Mem Cogn (2016) 44:469–487
DOI 10.3758/s13421-015-0568-5

Transitive reasoning distorts induction in causal chains

Momme von Sydow[1,4,5] · **York Hagmayer**[2] · **Björn Meder**[3]

Published online: 30 November 2015

Abstract A probabilistic causal chain $A \rightarrow B \rightarrow C$ may intuitively appear to be transitive: If A probabilistically causes B, and B probabilistically causes C, A probabilistically causes C. However, probabilistic causal relations can only guaranteed to be transitive if the so-called Markov condition holds. In two experiments, we examined how people make probabilistic judgments about indirect relationships $A \rightarrow C$ in causal chains $A \rightarrow B \rightarrow C$ that violate the Markov condition. We hypothesized that participants would make transitive inferences in accordance with the Markov condition although they were presented with counterevidence showing intransitive data. For instance, participants were successively presented with data entailing positive dependencies $A \rightarrow B$ and $B \rightarrow C$. At the same time, the data entailed that A and C were statistically independent. The results of two experiments show that transitive reasoning via a mediating event B influenced and distorted the induction of the indirect relation between A and C. Participants' judgments were affected by an interaction of transitive, causal-model-based inferences and the observed data. Our findings support the idea that people tend to chain individual causal relations into mental causal chains that obey the Markov condition and thus allow for transitive reasoning, even if the observed data entail that such inferences are not warranted.

Keywords Transitivity · Causality · Markov condition · Mixing of causal relationships · Probabilistic reasoning · Causal induction · Knowledge-based induction · Causal learning · Causal coherence · Transitive distortion effects · Categorization

Transitive reasoning enables judgments about unobserved relationships based on indirect evidence. If one observes that object A is heavier than object B, and that B is heavier than C, one can infer that A is heavier than C. Not all relations, however, are transitive. If A is the mother of B, and B is the mother of C, this does not mean that A is the mother of C.

We investigate whether and to what extent people reason transitively about causal relations, even when the conditions for transitive inferences do not hold true. We focus on probabilistic causal chains of the type $A \rightarrow B \rightarrow C$, where individual relations $A \rightarrow B$ and $B \rightarrow C$ can be combined to form a chain $A \rightarrow B \rightarrow C$ to make probabilistic inferences from the chain's initial event A to the terminal event C. First, we specify the conditions under which transitive reasoning in causal chains is valid. We then report the findings of two experiments investigating whether people make transitive inferences even when the available data entail that such inferences are not warranted.

Our research builds on the idea that people represent the world in terms of mental causal models (Sloman, 2005; Waldmann, 1996; Waldmann, Cheng, Hagmayer, & Blaisdell, 2008; Waldmann & Hagmayer, 2001) that share key characteristics with causal Bayes nets, which originated in philosophy and machine learning (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). The key question of the present research is whether judgments about indirect relations in

✉ Momme von Sydow
Momme.von-Sydow@uni-heidelberg.de; Momme.von-Sydow@lrz.uni-muenchen.de

[1] Department of Psychology, Ruprecht-Karls-Universität Heidelberg, Hauptstr. 47, 69117 Heidelberg, Germany

[2] Department of Psychology, University of Göttingen, Göttingen, Germany

[3] Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany

[4] Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Heidelberg, Germany

[5] Munich Center for Mathematical Philosophy (MCMP), University of Munich, Munich, Germany

causal chains are influenced by transitive reasoning even when the data show that the chain is intransitive.

Transitive reasoning in causal chains

When is transitive reasoning in causal chains valid? Consider a researcher conducting two studies with knockout mice to investigate the causal relations between a certain gene, the level of a particular neurotransmitter, and a behavioral phenotype. The first study finds that knockout mice tend to have an elevated neurotransmitter level,

$$P(\textit{elevated transmitter}|\textit{knockout mice}) > P(\textit{elevated transmitter}|\textit{normal mice})$$

The second study finds that an elevated transmitter level raises the probability of showing anxious behavior, that is,

$$P(\textit{anxiety}|\textit{elevated transmitter}) > P(\textit{anxiety}|\textit{no elevated transmitter})$$

Given these findings, what can be concluded regarding the relation between gene and behavior? A transitive inference may seem intuitively plausible, namely, inferring that knockout mice are more likely to show anxiety than normal mice, that is,

$$P(\textit{anxiety}|\textit{knockout mice}) > P(\textit{anxiety}|\textit{normal mice})$$

Note, however, that neither study has directly assessed this indirect relation. Rather, the two direct relations are integrated into a causal chain that guides the inference about the indirect relation.

One way to formalize causal chains is to use causal Bayes net theory (Pearl, 2000; Spirtes et al., 1993). The framework couples directed acyclic graphs with probability distributions, with the directed edges representing the causal dependencies between the domain variables. A central assumption of the causal Bayes net framework is the *causal Markov condition*, which states that a variable conditioned on its direct causes is independent of all other variables in the causal network, except for its causal descendants (Hausman & Woodward, 1999, 2004; Pearl, 2000; Spirtes et al., 1993; Spohn, 2001). It follows from the Markov condition that the probability distribution over the variables in the causal graph factors such that

$$P(X_i) = \prod_{X_i \in X} P(X_i|pa(X_i)), \quad (1)$$

where the joint probability of variables, $P(X_i) = P(X_1, \ldots X_n)$, is equal to the product of the probabilities of these variables, conditional on $pa(X_i)$, which denotes the set of direct causes of variables X_i in the graph (see, e.g., Hausman & Woodward, 1999, p. 531ff.). Applying the Markov condition to a causal chain $A \rightarrow B \rightarrow C$ entails A and C being conditionally independent given B, that is, $P(C|B \wedge A) = P(C|B \wedge \neg A)$ and $P(C|\neg B \wedge A) = P(C|\neg B \wedge \neg A)$. In other words, the probability of C depends only on the state of variable B and not on the state of variable A.

Importantly, if the Markov condition holds, the conditional probability $P(C|A)$ can be calculated from the parameters of the two direct causal relations with Eq. 2:

$$P(C|A) = P(B|A)P(C|B) + P(\neg B|A)P(C|\neg B). \quad (2)$$

We refer to such inferences about indirectly related events in causal chains as *transitive inferences*. Formally, probabilistic transitive inferences in chains are valid if the Markov condition holds (Bonnefon, Da Silva Neves, Dubois, & Prade, 2012). If the Markov condition does not hold, probabilities inferred through transitive inferences via Eq. 2 may deviate from the actual relations in the data.

Markov violations and category-based transitive inference

The Markov condition enables transitive causal inferences. Its normative and descriptive status, however, is highly disputed. Some advocates of Bayes nets have defended the Markov condition as being a universal characteristic of causal relations in the world or of their representations (Hausman & Woodward, 1999, 2004; Spohn, 2001). Others have criticized the condition's ontological or epistemological necessity (Arntzenius, 2005; Cartwright, 2001, 2002, 2006, 2007; Sober, 1987, 2001; Sober & Steel, 2012; Steel, 2006). However, even advocates of a universal Markov condition concede that the latter can be violated psychologically when (i) the event categories used are inadequate, (ii) there is a mismatch between causal representations and the true causal structure, or (iii) hidden external variables are correlated (Hausman & Woodward, 1999, 2004; Spohn, 2001). In short, both advocates and critics agree that the Markov condition can be violated in practice when descriptions of the world are incomplete or inadequate.

In this paper, we focus on situations in which the categories used to classify the instances involved in the causal relations result in causal chains that violate the Markov condition and also do not warrant transitive inferences. Consider again the example of the causal relations between a gene, a neurotransmitter level, and a behavioral phenotype. Assume that six mice serve as a sample: three knockout mice and three normal mice (A vs. $\neg A$). The neurotransmitter level (e.g., elevated vs. normal, B vs. $\neg B$) and the anxiety level (e.g., high vs. low, C vs. $\neg C$) of each mouse are assessed. Hence, the six mice are classified according to their genetic makeup, their

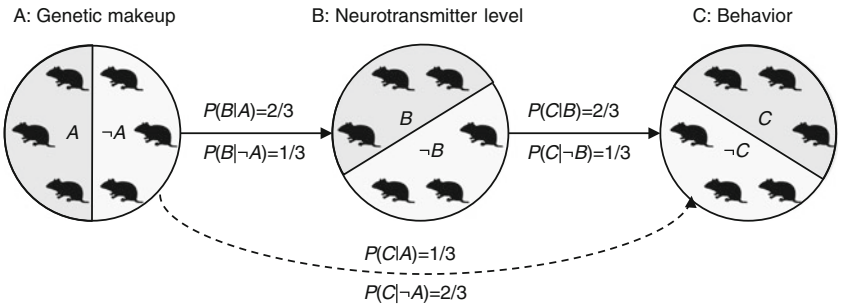

Fig. 1 Causal chain $A \rightarrow B \rightarrow C$ (gene → neurotransmitter → behavior) with positive relations between A and B and between B and C (solid arrows), but a negative relation between A and C (dashed arrow). Each symbol denotes an individual mouse (or mouse population), which differ in the presence versus absence of the three variables

neurotransmitter level, and their anxiety level. A plausible hypothesis might be that the gene causally influences a mouse's anxiety through regulating the neurotransmitter level. Figure 1 illustrates these temporally ordered variables and the parameters estimated from the data. Knockout mice are more likely to have an elevated neurotransmitter level than normal mice, $P(B|A) = 2/3 > P(B|\neg A) = 1/3$. Also, mice with an elevated transmitter level are more likely to be anxious than mice with a normal transmitter level, $P(C|B) = 2/3 > P(C|\neg B) = 1/3$. Despite these two positive relations, however, it does not hold that knockout mice are more likely to be anxious than normal mice. In fact, knockout mice are *less* likely to be anxious than normal mice, $P(C|A) = 1/3 < P(C|\neg A) = 2/3$.[1]

Whereas the data indicate a negative relation between gene and behavior, a different conclusion results from transitive reasoning. Inferring the probability $P(C|A)$ from the parameters of the chain's direct relations via Eq. 2 yields erroneous estimates of $P(C|A) = .56$ and $P(C|\neg A) = .44$, indicating that knockout mice are more likely to show high anxiety levels than normal mice. This discrepancy results from the causal chain being intransitive due to a violation of the Markov condition. Inferring $P(C|A)$ via Eq. 2 is normally only valid if A and C are independent conditional on B, which is not the case, as $P(C|B \wedge A) = .5$, but $P(C|B \wedge \neg A) = 1$.

In this example the Markov condition is violated at the category level due to mixing heterogeneous items (or subclasses) with varying deterministic relationships (Cartwright, 2001; Hausman & Woodward, 1999; Spirtes et al., 1993). For instance, for a mouse (here symbolized by a circle) in the top left corner of Fig. 1, the relations $A \rightarrow B \rightarrow C$ and $A \rightarrow C$ hold. Conversely, for the mouse in the bottom left corner, $A \rightarrow \neg B \rightarrow \neg C$ and $A \rightarrow \neg C$ hold. Thus, on the item level there is no discrepancy between the direct relations $A \rightarrow B$ and $B \rightarrow C$ and the indirect relation $A \rightarrow C$. Causal relations, however, are typically defined at the level of categories, for example, mice that have gene A versus mice that do not. On this level, however, the causal chain $A \rightarrow B \rightarrow C$ is intransitive. The key question of our studies is whether people make transitive inferences on the category level even when the observed causal chain is intransitive.

