Cogn Process (2010) 11:143–158
DOI 10.1007/s10339-009-0267-x

RESEARCH REPORT

The tight coupling between category and causal learning

Michael R. Waldmann · Björn Meder · Momme von Sydow · York Hagmayer

Received: 18 November 2008 / Accepted: 4 June 2009 / Published online: 27 June 2009

Abstract The main goal of the present research was to demonstrate the interaction between category and causal induction in causal model learning. We used a two-phase learning procedure in which learners were presented with learning input referring to two interconnected causal relations forming a causal chain (Experiment 1) or a common-cause model (Experiments 2a, b). One of the three events (i.e., the intermediate event of the chain, or the common cause) was presented as a set of uncategorized exemplars. Although participants were not provided with any feedback about category labels, they tended to induce categories in the first phase that maximized the predictability of their causes or effects. In the second causal learning phase, participants had the choice between transferring the newly learned categories from the first phase at the cost of sub-optimal predictions, or they could induce a new set of optimally predictive categories for the second causal relation, but at the cost of proliferating different category schemes for the same set of events. It turned out that in all three experiments learners tended to transfer the categories entailed by the first causal relation to the second causal relation.

M. R. Waldmann (✉) · B. Meder · M. von Sydow · Y. Hagmayer
Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany
e-mail: michael.waldmann@bio.uni-goettingen.de

Y. Hagmayer
e-mail: york.hagmayer@bio.uni-goettingen.de

B. Meder
Max Planck Institute for Human Development, Berlin, Germany

Introduction

Cognitive psychology has a tendency to compartmentalize research into different areas, such as memory, learning, decision-making, or categorization. Unfortunately, there is little contact between these fields. Although in each of these areas a wealth of theoretical and empirical knowledge has been gathered in the past decades, this strategy of *divide and conquer* led to notable blind spots. For example, causal knowledge plays an important role in learning, categorization, perception, decision-making, problem solving, and text comprehension. In each of these fields separate theories have been developed to investigate the role of causal knowledge. However, it remains unclear how these theories and empirical findings can be united. If causal knowledge underlies decision-making, for example, then it seems plausible to assume that our learning should be sensitive to this important goal (see Hagmayer and Sloman 2009; Meder and Hagmayer 2009).

Categorization and causal learning: the neglect of their tight coupling

Our aim in the present research project was to close the gap between two of these areas in which causality plays a crucial role, learning and categorization. Although the relevance of causal knowledge has been highlighted in both of these areas, their tight coupling has not received much attention until recently (see Lien and Cheng 2000; Kemp et al. 2007; Marsh and Ahn 2009; Waldmann and Hagmayer 2006).

Research on *causal and associative learning* has typically neglected the role of categorization by using tasks in which the stimuli are already pre-categorized. For example, Waldmann (2000, 2001) presented subjects with learning

tasks in which they had to acquire knowledge about contingencies between fictitious substances, which either could be present or absent, and the presence or absence of a novel disease, *Midosis*. Thus, both cues and outcomes were already categorized prior to learning so that the only remaining task was to learn about their statistical relations (see Shanks et al. 1996; De Houwer and Beckers 2002, for other examples). Although this research has yielded many interesting insights about learning, the role of categories is neglected. But our causal learning input does not always come with pre-categorized causes and effects. We often are confronted with exemplars of causes (e.g., different exemplars of biological entities) and exemplars of effects (e.g., different symptom patterns in a set of patients) so that we need to simultaneously learn about categories of causes, categories of effects, and the statistical relations indicating causal relations between these categories.

A similar shortcoming characterizes research on *categorization*. Although traditional similarity-based theories have been augmented by the theory-based view, which claims that natural concepts are often grounded in causal knowledge (Murphy and Medin 1985; see also Murphy 2002), the tight coupling between causal and category learning has largely been neglected. Instead, research in this area has mainly focused on how knowledge about the internal causal structure connecting the features of category members affects categorization. For example, disease categories frequently refer to common-cause models of diseases with the category features representing causes (e.g., viruses) and effects (e.g., symptoms). Many studies using these and similar materials have shown that the type of causal model connecting otherwise identical cause and effect features influences learning, typicality judgments, or induction (Rehder 2003a, b; Rehder and Hastie 2001, 2004; Waldmann et al. 1995; Waldmann 1996, 2000, 2001). In these studies, however, cause and effect features within the causal models were again treated as fixed, pre-categorized entities, which already existed prior to the learning or reasoning context. The fact that these feature categories may need to be induced from exemplar information provided in the learning input has not been acknowledged until fairly recently.

