayer (2005) went one step further in the direction of learning

(see Blaisdell, Sawa, Leising, & Waldmann, 2006, for an analogous study with rats). Participants were first shown diagrams similar to the one depicted in Figure 1 (without the numbers), which provide representations of *causal models*, that is, hypotheses about the structure of causal systems. Subsequently learners were handed a tabulated list of individual cases on a sheet of paper. The results not only showed that people are indeed sensitive to the distinction between observational and interventional predictions but also that learners took into account the parameters of the causal models (e.g., the strength of the causal relations), which they had gleaned from the presented data.

Although Waldmann and Hagmayer (2005) provided their participants with learning data it could be argued that, due to the presentation of tabulated data, the tasks were more like reasoning than standard learning tasks. Moreover, some researchers have claimed that processing of tabulated data is handled by different learning mechanisms than trial-by-trial learning (e.g., Shanks, 1991). Therefore, one novel feature of the present study is the use of a passive trial-by-trial observational learning paradigm to test whether learners are capable of deriving interventional predictions from passive observations of a set of separated learning trials.

Studying inferences based on trial-by-trial learning also introduces cues to causal structures that might compete with the initially instructed causal model. According to causal-model theory (Waldmann, Hagmayer, & Blaisdell, 2006) people primarily induce hypotheses about causal structures on the basis of cues, such as temporal order or verbal instructions (Lagnado & Sloman, 2006; Lagnado, Waldmann, Hagmayer, & Sloman, 2007). In most

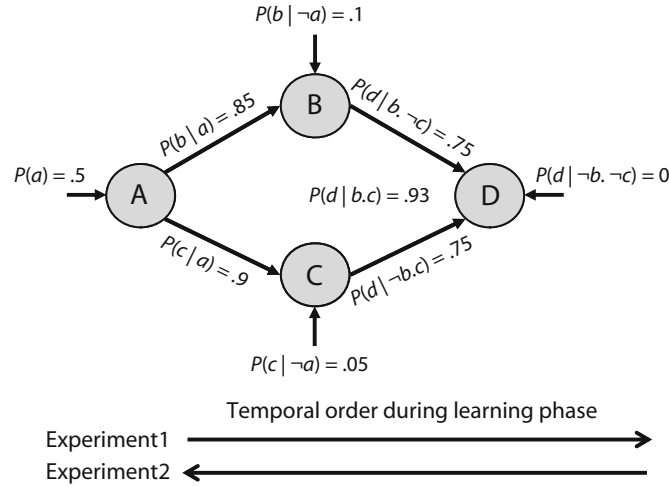


Figure 1. The parameterized causal model used in Experiments 1 and 2. Arrows indicate causal relations between variables; conditional probabilities encode the strength of these relations. All parameters were preset except $P(d|b, c)$, which is computed by a noisy-OR-gate (Pearl, 1988).

of our previous experiments learners were provided with structure hypotheses solely on the basis of initial instructions. However, there are cases in which temporal and causal order mismatch. For example, physicians often observe symptoms (i.e., effects) prior to learning about their causes. In these cases it is crucial that the temporal order of experiencing events is ignored as a cue to causal structure. Therefore, a second novel feature of the experiments involves the manipulation of temporal cues while holding the instructed causal model constant. In Experiment 1 the temporal order during learning conforms to the causal order (predictive learning from causes to effects), whereas in Experiment 2 the temporal order is reversed (diagnostic learning from effects to causes). Thus, in the diagnostic learning condition there is a mismatch between the temporal order implied by the causal model and the experienced learning order. Based on previous research on causal-model theory we expected participants to be capable of letting causal order override learning order, that is, we expected learners to stick to the initially suggested causal structure. However, the diagnostic learning context might influence the acquisition of the causal model's parameters thereby jeopardizing learners' competency to derive interventional predictions.