[1] Likewise, causal strength measures such as ΔP (Allan, 1980; Jenkins & Ward, 1965; cf. White, 2003) or causal power (Cheng, 1997; Griffiths & Tenenbaum, 2005; Meder, Mayrhofer, & Waldmann, 2014) indicate that the individual links are positive, whereas the causal strength for the indirect relation of A and C is negative when marginalizing over B.

Previous research on transitive inferences in causal chains

Previous research on causal chains has shown that people make transitive inferences from A to C when observing only relations $A \rightarrow B$ and $B \rightarrow C$ (Ahn & Dennis, 2000; Baetu & Baker, 2009). Ahn and Dennis (2000) used a sequential learning paradigm, providing participants with evidence on the covariation of events A and B (fertilizer→level of chemicals in soil) intermixed with evidence about events B and C (chemicals→blooming of flower). They additionally studied trial-order effects by varying whether positive or negative evidence for local contingencies (between A and B, and between B and C) was presented first. Participants received no data on the relation between fertilization (A) and blooming (C) but were asked to judge this indirect relation. Average causal judgments were positive, with higher judgments in the positive-evidence-first condition. A primacy effect when constructing the local relations in conjunction with transitive reasoning may explain this.

Baetu and Baker (2009) investigated the influence of transitive reasoning with chains in more detail. In their studies, positive, negative, and zero contingencies for $A \rightarrow B$ and $B \rightarrow C$ were combined. Participants learned about the contingency between two lights A and B (while light C was covered), and between lights B and C (while A was covered); trials occurred intermixed. Subsequently, participants first rated the global $A \rightarrow C$ relationship and then the local relationships $A \rightarrow B$ and $B \rightarrow C$. Baetu and Baker used $\Delta P = P(\text{effect}|\text{cause}) - P(\text{effect}|\neg\text{cause})$ as a measure of causal strength. If the causal Markov condition holds, $\Delta P_{A \rightarrow C} = \Delta P_{A \rightarrow B} \times \Delta P_{B \rightarrow C}$ (Baetu & Baker, 2009, Appendix). Although participants never observed the contingency $A \rightarrow C$, their judgments were consistent

with a multiplication of the individual contingencies, indicating transitive reasoning.

Goals and scope

The studies of Ahn and Dennis (2000) and Baetu and Baker (2009) indicate that in the absence of direct evidence on the relation $A \rightarrow C$, people make transitive inferences as if they are inducing causal chains that obey the Markov condition. In these studies, no contradictory data were available, so it seems like a reasonable default assumption for learners.

Our research goes beyond these studies by investigating reasoning with intransitive chains in which the Markov condition is violated. Intransitive chains provide a stronger test for the hypothesis that people integrate the links into causal models and tend to reason causally coherent, deviating from the data, as if the Markov condition holds (causal coherence hypothesis). Also, we did not use a sequential learning task (Ahn & Dennis, 2000; Baetu & Baker, 2009; cf. Hebbelmann & von Sydow, 2014; von Sydow, Hagmayer, Meder, & Waldmann, 2010) but presented all items in an overview format. This type of format allows participants to detect that the Markov condition does not hold on the category level, because subclasses of items with different contingencies are mixed. In addition to eliciting probability judgments on the category level, we obtained judgments about individual items to investigate the relationship between category-based and data-driven inferences.

Figure 2 illustrates our experimental paradigm. Each black square represents an individual item defined by the feature dimensions "size" and "grayscale" (see Fig. 3). The item space is partitioned into three categories $A/\neg A$, $B/\neg B$, and $C/\neg C$, with different boundaries. The data imply a positive contingency between A and B, $P(B|A) = .75 > P(B|\neg A) = .25$, as well as between B and C, $P(C|B) = .75 > P(C|\neg B) = .25$. However, A and C are independent, $P(C|A) = P(C|\neg A) = .5$, as indicated by the orthogonal category boundaries of A and C. The Markov condition is violated because $P(C|B \wedge A) = 2/3$ but $P(C|B \wedge \neg A) = 1$. Likewise, $P(C|\neg B \wedge A) = 0$ but $P(C|\neg B \wedge \neg A) = 1/3$ (see Fig. 2). This is due to subclasses of A items having different probabilities of co-occurring with event C. For instance, A items that are bright and small always lead to C, but A items that are dark and small never do.

The participants' task was to judge the conditional probability of C given A, $P(C|A)$, after being presented with the data.[2] Our key question was whether people would recognize the independence of A and C based on the observed data, or whether they would induce a Markov-coherent causal chain and transitively infer a positive relation. If they based their inferences solely on the available data, participants should judge that A and C are independent, $P(C|A) = P(C|\neg A) = .5$. In contrast, if people induce a causal chain $A \rightarrow B \rightarrow C$ and assume the Markov condition, they should infer that A and C are positively related, with $P(C|A) = .625$ (see Eq. 2). In Experiment 1, we elicited judgments on the category and item level. In Experiment 2 we compared several intransitive and transitive chains and investigated additional boundary conditions (e.g., judgment order).

[2] Conditional probabilities are simple, uncontroversial measures (Evans & Over, 2004; Oberauer, Weidenfeld, & Fisher, 2007). Measures of causality (ΔP, causal power) yield qualitatively similar predictions for the investigated contingencies.

Experiments 1a and 1b

In both experiments, participants were sequentially provided with data regarding the relations $A \rightarrow B$ and $B \rightarrow C$, which also allowed them to observe $A \rightarrow C$. The data entailed a violation of the Markov condition, rendering A and C statistically independent. Participants first judged $P(B|A)$ and $P(C|B)$ for the individual relations and finally estimated $P(C|A)$. In Experiment 1a, data were removed before participants judged $P(C|A)$. In Experiment 1b, the data remained visible. The question was whether participants would recognize that A and C were independent (e.g., by realizing that the category boundaries were orthogonal), or whether they would derive estimates for $P(C|A)$ from their causal model representation and reason transitively from A to C, concluding a positive relation. In a control condition participants were presented with only data on $A \rightarrow C$. Since the mediating event B was omitted, they were expected to realize the independence of A and C. We also requested judgments for individual items; the goal was to investigate to what extent people's judgments were sensitive to the varying contingencies on the item level.

Method

Participants and design. One hundred twenty-eight students from the University of Göttingen participated, in exchange for candy and participation in a lottery where they could win €50. There were sixty-four participants in each experiment (Experiment 1a: 68% female, M_{age} = 24 years; Experiment 1b: 45% female, M_{age} = 23 years). Participants were randomly assigned to either the chain condition ($A \rightarrow B \rightarrow C$) or the control condition ($A \rightarrow C$). In Experiment 1a, one participant who had previously participated in a related study was replaced.

Materials and procedure Both experiments used the same materials and counterbalancing conditions. The procedure was also identical except for varying whether the data were removed before judging $P(C|A)$ (Experiment 1a) or remained visible (Experiment 1b). Participants were asked to take the role of a developmental biologist investigating three

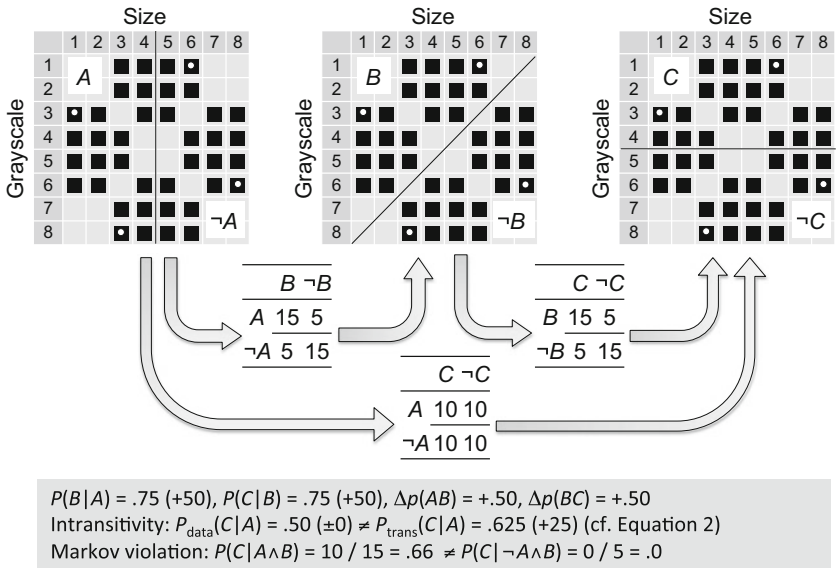

Fig. 2 Item space and examples for the event structure in the experimental conditions of Experiments 1a and 1b. Each black square represents one stimulus item ("microbe"; see Fig. 3), created by combining different levels of grayscale (1=bright, 8=dark) and size (1=small, 8=large). Squares with a white dot denote the four individual test items. The black lines bisecting the item space denote the category boundaries. Below the squares, contingencies and some resulting measures are shown. After the conditional probabilities the corresponding values on the used scale are shown in parentheses. $\Delta p(AB)$ denotes Delta P as a measure of causal strength between two variables, $P_{data}(C|A)$ represents the conditional probability of C given A, as entailed by the data, and $P_{trans}(C|A)$ represents the conditional probability as estimated transitively, based on $P_{data}(B|A)$, $P_{data}(B|\neg A)$, $P_{data}(C|B)$ and $P_{data}(C|\neg B)$, and Eq. 2

developmental stages of microbes. Specifically, the relations between the kinds of carotene developed by the microbes in three consecutive stages (α-carotene, β-carotene, and γ-carotene) were of interest.