The interaction of category and causal learning

One of the first studies investigating the tight coupling between category and causal learning was by Lien and Cheng (2000). In their experiments, Lien and Cheng presented exemplars to learners, which could be classified by different features at different hierarchical levels of abstraction. While in traditional supervised category learning studies explicit category feedback is provided, in Lien and Cheng's experiments no category labels were presented. Instead participants were provided only with information regarding which exemplars generated a specific causal effect and which did not. Learners in the experiments received pictures of substances that varied in color and shape along with information about which of these substances made flowers bloom, and which failed to do so. The results showed that learners categorized the cause events (substances) at the hierarchical level that were maximally predictive for the effect (i.e., blooming). Thus, the induced category scheme was determined by its suitability for predicting the effect. Lien and Cheng (2000) interpreted this as evidence for their *maximal-contrast hypothesis*: People tend to induce categories that maximize their causal predictability (i.e., contingencies).

Kemp et al. (2007) have developed a computational model implementing simultaneous learning of causal relations and categories ("causal schemata") in the absence of feedback about category labels. Following the basic idea of Lien and Cheng (2000) the model induces categories of causes and effects that allow to optimally predict the effect categories based on the cause categories.

Recently, Marsh and Ahn (2009) have also investigated the question of how causal learning may influence category formation in the absence of explicit category feedback. In a series of experiments they presented participants with cause events that varied on a continuous dimension (e.g., high, intermediate, low). In the experiments Marsh and Ahn manipulated the assignment of the causes to a binary effect. For example, in one condition high and intermediate values may cause the effect, whereas in the contrast condition, only high values but not the two others are causal. The results show that learners tended to categorize the causes according to the boundaries created by the effect. Thus, the ambiguous intermediate value was classified together with the high value when both caused the effect; otherwise it was classified with the low value. These categories affected both contingency and similarity judgments. These results can be predicted by Lien and Cheng's (2000) theory in that they show that learners try to induce categories that maximize predictability. Interestingly, in further studies Marsh and Ahn (2009) have shown that these classifications remained stable even when the mapping between the ambiguous exemplars and the effect was later changed.

In all of the studies mentioned above the interaction of causal and category learning was investigated with respect to a single causal relation. Optimality can easily be defined as maximal predictiveness as long as only one cause-effect relation is considered (Lien and Cheng 2000). The situation is more complex when the same events are involved in multiple causal relations. In these situations every causal relation may entail a different optimal categorical scheme. Figure 1 gives two examples of such situations. The first

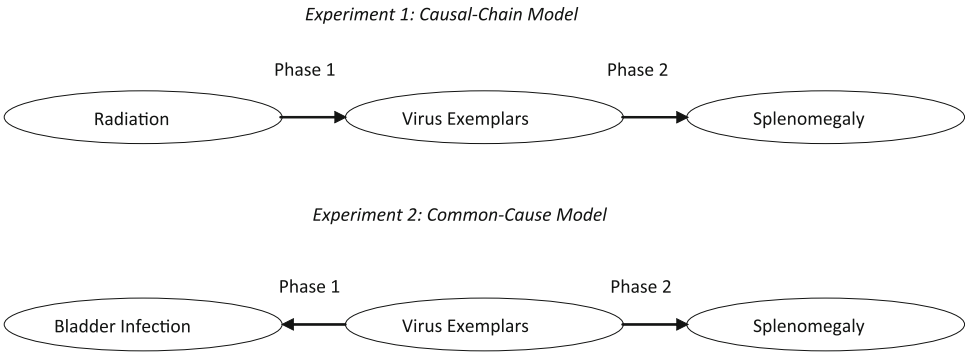

Fig. 1 Causal models in Experiments 1 and 2

model shows a causal chain in which the initial cause, radiation, influences the middle event, viruses, which in turn cause a swelling of the spleen (i.e., splenomegaly). The second model presents three events in a common-cause structure in which the viruses play the role of a common cause for its two effects, bladder infection and splenomegaly. In both causal models the virus exemplars are part of two different causal relations.