Seeing Versus Doing:

Modeling Observations and Interventions

Normatively, predictions based on observed values of a variable often differ from inferences drawn from the very same states brought about by active interventions. For example, if we arbitrarily change the reading of a barometer, our action renders the reading independent of its usual cause, atmospheric pressure. Such "strong" interventions (Woodward, 2003) can be formalized by Pearl's (2000) "do-operator," written as $\text{do}(\bullet)$ (see Spirtes, Glymour, & Scheines, 1993, for an alternative notation). Whereas the expression $P(a|c)$ refers to the probability of $A = a$, given that $C = c$ was observed, the expression $P(a|\text{do } c)$ refers

to the probability of $A = a$, given that C is fixed to state c by means of an intervention.¹

Correct predictions about the outcomes of interventions require that learners are sensitive to the structure of causal models and the difference between observations and interventions. In the present experiments we use the diamond-shaped causal model shown in Figure 1. Within this model the observed states of event C provide diagnostic evidence for the state of its cause A , thus $P(a|c) > P(a|\neg c)$. (See the Appendix for the formal derivations.) In contrast, manipulations of effects do not change their causes; hence, the probability of A remains at its base rate when the presence or absence of C is generated by means of an intervention [i.e., $P(a|\text{do } c) = P(a|\text{do } \neg c)$]. The fact that (strong) interventions create independence also implies a difference between observations and interventions in predictive reasoning from C to D . Obviously, there is the direct causal link connecting C to D , but there is also a second cause of D , event B . Observational predictions need to take the alternative pathway $A \rightarrow B \rightarrow D$ into account because observed values of C provide diagnostic evidence for A and, therefore, also include the influence of B on D . For example, observing C to be absent indicates that A , and therefore also B , is likely to be absent. However, the situation is different when C is not merely observed to be absent, but is actively prevented from occurring. Although this intervention ensures that event D is not influenced by C , the model's initial event A will still occur with its base rate probability and influence D by way of B . As a consequence, the probability of the final effect D is lower when C is merely observed to be absent than when the occurrence of C is prevented by means of an intervention [i.e., $P(d|\neg c) < P(d|\text{do } \neg c)$].

EXPERIMENT 1

The goal of the first experiment was to investigate whether learners are capable of deriving predictions for

hypothetical interventions after an observational trial-by-trial learning phase in which learning order corresponds to causal order.

Method

Participants and Design

Twenty-four students from the University of Göttingen, Germany, participated for course credit. The factors *intervention vs. observation* and *presence vs. absence of C* were varied within subjects.

Materials and Procedure

Causal model instructions. The variables of the causal model depicted in Figure 1 were introduced as four fictitious chemical substances causally interacting in wine casks. In addition, participants were shown a graphical representation of the causal structure similar to Figure 1 (without information about the model's parameters). They were instructed to learn about the strength of the causal relations from the learning data shown in Table 1, which embodied the probabilities shown in Figure 1. The kind of questions learners would have to answer after the learning phase was not mentioned until the test phase. Learners did not see the figure showing the model during either the learning or test phase.

Observational learning phase. The learning phase consisted of 40 trials presenting information about the states of the four variables on a computer screen, with each trial referring to a different wine cask. Each chemical substance was represented by a circle with the label of the corresponding substance; the circles were spatially arranged as a diamond-shaped form but without arrows (see Figure 1). At the beginning of each trial, all four circles were labeled with question marks indicating that the variables' states in the wine cask were not yet known. Then temporally ordered information about the presence and absence of the four variables was given. The presence of a chemical substance was signaled by a colored circle, its absence by a crossed-out circle. First, information about variable *A* was presented; then, simultaneously, variables *B* and *C* were shown; and, finally, information about event *D* was given. The interstimulus interval was 1 sec; the whole pattern stayed for another 2 sec on the screen before the next trial automatically began. Participants passively observed the unfolding events without making overt predictions.

Test phase. The learning phase was followed by a test phase in which participants were requested to answer four questions about hypothetical observations and four questions about hypothetical interventions. For the two observational diagnostic questions, participants were instructed to imagine observing the presence [absence] of substance *C* in 40 previously unseen wine casks and to estimate the number of casks in which substance *A* would also be found, that is, they had to estimate $P(a|c)$ and $P(a|\neg c)$ in a conditional frequency format. The two interventional diagnostic questions asked learners first to imagine that substance *C* was added to 40 casks, or that *C* was prevented in 40 casks, and then to estimate the number of casks in which substance *A* would also be found [i.e., participants estimated $P(a|do\ c)$ and $P(a|do\ \neg c)$]. The same set of questions was asked about *D*, the effect of *C*; that is, learners were requested to estimate the predictive probabilities $P(d|c)$, $P(d|\neg c)$, $P(d|do\ c)$,