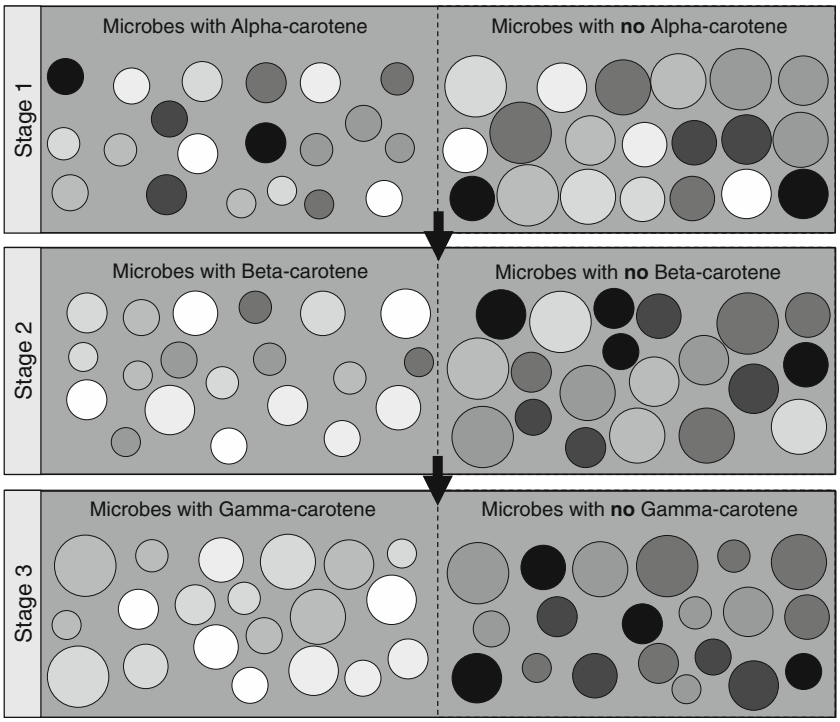

Fig. 3 Stimuli ("microbes") of Experiments 1 and 2 as shown to participants. Participants were instructed that microbes did not change their appearance across stages. Microbes were sorted in each stage according to whether they produced the stage-specific carotene (alpha-, beta-, or gamma-carotene). Participants' task was to judge the conditional probability of one type of microbe later becoming another specific type of microbe, that is, to judge $P(B|A)$, $P(C|B)$, and $P(C|A)$

Stimuli consisted of 40 individual "microbes" represented as circles (Fig. 3), varying on the dimensions "grayscale" and "size" (Fig. 2). A pretest showed that participants could accurately distinguish the individual items. Categories *A*, *B*, and *C* were created by rotating the category boundary, resulting in orthogonal categories *A* and *C*. To permit three linearly separable categories, some feature combinations were eliminated (Fig. 2).

Eight counterbalancing conditions with identical contingencies were created by rotating the three category boundaries in steps of 45° (Fig. 2). Accordingly, categories *A* and *C* had a one-dimensional boundary in four counterbalancing conditions, whereas category *B* involved a two-dimensional boundary. In the other four conditions, *A* and *C* had a two-dimensional and *B* a one-dimensional category boundary. In all conditions, there was a positive contingency between *A* and *B,* and *B* and *C*, respectively, whereas *A* and *C* were independent: $P(C|A) = P(C|\neg A) = .5$.

Participants were first presented with data regarding the first and second developmental stages, printed on large panels (Fig. 3). Data for the first stage arranged the 40 microbes according to whether they produced α-carotene (*A*) or not (¬*A*). The same microbes in the second stage were arranged according to whether they generated β-carotene (*B*) or not (¬*B*). Panels for the first and second stages were displayed simultaneously.

Participant first judged the conditional probability $P(B|A)$; that is, they judged whether microbes with α-carotene (*A*) tended to produce β-carotene (*B*) or not (¬*B*). The experimenter pointed to the corresponding categories on the panels. We used an 11-step rating scale of −100 to +100, with a midpoint of 0 indicating the independence of *A* and *B* (Fig. 4). On this scale $P(B|A) < .5$ corresponds to values below zero, $P(B|A) > .5$ corresponds to values above zero, and $P(B|A) = .5$ corresponds to zero.

After participants judged $P(B|A)$, the *C* panel was added, showing the microbes that had produced γ-carotene (*C*) and those that had not (¬*C*; Fig. 3). Using the same rating scale, participants judged the conditional probability $P(C|B)$, that is, whether microbes producing β-carotene (*B*) would or would not produce γ-carotene.

In Experiment 1a, after participants judged $P(C|B)$, all data panels were removed and participants had to judge $P(C|A)$ on the same 11-step rating scale. In Experiment 1b, after participants judged $P(C|B)$, all data panels remained visible when participants judged $P(C|A)$, so that the data showing the independence of *A* and *C* were available during the judgment.

The procedure and materials in the control conditions of both experiments were identical, except that participants were shown data only for categories *A* and *C*. Accordingly, they provided an estimate of only $P(C|A)$, with panels for *A* and *C* being visible during judgment.

After participants completed all probability judgments, the data panels were removed and participants were presented with four individual microbes (presented in one of two random orders). Figure 2 indicates the location of these test items with a white spot. We selected these items since they were at least one step away from the category boundary in all counterbalancing conditions, and in all counterbalancing conditions each item belonged to one of four combinations of categories ($A \wedge C$, $\neg A \wedge C$, $A \wedge \neg C$, and $\neg A \wedge \neg C$). For instance, in Fig. 2 the item with size = 1 and grayscale = 3 is of type $A \wedge C$ (i.e., the microbe produced α-carotene and γ-carotene). The goal was to investigate whether judgments were influenced by observed contingencies on the item level and/or transitive inferences based on the category level.

For each test item, participants judged the probability that it would generate γ-carotene. To eliminate uncertainty regarding the category membership of *A*, information was provided for each item on whether it had or had not produced α-carotene. Again, a rating scale of −100 to +100 was used, labeled "This [non] alpha microbe tends not to develop gamma-carotene later" on the left side and "tends to develop gamma-carotene later" on the right side. Different ratings for *A* and ¬*A* items yielding the same effect *C* would indicate an influence of category-based transitive reasoning. Different ratings for items belonging to categories *C* or ¬*C* would indicate an influence of the actually observed contingency on the item level.

Results

For the analyses, the eight counterbalancing conditions were recoded to match the data structure depicted in Fig. 2. Our

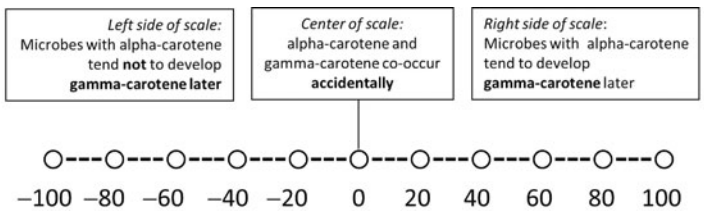

Fig. 4 Example scale used to elicit conditional probability judgments in Experiments 1a and 1b

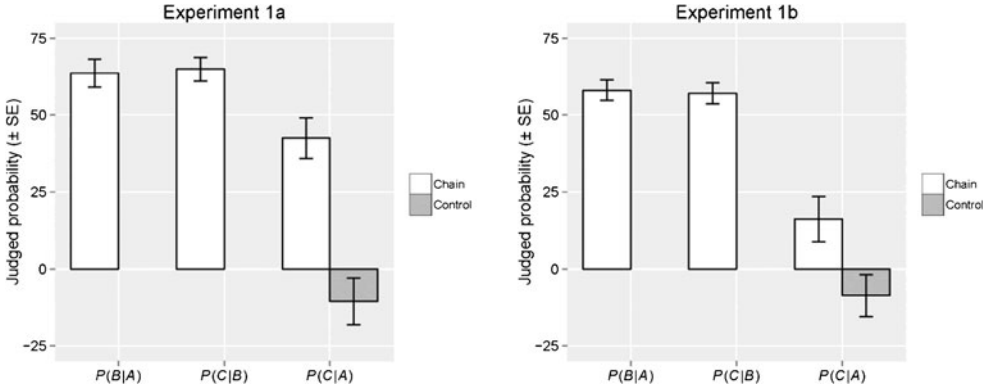

Fig. 5 Mean judgments (±SE) on the category level in Experiments 1a and 1b. Judgments were given on an 11-step scale of −100 to +100. $P(C|A)$ was judged with data being removed in Experiment 1a and data being present in Experiment 1b. For the local relations, $P(B|A) = P(C|B) = .75$ holds in the data, corresponding to +50 on the scale used. The observed data entails $P(C|A) = P(C|\neg A) = .5$, corresponding to a value of 0 on the scale used. Deriving $P(C|A)$ from a causal chain via Eq. 2 yields .625, corresponding to a value of +25 on the scale used. Judgments of $P(B|A)$ and $P(C|B)$ were not collected in the control condition

predictions for transitive inferences for the global $A \rightarrow C$ relation rely on qualitatively correct judgments of the local $A \rightarrow B$ and $B \rightarrow C$ relationships. Therefore, following previous research (Baetu & Baker, 2009), we included only participants who correctly judged both local relationships to be positive. In Experiment 1a, 25% of participants failed to meet this criterion in each of the two relations; in Experiment 1b an average of 19% failed in each of the two relations.[3]

Figure 5 shows participants' mean probability judgments. In both studies, judgments for the local relations $P(B|A)$ and $P(C|B)$ reflect the probabilities in the data. (The objective probabilities $P(B|A) = P(C|B) = .75$ correspond to a value of +50 on the scale used.) The key finding is that in both experiments participants in the chain condition judged $P(C|A)$ to be greater than zero, Experiment 1a: $t(15) = 6.49$, $p < .001$; Experiment 1b: $t(20) = 2.19$, $p < .05$. Thus, although A and C were independent in the observed data (corresponding to zero on the scale used), participants judged this relation to be positive. In contrast, in the control conditions in which the intermediate event B was omitted, judgments did not differ from zero, Experiment 1a: $t(31) = -1.39$, $p = .17$; Experiment 1b: $t(31) = .21$, $p = .22$. Judgments for $P(C|A)$ differed between experimental and control condition in both Experiment 1a, $t(46) = 4.52$, $p < .001$, and Experiment 1b, $t(51) = 2.84$, $p < .05$.[4]

These findings indicate that participants relied on transitive reasoning to judge the relation between A and C. Despite being able to detect that these categories of items were independent when B was omitted, they inferred a positive relation when B was the intermediate event. The lower judgments for $P(C|A)$ in Experiment 1b indicate that making the data available during judgment increased people's sensitivity to the independence of A and C.

We next analyzed participants' probability judgments regarding C for the individual test items, which consisted of four items combining membership of categories A and C (i.e., items of type A and C, A and $\neg C$, $\neg A$ and C, and $\neg A$ and $\neg C$; see Fig. 2). Theoretically, if judgments are based solely on transitive reasoning on the category level, there should be identical positive judgments for A items around +25, regardless of whether an item belonged to category C or $\neg C$. The two $\neg A$ items ($\neg A$ and C and $\neg A$ and $\neg C$) should receive negative ratings of around −25. Conversely, if participants' judgments are driven solely by the observed data, C items should receive a rating of +100, and $\neg C$ items one of –100, independent of whether they are type A or $\neg A$. Note that the maximal "bottom-up" effect (C vs. $\neg C$ items) is larger than the maximal "top-down" effect (A vs. $\neg A$ items).

[3] The predictions for people who *failed* to meet the selection criteria are not clear, since they may have failed for various reasons, such as a lack of concentration or because they recognized the intransitivity. To explore whether participants used the local relations as predictors, we correlated the empirically found estimates of $P(C|A)$ with the product $P(C|B) \times P(B|A)$ and with $P(C|B) \times P(B|A) + (1\text{-}P(C|B)) \times (1\text{-}P(B|A))$. The latter estimate relies on a symmetry assumption in line with the observed data: $P(C|B) = P(\neg C|\neg B)$. In Experiment 1a, both correlations were positive and large when we included all participants ($r = .53$ and $r = .55$) or only those meeting the selection criterion ($r = .63$ and $r = .63$). In Experiment 1b, in contrast, we obtained *no* correlation when considering *all* participants ($r = .12$ and $r = .08$) but strong positive correlations for participants meeting the criterion ($r = .55$ and $r = .53$). This suggests that people in the group meeting the selection criterion were often guided by transitivity, but that at least in Experiment 1b many of the participants failing to meet the criterion cannot be modeled by transitivity.