Table 2
Conditional Frequency Estimate Results Referring
to 40 Cases in Experiment 1 (Predictive Learning)
and Experiment 2 (Diagnostic Learning)

	Causal Model Predictions	Experiment 1		Experiment 2	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Diagnostic Inferences					
Observation					
$P(a c)$	38	30.5	7.6	33.5	8.6
$P(a \neg c)$	4	17.1	10.4	15.4	11.2
Intervention					
$P(a \text{do } c)$	20	25.5	10.6	25.5	11.2
$P(a \text{do } \neg c)$	20	27.3	8.6	22.4	10.3
Predictive Inferences					
Observation					
$P(d c)$	36	29.7	10.0	30.5	10.1
$P(d \neg c)$	5	14.8	11.6	18.3	13.3
Intervention					
$P(d \text{do } c)$	33	27.5	11.6	29.7	10.9
$P(d \text{do } \neg c)$	14	21.6	12.6	20.3	11.7

Note—For the derivation of the causal model predictions, see the Appendix.

and $P(d|do\ \neg c)$. Interventional and observational questions were blocked; the order of blocks was counterbalanced.

Results and Discussion

Table 2 shows the means of the conditional frequency estimates along with the values derived by a causal model analysis (see the Appendix).

Diagnostic Inferences

Whereas observed states of *C* are diagnostic for its cause *A*, generating or preventing *C* by means of intervention renders the event independent of *A*. Therefore, the observational probabilities should differ, whereas the interventional probabilities should stay at a constant level. To test these predictions, we conducted a number of planned within-subjects comparisons. Consistent with the predictions participants' mean estimates for the two observational probabilities differed [$F(1,23) = 36.51$, $MS_e = 59.17$, $p < .001$], but they judged the interventional probabilities to be at the same level ($F < 1$). Moreover, the estimates for the observational and interventional probabilities differed: $P(a|c)$ received higher estimates than $P(a|do\ c)$ did [$F(1,23) = 4.61$, $MS_e = 63.93$, $p < .05$]. Conversely, $P(a|\neg c)$ received lower estimates than $P(a|do\ \neg c)$ did [$F(1,23) = 21.03$, $MS_e = 58.99$, $p < .001$]. Although participants' estimates did not perfectly match the quantitative causal model predictions, the qualitative pattern of the results provides clear evidence for participants' sensitivity to the difference between seeing and doing in diagnostic judgments.

Predictive Inferences

Within the chosen causal model predictive inferences are more complicated than the requested diagnostic inferences. Whereas the latter only require considering the direct causal relation between *A* and *C*, the inferences concerning variable *D* require taking into account the complete model, in particular the alternative causal pathway $A \rightarrow B \rightarrow D$. If *C* is observed to be present, then its cause *A*

Table 1
Learning Data of Experiments 1 and 2 (40 trials)

Data Pattern	Frequency	Data Pattern	Frequency
<i>a. b. c. d</i>	14	$\neg a. b. c. d$	0
<i>a. b. c. \neg d</i>	1	$\neg a. b. c. \neg d$	0
<i>a. \neg b. c. d</i>	2	$\neg a. \neg b. c. d$	1
<i>a. \neg b. c. \neg d</i>	1	$\neg a. \neg b. c. \neg d$	0
<i>a. b. \neg c. d</i>	2	$\neg a. b. \neg c. d$	1
<i>a. b. \neg c. \neg d</i>	0	$\neg a. b. \neg c. \neg d$	1
<i>a. \neg b. \neg c. d</i>	0	$\neg a. \neg b. \neg c. d$	0
<i>a. \neg b. \neg c. \neg d</i>	0	$\neg a. \neg b. \neg c. \neg d$	17

and, therefore, the second causal pathway is likely to be instantiated. Therefore, large differences between the absence and presence of C are to be expected. In contrast, an intervention on C leaves the second pathway untouched; hence a smaller difference is to be expected for the intervention questions. As can be seen from Table 2, participants were indeed sensitive to these differences. As shown by the significant interaction contrast, the difference between the responses to the observational questions proved larger than the difference between the responses to the interventional questions [$F(1,23) = 8.73$, $MS_e = 54.65$, $p < .01$]. A second crucial test is provided by the comparison of $P(d|\neg c)$ and $P(d|\text{do } \neg c)$. Consistent with the causal model predictions, participants judged the probability of the occurrence of D to be significantly higher when C was prevented by an intervention than when it was merely observed to be absent [$F(1,23) = 9.57$, $MS_e = 57.83$, $p < .01$]. In accordance with the parameterization of the causal model, there was only a slight, nonsignificant difference between $P(d|c)$ and $P(d|\text{do } c)$ [$F(1,23) = 1.0$, $MS_e = 53.75$, $p = .33$]. This test is important, as there might have been a general tendency to answer interventional questions differently from observational questions. In summary, both diagnostic and predictive inference reflected a remarkable grasp of the observational and interventional inferences afforded by the presented causal model.

EXPERIMENT 2

The main goal of Experiment 2 was to test whether people access causal models adequately when learning order does not match causal order. The same experimental design, cover story, and instructions were used as in Experiment 1.

Method

Again, 24 participants from the University of Göttingen participated; none of them had taken part in Experiment 1. In contrast to Experiment 1, the temporal order of learning events did not match their causal order (i.e., diagnostic learning from effects to causes). In each trial, participants were first informed about the status of effect D , then simultaneously about the mediating variables B and C , and finally about the initial cause A . As in Experiment 1, learners were requested to estimate the conditional frequencies of A and D , given observations of or interventions in C .

Results and Discussion

Diagnostic inferences. As can be seen from Table 2, learners' responses to the diagnostic inference questions closely resemble the ones in Experiment 1. The observational questions differed significantly [$F(1,23) = 63.88$, $MS_e = 61.43$, $p < .001$], but there was no difference between the interventional questions [$F(1,23) = 1.52$, $MS_e = 75.26$, $p = .23$]. As in Experiment 1, learners' estimates for the observational and interventional probabilities differed: Whereas $P(a|c)$ received higher estimates than $P(a|\text{do } c)$ did [$F(1,23) = 21.28$, $MS_e = 35.72$, $p < .001$], $P(a|\neg c)$ was judged lower than $P(a|\text{do } \neg c)$ was [$F(1,23) = 13.15$, $MS_e = 45.24$, $p < .01$]. The diagnostic inferences show again a remarkable grasp of the difference between seeing and doing despite the added complexity entailed by the diagnostic learning procedure.

Predictive inferences. As in Experiment 1, participants were asked to estimate the probability of D when C was observed or manipulated by an external intervention. However, in contrast to Experiment 1, the difference between the observational questions proved statistically equivalent to the difference between the interventional questions ($F < 1$). Moreover, in contrast to Experiment 1 the crucial test between $P(d|\neg c)$ and $P(d|\text{do } \neg c)$ failed to reach significance ($F < 1$). Thus, there was no evidence that participants correctly differentiated between seeing and doing in the more complex predictive task.

GENERAL DISCUSSION

Taken together, the results of the two experiments provide convincing evidence that learners are capable of deriving correct predictions for hypothetical interventions after trial-by-trial observational learning. Thus, the competency discovered in previous studies is clearly not restricted to causal reasoning based on tabulated data (Waldmann & Hagmayer, 2005) or descriptions of causal situations (Sloman & Lagnado, 2005). The present findings strongly support the role of causal models in causal learning, and demonstrate a competency that goes beyond the expressive power of associative theories.

Although the qualitative pattern of responses conforms to the predictions of the causal model analyses, the quantitative estimates were not perfect, of course. For example, learners had difficulties with correctly assessing cases in which events were observed to be absent [e.g., when estimating $P(a|\neg c)$]. We think that these deviations are mainly rooted in imperfect parameter estimation processes. After observing the results of Experiment 1, in Experiment 2 we additionally asked participants to estimate the base rate of event A after they had answered the observational and interventional questions. Interestingly, participants consistently overestimated this parameter: Whereas the actual value of $P(a)$ was .5, learners' mean probability judgment was .65. Since, normatively, the posterior probability $P(a|\neg c)$ increases with an increase of $P(a)$, such an overestimation is likely to have contributed to the quantitative mismatch.