[4] Additionally, we controlled for the two patterns of one-dimensional (1D) and two-dimensional (2D) category boundaries used in the eight control conditions (1D–2D–1D vs. 2D–1D–2D categorization type). Descriptively, the distortion effect was larger in the 2D–1D–2D condition. However, a two-way ANOVA concerning the $P(C|A)$ judgments showed a significant main effect only of experimental condition vs. control, Experiment 1a, $F(1, 44) = 19.50$, $p < .001$; Experiment 1b, $F(1, 49) = 5.15$, $p < .05$. There was no effect of categorization type, Experiment 1a, $F(1, 44) = 1.7$, $p = .20$; Experiment 1b, $F(1, 49) < 1$, $p = .99$, and no interaction, Experiment 1a, $F(1, 44) < 1$, $p = .94$; Experiment 1b, $F(1, 49) = 1.43$, $p = .24$.

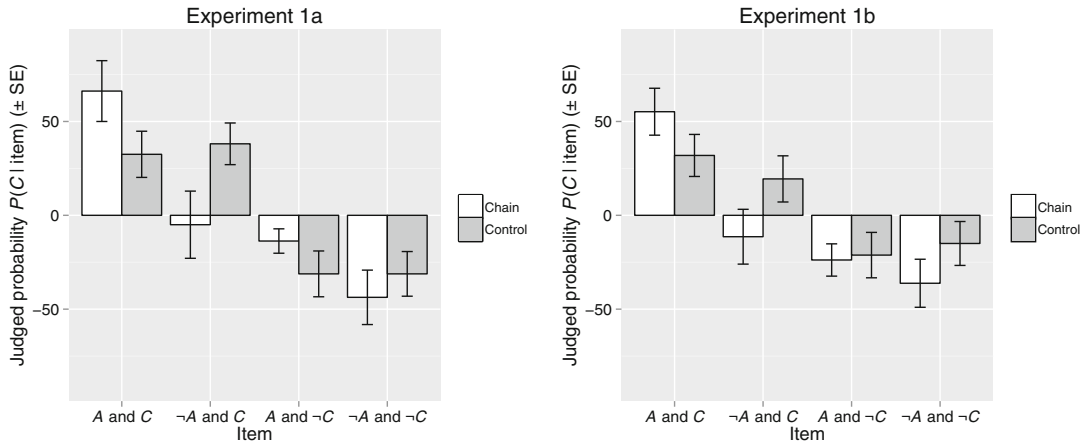

Fig. 6 Mean judgments (±*SE*) of *P*(*C*|item) in Experiments 1a and 1b. Each item corresponds to a combination of categories *A*/¬*A* and *C*/¬*C* (see Fig. 2 for locations in the item space). For each item, information on *A* vs. ¬*A* was given and the task was to infer the probability of *C*. Judgments were given on an 11-step scale of −100 to +100

Figure 6 shows participants' mean judgments for the four test items regarding the probability of *C*. For these judgments the learning data had been removed in both studies. Judgments in the control condition serve as baseline, showing the influence of the data on participants' judgments without intermediate event *B*. In the control condition, in both studies judgments varied only as a function of whether an item did or did not belong to category *C*, irrespective of whether the item belonged to category *A* or ¬*A*. Consequently, in both experiments an analysis of variance (ANOVA) for the control condition, with item types *A* versus ¬*A* and *C* versus ¬*C* as within-subject factors, yielded a significant effect only for items of type *C* versus ¬*C*, Experiment 1a, $F(1, 31) = 17.90$, $p < .001$, $\eta_p^2 = .37$; Experiment 1b, $F(1, 31) = 6.33$, $p < .05$, $\eta_p^2 = .17$; there was no influence of type *A* versus ¬*A*, $F(1, 31) < 1$, and no interaction, Experiment 1a, $F(1, 31) < 1$; Experiment 1b, $F(1, 31) = 1.27$, $p = .27$.

A different pattern of judgments was obtained in the chain conditions. For instance, the two items belonging to category *C* were judged differently depending on the status of *A*, with higher judgments for *A* than for ¬*A* items. Analogously, items that belonged to category ¬*C* received higher judgments when belonging to *A* than when belonging to ¬*A*. The pattern also reflects an influence of the data, as the judgments for the two *A* items and the two ¬*A* items varied as a function of true membership regarding *C*. Accordingly, in Experiment 1a, a main effect of type *C* versus ¬*C*, $F(1, 15) = 11.67$, $p < .01$, $\eta_p^2 = .43$, a main effect of type *A* versus ¬*A*, $F(1, 15) = 6.04$, $p < .05$, $\eta_p^2 = .28$, and an interaction, $F(1, 15) = 9.72$, $p < .05$, $\eta_p^2 = .39$, resulted. Thus, in the chain condition participants' judgments were influenced by both the category level relations and observations on the item level.

In Experiment 1b, the pattern was qualitatively similar, although less pronounced. The ANOVA for the chain condition yielded a significant effect of items of type *C* versus type ¬*C*, $F(1, 20) = 14.69$, $p < .01$, $\eta_p^2 = .46$, as well as—at least for a one-tailed test of our prediction—an influence of type *A* versus ¬*A*, $F(1, 20) = 3.80$, $p_{\text{one-tailed}} < .05$, $\eta_p^2 = .16$, but no interaction, $F(1, 20) < 1$. The two main effects indicate that judgments on the item level were still influenced by the inferred positive relation between *A* and *C* on the category level and the observed data concerning category *C* versus ¬*C*. As in Experiment 1a, the effect of the category membership of *C* was larger than that of category *A*, consistent with the theoretically maximal effect of these factors.

Table 1 shows the distribution of judgments, that is, the proportion of participants judging *P*(*C*|*A*) to be positive, zero, or negative. Although a larger proportion of participants in the experimental conditions detected the zero contingency in Experiment 1b than in Experiment

Table 1 Percentage (and frequency) of participants in Experiments 1a and 1b judging the relation between *A* and *C* to be negative, zero, or positive on a scale of −100 to +100

Experiment	Condition	Negative	Zero	Positive
1a	Intransitive chain	0% (0)	6% (1)	**94% (15)**
	Control	28% (9)	**56% (18)**	16% (5)
1b	Intransitive chain	19% (4)	33% (7)	**48% (10)**
	Control	28% (9)	**56% (18)**	16% (5)

Note. Boldface entries correspond to the prediction of transitive distortion effects in chain conditions and zero judgments in the control condition. Participants' judgments were classified as positive for the interval $0 < x \leq 100$, negative for $-100 \leq x < 0$, and zero if and only if they answered "zero."

1a (exact Fisher test, $p < .05$), both experiments show a larger proportion of positive answers in these conditions than in the respective control conditions (Experiment 1a: $p < .001$; Experiment 1b: $p < .05$).

Discussion

The studies show that participants who made correct judgments regarding the two local relations $A \rightarrow B$ and $B \rightarrow C$ judged $P(C|A)$ to be larger than zero, indicating a transitive inference from the chain's initial event A to the final event C. These judgments strongly differed from those in the control conditions without intermediate variable B, in which judgments around zero were obtained. The size of the effect was modulated by the specific testing conditions, with a stronger influence of the category-based transitive inference when the data were not visible during the judgment. Probability judgments for individual test items were influenced by transitive inferences on the category level and item-specific knowledge. The results support the idea that participants induced a causal chain $A \rightarrow B \rightarrow C$ and reasoned transitively from A to C, although the Markov condition was violated and transitivity did not hold in the data.

Experiment 2

The goal of Experiment 2 was to investigate reasoning with intransitive chains in a wider array of circumstances. In addition to the intransitive chain of Experiment 1 ($A \rightarrow B$, $B \rightarrow C$, A independent of C, henceforth denoted ++0) we included a new intransitive chain involving two preventive causal relations ($A \rightarrow \neg B$, $\neg B \rightarrow C$, A independent of C, denoted −−0). The goal was to rule out that positive judgments for the local relations created a response bias toward a positive rating for the $A \rightarrow C$ relation. Additionally, we used a new control condition that matched the complexity of the intransitive chains. This involved a positive $A \rightarrow B$ contingency, followed by independent B and C variables, and likewise independent variables A and C (denoted +00).

We also included two transitive chains that obeyed the Markov assumption ($A \rightarrow B$, $B \rightarrow C$, $A \rightarrow C$, denoted +++; and $A \rightarrow \neg B$, $\neg B \rightarrow C$, $A \rightarrow C$, denoted −−+) as a comparison for the respective intransitive chains (++0 and −−0). This allowed us to investigate whether participants would use the observable evidence in addition to transitive reasoning to judge the indirect relation between A and C. Finally, we included a Markov-coherent chain with a negative overall relation (−+−) to examine whether people correctly learn a negative overall relationship.

Based on the findings of Experiment 1 we expected distortion effects due to transitive inferences in conditions ++0 and −−0. However, we expected to find higher ratings in the respective transitive conditions +++ and −−+, due to an additional influence of the observable positive relation between A and C. In the control condition +00, both transitive inferences and the learning data entailed a zero contingency. Therefore we expected participants to correctly detect A's and C's independence.

Finally we controlled for question order. In the *local–global* conditions, similar to Experiment 1, participants rated the individual causal links before judging the conditional probability of C given A. In the *global–local* conditions, this order was reversed.

Method

Participants One hundred twenty-four participants (56% female; $M_{age} = 23$ years), mostly students from the University of Heidelberg, took part in exchange for chocolate and participation in a lottery where they could win €50. Three participants were excluded from the analyses (two clear outliers in the time used and one who gave a rating of −100 in all local judgments).

Design The experiment had a 2 (Judgment Order: local–global vs. global–local, between-subjects) × 6 (Contingency Condition, within-subject) mixed factorial design. The six within-subject conditions, which involved different $A \rightarrow B$, $B \rightarrow C$, and $A \rightarrow C$ contingencies, were presented in random order. Figure 7 shows the item space and resulting contingencies for the six contingency conditions. Conditions ++0 and −−0 were intransitive and Markov-incoherent. The +++, −−+, and −+− conditions were transitive and Markov-coherent. The neutral control condition +00 entailed a zero $A \rightarrow C$ contingency, both in the data and when using Eq. 2. For each condition (see Fig. 7), participants were randomly assigned to one of eight counterbalancing conditions created by rotating the category boundaries, resulting in 48 data sets, each with three data panels presenting 40 microbes.[5]

Materials and procedure We used the same scenario and materials as in Experiment 1 (see Fig. 3), but in a computer-based experiment. As in Experiment 1b, judgments about $P(C|A)$ were elicited in the presence of the learning data. Since participants were presented with several contingency conditions, we omitted the single-item test.

First, participants were told about the microbes, their three developmental phases, the three types of carotenes (α-carotene, β-carotene, γ-carotene), and the question order. To elicit probability judgments, we again used an 11-step scale but with different labels and no numbers (see Fig. 8). For reasons of comparability we report results using a scale of −100 to +100.

Participants were randomly assigned to a question order. In the local–global condition, participants were first shown the $A/\neg A$ panel and the $B/\neg B$ panel and asked to judge $P(B|A)$. Subsequently, the $C/\neg C$ panel was added and participants judged $P(C|B)$. Finally, they estimated $P(C|A)$, with the

[5] With 40 microbes, some conditions only approximate the Markov condition. We did not increase the number of items in order to retain comparability with Experiment 1 and to ensure that the task did not become more difficult.

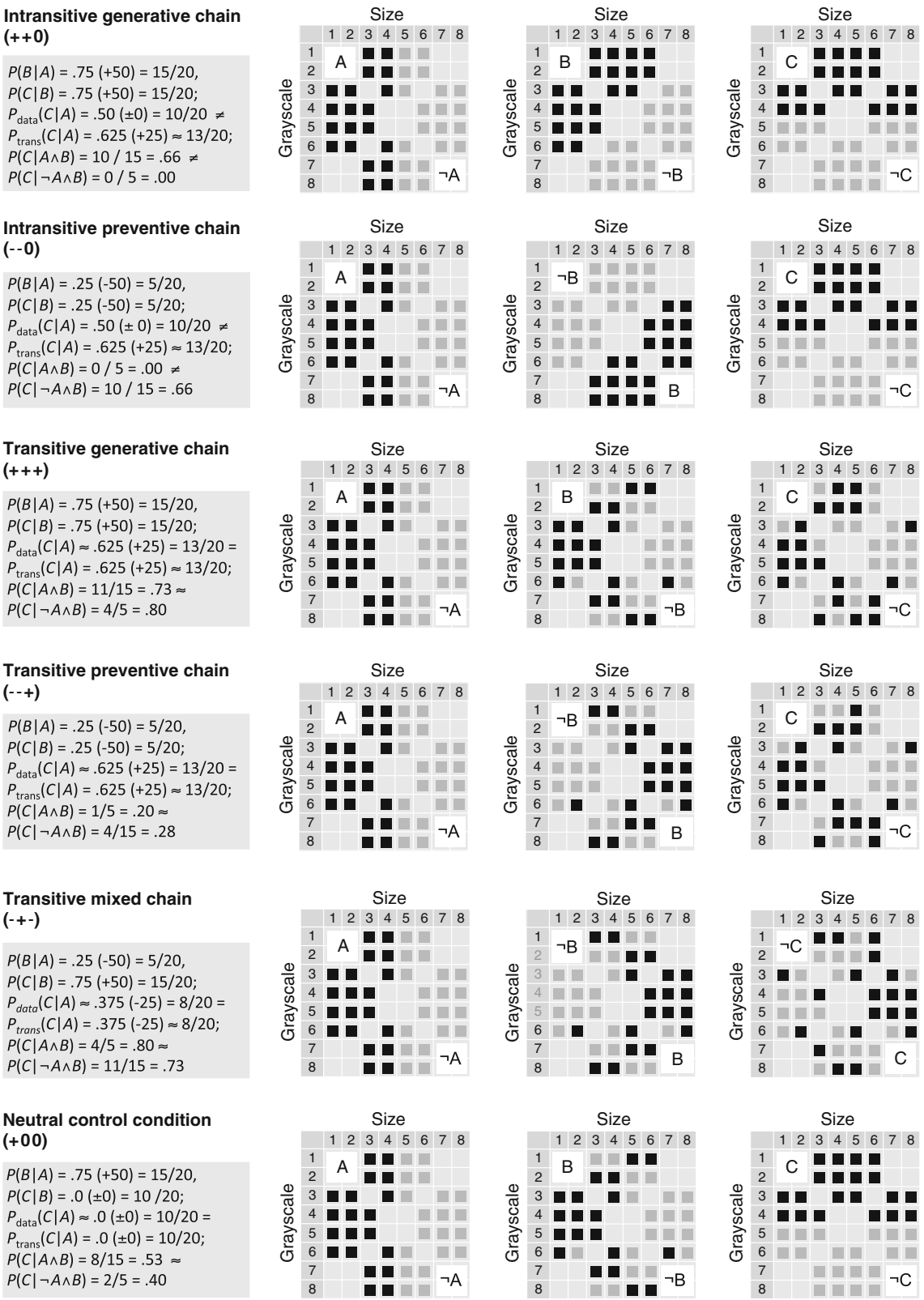

Fig. 7 Item space in the six contingency conditions of Experiment 2. Black squares represent stimulus items ("microbes") in each of the three developmental stages showing a corresponding carotene (*A*, *B*, *C*); lighter squares refer to items without the corresponding carotenes (¬*A*, ¬*B*, ¬*C*)

previous judgments and data remaining visible. In the global–local condition, all three panels were presented from the outset and participants first estimated $P(C|A)$. Subsequently, we asked for judgments for $P(B|A)$ and $P(C|A)$. Finally, we

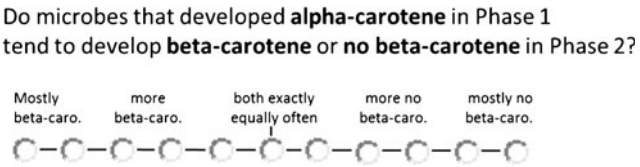

Fig. 8 Example of scale used for eliciting conditional probability judgments in Experiment 2. Judgments were given on an 11-step scale without values; for reasons of comparability with Experiment 1, we report our results using a scale of −100 (*mostly no beta-carotene*) and +100 (*mostly beta-carotene*). Note that the scale in Experiment 1 had −100 on the left and +100 on the right. Here, the equivalent of −100 (*mostly no beta-carotene*) is on the right and the equivalent of +100 (*mostly beta-carotene*) is on the left

included an item-sensitivity test to make sure that participants were able to distinguish between neighboring feature values of size or brightness (see Appendix 1).

Results

All answers were recoded to match the counterbalancing conditions shown in Fig. 7. We used the same selection criterion as before; that is, participants had to judge the two local links qualitatively correctly (cf. Baetu & Baker, 2009), because this is a prerequisite for testing the impact of transitive reasoning in intransitive chains. Judgments for positive local relations had to be in the interval $+20 \leq x \leq +100$ (i.e., one of the five farthest points to the right on the 11-point rating scale); for a negative relation in the interval $-100 \leq x \leq -20$ (i.e., one of the five farthest points to the left); and for the relation with a predicted zero mean in the interval $-40 \leq x \leq +40$ (i.e., one of the five points mid-scale). These intervals are centered on the predicted values for the respective relations: positive = 50, negative = −50, and null = 0.[6] Selections were made for each condition separately.

Figure 9 shows the mean estimates for judgments of $P(C|A)$. Judgments of $P(B|A)$ and $P(C|B)$ were close to the true values (for details, see Appendix 2, Table 5). Importantly, the intransitive chains ++0 and −−0 yielded positive judgments, despite A and C being independent.[7] In the Markov-coherent +00 condition, answers were close to zero, the transitive chains +++ and −−+ yielded strong positive judgments, and the transitive chain −+− yielded negative judgments. Question order (local–global vs. global–local) did not influence judgments.

We first analyzed whether ratings of $P(C|A)$ differed from zero (Table 2). Unsurprisingly, ratings differed from zero in the three conditions in which there was a relation between A and C (+++, −−+, and −+−). However, consistent with transitive reasoning, estimates also differed from zero in the Markov-incoherent conditions (++0 and −−0), although A and C were independent. They did not differ from zero in the control condition (+00).

Second, we compared estimates of $P(C|A)$ across the different conditions, controlling for question order.[8] Table 3 shows the results of respective 2 × 2 ANOVAs. For no comparison were Question order or the Order × Contingency interaction significant. The comparisons of the generative intransitive chain ++0 and the preventive intransitive chain −−0 with the neutral control condition +00 (see Table 3, upper two rows) indicate illicit transitive inferences. In both intransitive chain conditions, judgments were higher than in the control condition (differences of 18.7 and 13.8, respectively), although the difference between the preventive chain and the control condition was only marginally significant ($p = .06$, one-sided test).[9]

To investigate to what extent participants were sensitive to the observed data, ratings in the intransitive conditions (++0 and −−0) were compared to the corresponding transitive conditions (+++ and −−+; see Table 3, two middle rows). Judgments were higher in the latter cases when $P(C|A) > .5$ than when $P(C|A) = .5$. These results show that judgments in the intransitive conditions were influenced not only by transitive reasoning on the category level, but also by the observed contingencies.

Additionally, we examined potential response biases by comparing the ++0 condition with the −−0 condition, and the +++ condition with the −−+ condition (see Table 3, bottom two lines). If observing two positive relationships for the direct links or giving two positive judgments creates a tendency to judge the indirect relation positively, too, judgments of $P(C|A)$ should differ between conditions. The analyses show that this was not the case: There was no response bias and no effect of question order (see also Fig. 9).

Furthermore, Table 4 presents an analysis on the individual level for the target inference $P(C|A)$. Each participant was classified according to judging a particular relation qualitatively as positive, negative, or zero. In both the generative intransitive condition (++0) and the preventive intransitive chain condition (−−0) the overall proportion of positive answers was higher than in the neutral control condition (+00; four field χ^2 tests, $p < .001$, $p < .05$).[10] There was an even higher proportion of positive $P(C|A)$ judgments in the corresponding transitive conditions +++ and −−+ (χ^2 tests, $p < .05$, $p < .001$). The proportion of negative answers in the mixed

[6] Across all conditions, an average of 29% of answers concerning local relations fell into the six excluded levels of the scale (out of 11 levels).

[7] We also tested if the difference between the categorization types modulated the size of the distortion effect. There was no significant difference in the two relevant intransitive conditions, ++0: $F(1, 71) = 2.55$, $p = .11$ and −−0: $F(1, 64) = 3.35$, $p = .07$, but descriptively the distortion effect was larger in the 2 dimension-1 dimension-2 dimension (2D–1D–2D) condition.

[8] We did not conduct a global ANOVA on question order because applying the selection criterion of qualitatively correct local judgments to all conditions simultaneously would have excluded too many participants.

[9] Note that this test has a lower statistical power than the test against zero, because applying the selection criterion to both conditions lowered the number of participants involved.

[10] However, for the preventive intransitive chain condition this difference seems to have been driven mainly by the local–global condition where the difference between positive and negative answers became significant (exact binomial test, $p < .05$).