In a similar vein, we believe that the impaired performance regarding the predictive inferences in Experiment 2 is rooted in the parameter estimation processes. For example, participants observed the probability of B and C given D but had to infer the probability of D given B and C as a parameter of the causal model. Therefore, the learning process may have led to inadequate estimates of the model's parameters, which is particularly likely to affect the complex predictive inferences. Whereas the diagnostic questions could be correctly answered by recognizing that interventions render the manipulated variables independent of their actual causes, correct estimates for the predictive inferences require considering all of the model's parameters. Thus, if the parameters are not acquired correctly during learning the inferences are likely to be wrong.

One reviewer raised the question whether our results demonstrated learning of parameters or could be predicted on the basis of knowledge about the causal struc-

ture alone. One finding casting doubt on this hypothesis is the difference between Experiment 1 and 2 which shows that, despite of identical structures, the arrangement of the learning input matters. We also have completed further experiments (unpublished) in which we varied the size of the model's parameters which systematically influenced learners' inferences (cf. Waldmann & Hagmayer, 2005).

There are also open questions which future studies need to address. For example, one important question is how people exactly learn to parameterize causal models, particularly when data is presented in a trial-by-trial learning procedure. Recent work on dynamical models of causal learning has begun to address this issue (e.g., Danks, Griffiths, & Tenenbaum, 2003). Another question is how knowledge about causal structures guides causal reasoning. One idea is that people use causal model representations to run mental simulations of real-world situations. By mirroring the causal features of the represented domain, such simulations constrain people's intuitive causal reasoning and allow them to derive causal judgments without the computationally demanding processes implied by a quantitative Bayesian analysis (Hagmayer & Waldmann, 2000).

In summary, the present studies provide clear evidence for the use of causal model representations in causal learning and causal reasoning and challenge associative accounts. Future research will have to develop psychological models that integrate competence and performance in causal learning and causal reasoning.

AUTHOR NOTE

The present experiments formed part of the doctoral dissertation of the first author, which was supervised by the last author. Portions of this research were presented at the 2005 meeting of the Cognitive Science Society in Stresa, Italy. We thank D. Danks, D. Lagnado, B. Spellman, M. v. Sydow, and one anonymous reviewer for many helpful comments on an earlier version of this article. Correspondence concerning this article should be addressed to Björn Meder, Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany (e-mail: bmeder@uni-goettingen.de).

REFERENCES

- BLAISDELL, A. P., SAWA, K., LEISING, K. J., & WALDMANN, M. R. (2006). Causal reasoning in rats. *Science*, **311**, 1020-1022.
- DANKS, D., GRIFFITHS, T. L., & TENENBAUM, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 67-74). Cambridge, MA: MIT Press.
- DAWID, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, **70**, 161-189.
- DICKINSON, A. (2001). Causal learning: An associative analysis. *Quarterly Journal of Experimental Psychology*, **54B**, 3-25.
- GOPNIK, A., GLYMOUR, C., SOBEL, D. M., SCHULZ, L. E., KUSHNIR, T., & DANKS, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, **111**, 3-32.
- HAGMAYER, Y., & WALDMANN, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 214-219). Mahwah, NJ: Erlbaum.
- LAGNADO, D. A., & SLOMAN, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **32**, 451-460.
- 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). Oxford: Oxford University Press.
- PEARL, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.
- PEARL, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
- SHANKS, D. R. (1991). On similarities between causal judgments in experienced and described situations. *Psychological Science*, **5**, 341-350.
- SLOMAN, S. A., & LAGNADO, D. A. (2005). Do we "do"? *Cognitive Science*, **29**, 5-39.
- SPIRITES, P., GLYMOUR, C., & SCHEINES, P. (1993). *Causation, prediction, and search*. New York: Springer.
- STEYVERS, M., TENENBAUM, J. B., WAGENMAKERS, E. J., & BLUM, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, **27**, 453-489.
- WALDMANN, M. R., & HAGMAYER, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 216-227.
- 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.
- WOODWARD, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

NOTE

1. A more general representation of outside interventions within causal models is provided by augmenting causal model representations with intervention nodes representing additional cause variables (cf. Dawid, 2002). Within this framework, predictions of the outcomes of different types of interventions can be modeled as probabilistic inferences (so-called "explaining away," cf. Pearl, 1988). However, to simplify the derivations in this article, here we use the do-operator.