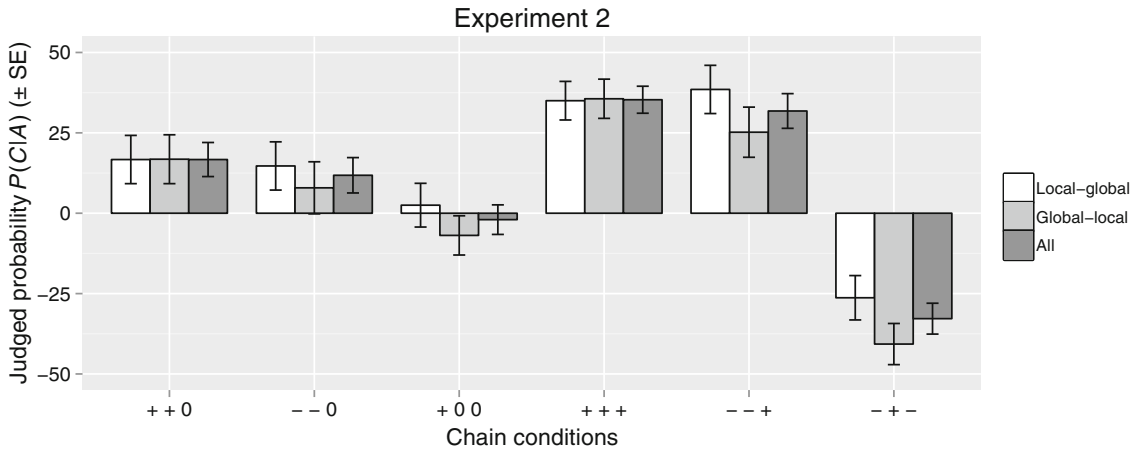

Fig. 9 Mean judgments of $P(C|A)$ ($\pm SE$) on the category level in Experiment 2. Judgments were given on an 11-step scale of −100 to + 100. Local–global: Participants first judged the local relations $P(B|A)$ and $P(C|B)$ before judging $P(C|A)$. Global–local: Participants first judged $P(C|A)$ and subsequently the local relations $P(B|A)$ and $P(C|B)$. All: Mean judgments aggregated across question order. See text for descriptions of chain conditions

transitive condition (−+−) was of a similar size to the proportion in the positive transitive conditions and higher than that in the control condition +00 (χ^2 test, $p < .001$).

Finally, the item-sensitivity test corroborated a reasonable ability of participants to distinguish between neighboring feature values of the items shown (microbes) and that differences in ability seem not to have driven the distortion effect (see Appendix 1).

Discussion

In Experiment 2 we replicated and extended the results of Experiment 1, while controlling for alternative explanations. In both intransitive chain conditions (++0 and −−0) participants' judgments deviated from the observed data, consistent with the idea that people tend to induce Markov-coherent causal chains and use them to reason transitively from A to C. The fact that participants gave higher ratings when reasoning with transitive chains suggests that judgments were also influenced by the learning data. In the transitive condition the ratings even seem a bit too high, but this may relate to the numberless rating scale in this experiment (see also Rehder & Burnett, 2005).

Our analyses also show that the distortion effect in the intransitive conditions cannot be explained by answer tendencies being due to influences from previous judgments or by previous beliefs about the global relation of A and C. First, the +00 control condition yielded judgments close to zero; second, the intransitive −−0 and ++0 conditions yielded similar judgments. Third, judgments in the positive (+++, −−+) and negative (−+−) transitive conditions had similar absolute positive or negative values, and fourth, the order in which judgments were elicited was irrelevant.

Table 2 t Tests (one-tailed) for judgments of $P(C|A)$ against zero in Experiment 2

Condition	$P(C\|A)$ M (SE)	df	t	p
++0	16.71 (5.30)	72	3.15	.002
−−0	11.81 (5.43)	65	2.16	.033
+00	−1.97 (4.62)	60	−0.43	.672
+++	35.29 (4.27)	84	8.27	.001
−−+	31.85 (5.48)	83	5.81	.001
−+−	−32.81 (4.85)	63	−6.77	.001

General discussion

Our goal was to investigate whether transitive inferences in probabilistic causal chains of the type $A \rightarrow B \rightarrow C$ distort the induction of the relationship between A and C when the transitive inference based on the independent combination of the observed local relationships $A \rightarrow B$ and $B \rightarrow C$, for instance, entails a positive indirect relationship, while the data directly shows that that A and C are independent. We studied the influence of transitive inference in intransitive chains that violated the Markov condition on the category level because heterogeneous subclasses of items were mixed. Our results show that people made judgments about $P(C|A)$ that systematically deviated from the observed data but were consistent with a transitive inference from A to C based on a mental causal model (illicitly) obeying the Markov assumption.

Experiment 1a demonstrated the influence of inappropriate transitive reasoning when participants learned consecutively about the two individual links and made judgments after the

Table 3 Analyses of variance (2 × 2) comparing judgments of $P(C|A)$ in the within-subject contingency conditions while controlling for between-subjects question order

Comparisons		Contingency			Question order			Contingency × Order			N
		F	η_p^2	p	F	η_p^2	p	F	η_p^2	p	
++0	+00	6.16	.12	.017	0.12	.00	.725	0.19	.00	.668	44
−−0	+00	2.57	.05	.116	2.42	.06	.127	0.06	.00	.805	42
++0	+++	9.61	.13	.003	0.04	.00	.848	0.18	.01	.674	63
−−0	−−+	7.66	.16	.008	0.14	.00	.641	0.10	.00	.748	42
++0	−−0	0.05	.00	.828	0.40	.00	.529	0.05	.00	.768	45
+++	−−+	0.46	.01	.500	0.22	.00	.995	1.99	.04	.165	47

Note. See Fig. 9 for mean judgments in the different conditions.

learning data were removed. Experiment 1b showed that this finding was also obtained when all data were available while judging $P(C|A)$, although in this case the judgments were influenced more strongly by the learning data. When the intermediate event B was omitted from the data, participants had no difficulty recognizing that A and C were unrelated. Further analyses showed that judgments on the level of individual items were influenced by transitive inferences on the category level and the item-level relations.

Experiment 2 investigated the robustness of these findings while controlling for alternative explanations, such as task complexity and possible answer tendencies. Similar findings to Experiments 1a and 1b were obtained. A direct comparison of intransitive versus transitive chains showed that participants' judgments on the category level were influenced not only by illicit transitive reasoning but also by the observed data. Results in the different control conditions refute the idea that these distortion effects are due to answer tendencies

Table 4 Percentage (and frequency) of participants in Experiment 2 judging the relationship between A and C to be negative, zero, or positive, on a scale of −100 to +100

Contingency condition	Order	Negative	Zero	Positive
Generative intransitive chain (++0)	Local–global	30% (11)	6% (2)	**64% (23)**
	Global–local	19% (7)	22% (8)	**59% (22)**
	All	25% (18)	14% (10)	**62% (45)**
Preventive intransitive chain (−−0)	Local–global	24% (9)	24% (9)	**53% (20)**
	Global–local	36% (10)	21% (6)	**42% (12)**
	All	29% (19)	23% (14)	**48% (32)**
Neutral control (+00)	Local–global	25% (8)	**41% (13)**	34% (11)
	Global–local	44% (13)	**28% (8)**	28% (8)
	All	34% (21)	**34% (21)**	31% (19)
Generative transitive chain (+++)	Local–global	14% (6)	7% (3)	**79% (35)**
	Global–local	10% (4)	10% (4)	**80% (33)**
	All	12% (10)	8% (7)	**80% (68)**
Preventive transitive chain (−−+)	Local–global	11% (3)	7% (2)	**81% (22)**
	Global–local	22% (6)	0% (0)	**78% (21)**
	All	17% (9)	4% (2)	**80% (43)**
Mixed transitive chain (−+−)	Local–global	**71% (25)**	11% (4)	17% (6)
	Global–local	**86% (25)**	10% (3)	3% (1)
	All	**78% (50)**	11% (7)	11% (7)

Note. Boldface entries indicate the main prediction for conditions assuming that participants engage in transitive reasoning without modeling subjective distributions over values. Participants' judgments were deemed positive for the interval $0 < x \leq 100$, negative for $-100 \leq x < 0$, and zero if and only if they answered "zero." All: Percentage and frequency of judgments aggregated across question order.

resulting from previous judgments (cf. atmosphere effects in syllogistic reasoning; Seels, 1936) or to prior beliefs concerning the global relation. Judgments of the indirect relation were correct and independent of question order in Markov-coherent, transitive conditions: positive for generative as well as preventive relations, negative for mixed relations, and zero in the neutral control condition.

The present research goes beyond previous studies (Ahn & Dennis, 2000; Baetu & Baker, 2009) that investigated the influence of transitive reasoning in the absence of data about *A* and *C*, so that learners could not assess whether the Markov condition held true. While these studies demonstrated that people made transitive inferences in the absence of data on the relation between *A* and *C*, our results suggest they do so even in the presence of counterevidence. Although participants could observe the indirect relation between *A* and *C*, judgments were substantially influenced by transitive reasoning.

Should one assume transitivity and the Markov condition when inducing causal structures?

Cartwright (2001, 2002, 2007) criticized the concept of assuming the Markov condition as a universal property of causal relations in the world. Even proponents of a universal assumption of the Markov condition concede that the condition need not hold for inadequate category schemes or incomplete causal structures actually in use (Hausman & Woodward, 1999, 2004; Spohn, 2001). Inspired by these ideas, we investigated the influence of transitive reasoning in intransitive chains that violate the Markov condition.

In our scenarios the Markov condition is violated due to mixing subclasses of items with different contingencies. Aggregating these subclasses into the same category results in a violation of the Markov condition and of transitivity, given the provided categories. However, categories play an indispensable role in causal induction and causal reasoning, as causal relations are typically defined on the category level (Lien & Cheng, 2000; Waldmann & Hagmayer, 2006; Waldmann, Meder, von Sydow, & Hagmayer, 2010; also Hagmayer, Meder, von Sydow, & Waldmann, 2011). Even if one assumes that causal relationships at a more fine-grained level adhere to the Markov condition, there is no guarantee that this is the case for a given category scheme. One rarely knows whether categories are homogeneous, and causal relationships may often involve mixtures of different causal relationships at some lower level or involve hidden variables. Thus it seems plausible for transitive distortion effects to play a substantial role in everyday as well as scientific reasoning.

Do our findings show that people's probabilistic inferences are generally flawed and error prone? The results do show that transitive reasoning that assumes an independent integration of causal links can systematically deviate from objective data. Yet every cognitive system needs to make inductive inferences about unobserved relations, and the virtue of the Markov condition is that it enables such inferences (Pearl, 2000; Spirtes et al., 1993). Moreover, the independence assumptions formalized in the Markov condition facilitate a parsimonious representation of relationships between variables (Domingos & Pazzani, 1997). Thus the Markov condition may provide a reasonable default assumption that guides human learning at least initially, even if the assumption does not hold (von Sydow et al., 2010; Jarecki, Meder, & Nelson, 2013). Although the Markov condition does not need to hold for the categories we used, it may provide reasonable guidelines for an ideal construction of causal relationships and categories (Hausman & Woodward, 1999, 2004; but cf. Cartwright, 2007). We focused here on chains where we find this idea convincing. Even if transitive distortion effects show that the independent integration of single links may lead learners astray, this is taken as support for the idea that people tend to assume that causal chains are transitive. Apart from resolving such situations by differentiating categories into different subclasses—for which we here found only weak evidence—a further way to prevent intransitive chains is that people may already induce categories in a way that allows for transitive reasoning (Hagmayer et al. 2011).

Transitive reasoning in causal chains: boundary conditions and future directions

Our findings suggest several avenues for future research. A key question concerns the boundary conditions for illicit transitive reasoning.

One way to eliminate transitive distortion effects for a chain with two nonzero local relations and a zero global contingency may be to highlight a possible direct relation between *A* and *C*. Although the temporal order of events in our experiments constrained the set of plausible causal models (Lagnado & Sloman, 2006), such constraints do not exclude a chain with an additional direct link between *A* and *C*, rather than assuming that these variables were only indirectly connected via intermediating event *B*. Our results suggest that people tend to induce a parsimonious chain model without an additional link. Future research should investigate whether better calibrated judgments are obtained if one would explicitly point out alternative causal structures.

The obtained distortion effects might have been caused by a focus on causal *relations*, and might have been attenuated by a focus on the involved *categories*. In fact, research on causal-based category induction (Lien & Cheng, 2000; Waldmann &

Hagmayer, 2006; Waldmann et al., 2010) suggests that people can use causal information to induce categories. The results of Hagmayer et al. (2011) suggest that people tend to continue to use categories from earlier nodes in a chain. They investigated the transfer of category schemes when learning causal chains $A \rightarrow B \rightarrow C$, where the dichotomous events A and C were precategorized but the intermediate event B consisted of uncategorized exemplars. They showed that the categorization of B based on A was subsequently used for the second causal relation, even if not optimal. In our task all three events were precategorized, and no transfer of categories occurred—otherwise participants would have realized that A and C were orthogonal. Nonetheless, tasks that focus more strongly on categories than on relationships between categories might reduce transitive distortion effects.

Similarly, an emphasis on different subclasses of items with different contingencies (mixing) might reduce transitive distortion effects. Although we used only a two-dimensional item space and—for the intransitive chains—deterministic relationships on the subclass level, a stronger emphasis on the existence of subclasses, communication of several different causes for categories (Bonnefon et al., 2012), or an even simpler item space (see Fig. 1) might reduce distortion effects.

The semantics and pragmatics of scenarios and judgments appear to provide an additional important dimension relevant to issues of causal intransitivity. For example, being hungry (A) causes one to eat (B), which in turn causes one to feel full (C). Here, a transitive inference would suggest, counterintuitively, that being hungry first causes one to feel full. In fact, research on verbally communicated causal relations suggests that people do regard some causal chains as intransitive (Bonnefon et al., 2012; Mayrhofer, Hildenbrand, & Waldmann, 2013). Future research should aim to investigate the conditions under which chains are considered to be transitive.

A further important direction for future research concerns the relationships to different models of (causal) learning. While our investigation of intransitive causal chains was motivated by the postulated central role of the Markov condition in causal Bayes nets (Cartwright, 2002, 2006; Hausman, & Woodward, 1999, 2004; Spohn, 2001; cf. Mayrhofer & Waldmann, 2015; Rehder & Burnett, 2005), an important question is to what extent associative models of learning could account for our findings. Although associative and causal learning differ with respect to important normative and descriptive issues (Goedert, & Spellman, 2005; Waldmann, 1996), there is also some overlap and convergence between associative and probabilistic models of contingency judgment (Chater, 2009; De Houwer & Beckers, 2002; Mitchell, De Houwer, & Lovibond, 2009; Pineño & Miller, 2007). For instance, in line with Marr's (1982) distinction between computational and algorithmic models, the Rescorla–Wagner model of associative learning (Rescorla & Wagner, 1972) converges under specific circumstances on the probabilistic contrast ΔP (Jenkins & Ward, 1965), a prominent measure of statistical contingency or causal strength (Chapman & Robbins, 1990; Chater, 2009; Cheng, 1997; Danks, 2003; Griffiths & Tenenbaum, 2005). Regarding our transitive distortion effects, associative approaches that model updating of associative strength in a pure bottom-up fashion based on directly observable contingencies between events cannot explain our results. However, associative approaches that additionally model "inferred" associations may be able to account for our results (e.g., Baetu & Baker, 2009). Future research on transitive reasoning should aim to investigate the different models and to characterize the relationships among them.

Finally, an important question is whether inferential distortion effects are restricted to causal chains. There is some preliminary evidence that they do not generalize to common-effect structures ($A \rightarrow B \leftarrow C$) with similar positive local $A \rightarrow B$ and $B \leftarrow C$ contingencies and zero contingencies between A and C (von Sydow et al., 2010). This would be expected from a causal Bayes net perspective. Another question concerns common-cause structures, which play a central role in the philosophical criticism of the Markov condition (Cartwright, 2007; Salmon, 1978; Sober, 1987; cf. Hausman & Woodward, 1999, 2004). According to Bayes nets, causal chains and common-cause structures are "Markov equivalent." This suggests identical inferential distortion effects for both structures. Empirically, however, the evidence on the direct psychological validity of the Markov condition for common-cause structures is mixed (Rehder & Burnett, 2005; see also Jarecki et al., 2013; Mayrhofer, Goodman, Waldmann, & Tenenbaum, 2008; Mayrhofer & Waldmann, 2015; Rottman & Hastie, 2013; von Sydow, 2011, 2013). Future research should compare reasoning with different causal structures when the data violate the Markov assumption (von Sydow et al., 2010).

Relations to and differences from other research

Although our results are novel in the causal domain, related findings in other fields point in a similar direction. For example, the Simpson paradox (Simpson,

1951) describes how statistical dependencies can vanish or even be reversed when moving from populations to subpopulations. Some studies (Fiedler, Walther, Freytag, & Nickel, 2003; Waldmann & Hagmayer, 2001) have demonstrated participants' problems in adequately controlling for a third, confounding variable that reverses the relation between two events. Whereas in our experiments participants integrated individual causal links and thereby misjudged the distal relation, participants in the mentioned experiments integrated subpopulations, violating the relation among variables in the overall population.

Other research has shown distortion of zero cue–outcome contingencies, based on high or low base rates of an outcome (Baker, Berbrier, & Vallée-Tourangeau, 1989, Experiment 3; Dickinson, Shanks, & Evandon, 1994; also Buehner, Cheng, & Clifford, 2003). Our results are neutral with regard to such an "outcome density bias," because we used no skewed outcomes that is, $P(A) = P(B) = P(C) = .5$, and, empirically, we did not find distortion effects in the zero-contingency control conditions.

Furthermore, so-called pseudo-contingencies have been discussed (Fiedler & Freytag, 2004; Fiedler, Freytag, & Meiser, 2009; Fiedler, Kutzner, & Vogel, 2013; Meiser & Hewstone, 2004; cf. Kutzner, Vogel, Freytag, & Fiedler, 2011), normally referring to illicit inferences about relations between events based on skewed marginal distributions. For instance, when many students in one class watch a lot of television, and many students in the same class show aggressive behavior, one might infer that students who watch a lot of television tend to be aggressive, even if the events are not correlated. Such pseudo-contingencies, however, are unlikely to apply in our scenarios, as our distributions (including the ones with zero contingency) were not skewed.

Concluding remarks

Our results contribute to a view that emphasizes the role of top-down or knowledge-based inference processes in induction. It has been argued in different fields, such as perception (Gregory, 1980), memory (Loftus & Hoffman, 1989), and language comprehension (Graesser, Singer, & Trabasso, 1994), that top-down processes favoring broadly coherent representations have a substantial and occasionally distorting impact on induction. Overall, our results corroborate the idea that people derive probability estimates by combining single causal links into complex causal models in a modular way (Waldmann et al., 2008). The present results, however, suggest that people base their probability judgments in causal structures not only on bottom-up data—even if observations are directly available during judgments—but also on transitive inferences based on mental causal models that obey the Markov condition, even if transitivity does not hold in the data.

Acknowledgments The work of M. v. S. and the running of the experiments were supported by a grant from the Deutsche Forschungsgemeinschaft (DFG Sy 111/2), as part of the priority program "New Frameworks of Rationality" (SPP 1516). B. M. was supported by grant ME 3717/2 from the same program. Portions of Experiments 1a and 1b were presented at the 2009 Cognitive Science conference in Amsterdam (von Sydow, Meder, & Hagmayer, 2009). We thank Alexander Wendt, Antonia Lange, Alina Greis, Christin Corinth, and Martine Vardar for assistance in data collection and Anita Todd and Martha Cunningham for correcting the manuscript. We are grateful to Ben Newell, Dennis Hebbelmann, Klaus Fiedler, Martha Cunningham, Ralf Mayrhofer, and Michael R. Waldmann for helpful comments on this research.

Appendix 1

Because Experiment 2 was run on a computer with a smaller screen than the paper used in Experiment 1a and 1b (with a DIN A4 page for each of the three panels), we ran an additional item-sensitivity test in Experiment 2 to control for each participant's ability to distinguish feature values of the microbes. The test involved two counterbalancing conditions using different items. For both conditions, participants compared six pairs of microbes according to size ("Is Item 1 bigger or smaller than Item 2?") and six other pairs according to grayscale. The comparisons each concerned five minimal (one-step) differences and one larger (four-step) difference in the 8 × 8 stimulus space (cf. Fig. 7).

The results of the item-sensitivity test showed a reasonable average error rate of 8%. There were no differences in error between size and grayscale (7% vs. 8%), and only small differences between clearly distinguishable (four-step) and hard-to-distinguish (one-step) items (5% vs. 8%). Thus, only a small proportion of errors can plausibly be attributed to an inability to distinguish similar items rather than to general noise. Additionally, the error rate in the item-sensitivity task did not correlate with the number of errors in the local judgments, $r(121) = .07$, or with the proportion of answers in line with our predictions, $r(121) = .09$. Participants' individual ability to differentiate between items does not seem to have mediated the distortion effects we found.