APPENDIX

The probability distribution associated with the model shown in Figure 1 can be factored by applying the causal Markov condition (Spirtes et al., 1993; Pearl, 2000) to the causal model:

$$P(A.B.C.D) = P(A) \cdot P(B|A) \cdot P(C|A) \cdot P(D|B.C). \quad (A1)$$

Modeling Observations

Based on the decomposed probability distribution, the probabilities implied by the observational data can be computed. For example, a diagnostic inference from C to A is computed using Bayes rule:

$$P(a|c) = \frac{P(c|a) \cdot P(a)}{P(c|a) \cdot P(a) + P(c|\neg a) \cdot P(\neg a)} = \frac{P(c|a) \cdot P(a)}{P(c)}, \quad (A2)$$

and

$$P(a|\neg c) = \frac{P(\neg c|a) \cdot P(a)}{P(\neg c|a) \cdot P(a) + P(\neg c|\neg a) \cdot P(\neg a)} = \frac{P(\neg c|a) \cdot P(a)}{P(\neg c)}. \quad (A3)$$

A more complex example is the prediction of variable D from observations of event C :

$$\begin{aligned} P(d|c) &= \sum_i P(A_i|c) \cdot P(B_i|A_i) \cdot P(d|B_i.c) \\ &= P(a|\neg c) \cdot P(b|a) \cdot P(d|b.\neg c) + P(a|\neg c) \cdot P(\neg b|a) \cdot P(d|\neg b.\neg c) \\ &\quad + P(\neg a|\neg c) \cdot P(b|\neg a) \cdot P(d|b.\neg c) + P(\neg a|\neg c) \cdot P(\neg b|\neg a) \cdot P(d|\neg b.\neg c), \end{aligned} \quad (A4)$$

and

$$\begin{aligned} P(d|\neg c) &= \sum_i P(A_i|\neg c) \cdot P(B_i|A_i) \cdot P(d|B_i.\neg c) \\ &= P(a|\neg c) \cdot P(b|a) \cdot P(d|b.\neg c) + P(a|\neg c) \cdot P(\neg b|a) \cdot P(d|\neg b.\neg c) \\ &\quad + P(\neg a|\neg c) \cdot P(b|\neg a) \cdot P(d|b.\neg c) + P(\neg a|\neg c) \cdot P(\neg b|\neg a) \cdot P(d|\neg b.\neg c). \end{aligned} \quad (A5)$$

By conditionalizing A on C , these computations take into account that observed states of C are diagnostic for the state of A . The probability of the final effect D reflects both the influence of B and C .

Modeling Interventions

Pearl's (2000) "do-operator" provides a formal means for representing the notion of strong interventions that fix the value of the target variable. For example, an intervention in C renders the event independent of its actual cause A , therefore,

$$P(a|\text{do } c) = P(a|\text{do } \neg c) = P(a). \quad (A6)$$

The probability of $D = d$ given that C is generated (do c) or inhibited (do $\neg c$) by means of an intervention, is given by

$$\begin{aligned} P(d|\text{do } c) &= \sum_i P(A_i) \cdot P(B_i|A_i) \cdot P(d|B_i.c) \\ &= P(a) \cdot P(b|a) \cdot P(d|b.c) + P(a) \cdot P(\neg b|a) \cdot P(d|\neg b.c) \\ &\quad + P(\neg a) \cdot P(b|\neg a) \cdot P(d|b.c) + P(\neg a) \cdot P(\neg b|\neg a) \cdot P(d|\neg b.c), \end{aligned} \quad (A7)$$

and

$$\begin{aligned} P(d|\text{do } \neg c) &= \sum_i P(A_i) \cdot P(B_i|A_i) \cdot P(d|B_i.\neg c) \\ &= P(a) \cdot P(b|a) \cdot P(d|b.\neg c) + P(a) \cdot P(\neg b|a) \cdot P(d|\neg b.\neg c) \\ &\quad + P(\neg a) \cdot P(b|\neg a) \cdot P(d|b.\neg c) + P(\neg a) \cdot P(\neg b|\neg a) \cdot P(d|\neg b.\neg c). \end{aligned} \quad (A8)$$

In contrast to the computations underlying the observational inferences, in the interventional inferences variable A is no longer conditionalized on C . On the right-hand side of the equations, only parameters that can be derived from observational learning are involved.