Appendix 2

Table 5 Detailed mean estimates (±*SE*) on the category level in Experiment 2

Condition	Prediction				Results			*N*
	P(*B*|*A*)	*P*(*C*|*B*)	*P*(*C*|*A*)		*P*(*B*|*A*)	*P*(*C*|*B*)	*P*(*C*|*A*)	
			Data	Trans.				
Local–global								
++0	50	50	0	25	57.2 (3.9)	52.8 (3.5)	16.7 (7.5)	36
−−0	−50	−50	0	25	−58.9 (3.1)	−59.4 (3.1)	14.7 (7.5)	38
+0 0	50	50	0	0	63.1 (3.7)	3.7 (4.5)	2.5 (6.8)	32
+++	50	50	25	25	59.1 (3.0)	50.9 (3.1)	35.0 (6.0)	44
−−+	−50	−50	25	25	−57.0 (3.9)	−54.8 (4.1)	38.5 (7.5)	27
−+−	−50	50	−25	−25	−52.6 (3.4)	51.4 (3.1)	−26.3 (6.9)	35
Global–local								
++0	50	50	0	25	51.3 (3.9)	48.6 (3.6)	16.8 (7.6)	37
−−0	−50	−50	0	25	−50.7 (5.1)	−64.3 (3.6)	7.9 (8.1)	28
+0 0	50	50	0	0	46.2 (4.1)	−8.2 (4.5)	−6.9 (6.1)	29
+++	50	50	25	25	51.7 (3.3)	47.3 (3.3)	35.6 (6.1)	41
−−+	−50	−50	25	25	−48.9 (3.6)	−39.3 (4.2)	25.2 (7.8)	27
−+−	−50	50	−25	−25	−57.2 (3.5)	46.2 (3.7)	−40.7 (6.4)	29
All								
++0	50	50	0	25	54.2 (2.7)	50.7 (2.5)	16.7 (5.3)	73
−−0	−50	−50	0	25	−55.4 (2.8)	−61.5 (2.4)	11.8 (5.5)	66
+0 0	50	0	0	0	55.1 (2.9)	−2.0 (3.2)	−2.0 (4.6)	61
+++	50	50	25	25	55.5 (2.3)	49.2 (2.2)	35.3 (4.2)	85
−−+	−50	−50	25	25	−52.9 (2.7)	−47.0 (3.1)	31.8 (5.4)	54
−+−	−50	50	−25	−25	−54.7 (2.4)	49.1 (2.4)	−32.8 (4.8)	64

Note. In the local–global conditions, participants rated the individual causal links before judging the conditional probability of *C* given *A*. In the global–local conditions, this order was reversed. All = Mean judgments aggregated across question order. Data = conditional probability as entailed by the data. Trans. = the conditional probability as estimated transitively. All conditional probability judgments are presented here using a scale of −100 to +100 although the scale was not explicitly numbered (see Fig. 8).

References

Ahn, W., & Dennis, M. (2000). Induction of causal chains. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the cognitive science society* (pp. 19–24). Mahwah, NJ: Erlbaum.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 415,* 147–149. doi:10.3758/BF03334492

Arntzenius, F. (2005). Reichenbach's common cause principle. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from http://plato.stanford.edu/archives/win2008/entries/physics-Rpcc/

Baetu, I., & Baker, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes, 35,* 153–168. doi:10.1037/a0013764

Baker, A. G., Berbrier, M. W., & Vallée-Tourangeau, F. (1989). Judgments of a 2 × 2 contingency table: Sequential processing and the learning curve. *Quarterly Journal of Experimental Psychology, 41B,* 65–97. doi:10.1080/14640748908401184

Bonnefon, J.-F., Da Silva Neves, R., Dubois, D., & Prade, H. (2012). Qualitative and quantitative conditions for the transitivity of perceived causation. *Annals of Mathematics and Artificial Intelligence, 64,* 311–333. doi:10.1007/s10472-012-9291-0

Buehner, M., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1119–1140. doi:10.1037/0278-7393.29.6.1119

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist, 84,* 242–264. doi:10.5840/monist20018429

Cartwright, N. (2002). Against modularity, the causal Markov condition, and the link between the two: Comments on Hausman and Woodward. *British Journal for the Philosophy of Science, 53,* 411–453. doi:10.1093/bjps/53.3.411

Cartwright, N. (2006). From metaphysics to method: Comments on manipulability and the causal Markov condition. *British Journal for the Philosophy of Science, 57,* 197–218. doi:10.1093/bjps/axi156

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511618758

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18,* 537–545. doi:10.3758/BF03198486

Chater, N. (2009). Rational and mechanistic perspectives on reinforcement learning. *Cognition, 113,* 350–364. doi:10.1016/j.cognition.2008.06.014

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

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology, 47,* 109–121. doi:10.1016/S0022-2496(02)00016-0

De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *The Quarterly Journal of Experimental Psychology. B, 55,* 289–310. doi:10.1080/02724990244000034

Dickinson, A., Shanks, D., & Evandon, J. (1994). Judgement of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 36A,* 29–50. doi:10.1080/14640748408401502

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29,* 103–130. doi:10.1023/A:1007413511361

Evans, S. B. T., & Over, D. E. (2004). *If.* Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198525134.001.0001

Fiedler, K., & Freytag, P. (2004). Pseudocontingencies. *Journal of Personality and Social Psychology, 87,* 453–467. doi:10.1037/0022-3514.87.4.453

Fiedler, K., Freytag, P., & Meiser, T. (2009). Pseudocontingencies: An integrative account of an intriguing cognitive illusion. *Psychological Review, 116,* 187–206. doi:10.1037/a0014480

Fiedler, K., Kutzner, F., & Vogel, T. (2013). Pseudocontingencies–Logically unwarranted but smart inferences. *Current Directions in Psychological Science*, 1–6. doi:10.1177/0963721413480171

Fiedler, K., Walther, E., Freytag, P., & Nickel, S. (2003). Inductive reasoning and judgment interference: Experiments on Simpson's paradox. *Personality and Social Psychology Bulletin, 29,* 14–27. doi:10.1177/0146167202238368

Goedert, K. M., & Spellman, B. A. (2005). Non-normative discounting: There is more to cue-interaction effects than controlling for alternative causes. *Learning & Behavior, 33,* 197–210. doi:10.3758/BF03196063

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101,* 371–395. doi:10.1037/0033-295X.101.3.371

Gregory, R. L. (1980). Perceptions as hypothesis. *Philosophical Transactions of the Royal Society of London, Series B, 290,* 181–197. doi:10.1098/rstb.1980.0090

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

Hagmayer, Y., Meder, B., von Sydow, M., & Waldmann, M. R. (2011). Category transfer in sequential causal learning: The unbroken mechanism hypothesis. *Cognitive Science, 35,* 842–873. doi:10.1111/j.1551-6709.2011.01179.x

Hausman, D., & Woodward, J. (1999). Independence, invariance, and the causal Markov condition. *British Journal for the Philosophy of Science, 50,* 521–583. doi:10.1093/bjps/50.4.521

Hausman, D., & Woodward, J. (2004). Modularity and the causal Markov condition: A restatement. *British Journal for the Philosophy of Science, 55,* 147–167. doi:10.1093/bjps/55.1.147

Hebbelmann, D., & von Sydow, M. (2014). Betting on transitivity in an economic setting. *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society* (pp. 2339–2344). Austin, TX: Cognitive Science Society.

Jarecki, J., Meder, B., & Nelson, J. D. (2013). The assumption of class-conditional independence in category learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 2650–2655). Austin, TX: Cognitive Science Society.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79,* 1–17. doi:10.1037/h0093874

Kutzner, F., Vogel, T., Freytag, P., & Fiedler, K. (2011). Contingency inferences driven by base rates: Valid by sampling. *Judgment and Decision Making, 6,* 211–221.

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

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40,* 87–137. doi: 10.1006/cogp.1999.0724

Loftus, E. F., & Hoffman, H. G. (1989). Misinformation and memory: The creation of memory. *Journal of Experimental Psychology: General, 118,* 100–104. doi:10.1037/0096-3445.118.1.100

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

Mayrhofer, R., Goodman, N. D., Waldmann, M. R., & Tenenbaum, J. B. (2008). Structured correlation from the causal background. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 303–308). Austin, TX: Cognitive Science Society.

Mayrhofer, R., Hildenbrand, I., & Waldmann, M. R. (2013). Causal dispositions and transitivity in causal chains. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (p. 4041). Austin, TX: Cognitive Science Society.

Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science, 39,* 65–95. doi:10.1111/cogs.12132

Meder, B., Mayrhofer, R., & Waldmann, M. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121,* 277–301. doi:10.1037/a0035944

Meiser, T., & Hewstone, M. (2004). Cognitive processes in stereotype formation: The role of correct contingency learning for biased group judgments. *Journal of Personality and Social Psychology, 87,* 599–614. doi:10.1037/0022-3514.87.5.599

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32,* 183–198. doi:10.1017/S0140525X09000855

Oberauer, K., Weidenfeld, A., & Fischer, K. (2007). What makes us believe a conditional? The roles of covariation and causality. *Thinking and Reasoning, 13,* 340–369. doi:10.1080/13546780601035794

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

Pineño, O., & Miller, R. R. (2007). Comparing associative, statistical, and inferential accounts of human contingency learning. *Quarterly Journal of Experimental Psychology, 60,* 310–329. doi:10.1080/17470210601000680

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50,* 264–314. doi:10.1016/j.cogpsych.2004.09.002

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

Rottman, B. M., & Hastie, R. (2013). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140,* 109–139. doi:10.1037/a0031903

Salmon, W. (1978). Why ask, "Why?"? *Proceedings of the American Philosophical Association, 51,* 683–705. doi:10.2307/3129654

Seels, S. B. (1936). The atmosphere effect: An experimental study of reasoning. *Archives of Psychology, 200,* 1–72.

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

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195183115.001.0001

Sober, E. (1987). The principle of the common cause. In J. Fetzer (Ed.), *Probability and causality* (pp. 211–228). Dortrecht, The Netherlands: Reidel.

Sober, E. (2001). Venetian sea levels, British bread prices, and the principle of the common cause. *British Journal for the Philosophy of Science, 52,* 331–346. doi:10.1093/bjps/52.2.331

Sober, E., & Steel, M. (2012). Screening-off and causal incompleteness: A no-go theorem. *British Journal for the Philosophy of Science, 52,* 1–38. doi:10.1093/bjps/axs021

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York, NY: Springer. doi:10.1007/978-1-4612-2748-9

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), *Stochastic dependence and causality* (pp. 157–172). Stanford, CA: CSLI.

Steel, D. (2006). Comment on Hausman & Woodward on the causal Markov condition. *British Journal for the Philosophy of Science, 57,* 219–231. doi:10.1093/bjps/axi154

von Sydow, M. (2011). The Bayesian logic of frequency-based conjunction fallacies. *Journal of Mathematical Psychology, 55,* 119–139. doi:10.1016/j.jmp.2010.12.001

von Sydow, M. (2013). Logical patterns in individual and general predication. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 3693–3698). Austin, TX: Cognitive Science Society.

von Sydow, M., Hagmayer, Y., Meder, B., & Waldmann, M. (2010). How causal reasoning can bias empirical evidence. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 2087–2092). Austin, TX: Cognitive Science Society.

von Sydow, M., Meder, B., & Hagmayer, Y. (2009). A transitivity heuristic of probabilistic causal reasoning. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 803–808). Austin, TX: Cognitive Science Society.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning* (pp. 47–88). San Diego, CA: Academic Press. doi:10.1016/s0079-7421(08)60558-7

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

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82,* 27–58. doi:10.1016/S0010-0277(01)00141-X

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology, 53,* 27–58. doi:10.1016/j.cogpsych.2006.01.001

Waldmann, M. R., Meder, B., von Sydow, M., & Hagmayer, Y. (2010). The tight coupling between category and causal learning. *Cognitive Processing, 11,* 143–158. doi:10.1007/s10339-009-0267-x

White, P. A. (2003). Making causal judgements from contingency information: The pCI rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 710–727. doi:10.1037/0278-7393.29.4.710