Assuming that the viruses are presented as a set of uncategorized exemplars, the question arises how categories should be induced on the basis of causal information. Using one of the two relations to induce categories yielding maximal contrasts will lead to optimal categories for this single relation, but these categories may not necessarily be optimal for the second relation. Thus, for example, in the causal-chain model the presence or absence of radiation may be used to induce virus categories which yield maximal statistical contingencies for radiation. But these newly induced virus categories may not generate maximal contingencies with respect to the final effect. If both relations are learned at once, it may be possible to induce categories that are globally optimal for predicting both related events, although they may not be locally optimal with respect to either.

This question is of particular importance since causal knowledge is rarely learned at once; we rather acquire it in fragments which later are tied together within complex causal models (see Lagnado et al. 2007; Waldmann et al. 2006, 2008; von Sydow et al. 2009). For example, learners may first learn about a single causal relation which determines how the events referring to causes or effects are categorized. In a later learning phase, the very same exemplars might be presented within a second relation. Would learners now continue to use the initially induced categories at the potential cost of suboptimal predictions with respect to the second, later learned relation, or would they induce a new set of categories for the second causal relation? For example, in the common-cause model learners could induce one set of virus categories which is optimal for predicting bladder infection, and a different set of virus categories for optimally predicting splenomegaly. Inducing two sets of categories would yield maximal contrasts for either effect but at the cost of having to activate different category sets for different effects.

Waldmann and Hagmayer (2006) presented a first set of studies investigating whether learners would transfer explicitly learned categories to a subsequent causal learning task in which cause exemplars were paired with effects. For example, they trained participants in a first learning phase (Phase 1) to categorize images representing different virus exemplars. They used a standard category learning paradigm with feedback in which category labels ("allovedic" vs. "hemovedic") were provided. Thus, in this phase learners acquired knowledge about two types of virus categories. In the second learning phase (Phase 2) the same virus exemplars were presented (without category labels) and paired with causal effect information (presence or absence of splenomegaly). Thus, Phase 2 closely resembled the task in the study of Lien and Cheng (2000) (see also Marsh and Ahn 2009). If only this learning phase was presented, categories should be induced which, according to Lien and Cheng, are maximally predictive for the effects. However, alternatively learners might also opt to continue to use the virus categories acquired in the initial learning phase, and transfer them to the subsequent causal learning, despite the fact that the previously learned categories were not optimal for predicting the effect in the causal learning phase. The results showed that learners often transferred category schemes to the second causal learning phase, rather than inducing new ones. Especially when the categories referred to natural kinds (e.g., viruses) learners tended to stick with the category scheme learned in Phase 1 rather than replacing it with a second set of categories re-classifying the same virus exemplars. One explanation is that people seem to believe that natural kinds refer to stable entities in the world even when the causal relations in which they are involved are only probabilistic (see also the General discussion).

Aim of the present studies

Both Lien and Cheng (2000) and Marsh and Ahn (2009) have shown that learners induce cause categories involved

in single causal relations on the basis of feedback about effects (i.e., effect-based categories). We will extend this research by investigating both effect-based categories (Experiments 2a, b) and cause-based categories (Experiment 1). The primary goal of our research is to go beyond the focus of Lien and Cheng (and Marsh and Ahn) on single causal relations, and investigate whether categories induced within causal learning contexts are transferred to other causal relations which overlap with the causal relation that was the basis of the induction of the categories. Thus, our focus will be on the interaction of category and causal learning when participants consecutively acquire knowledge about complex causal models. While Waldmann and Hagmayer (2006) have studied transfer in a task in which category learning was based on explicit feedback about category labels, the present studies will study the interplay between category and causal learning in tasks that provide only learning input about causes and effects rather than about category labels.

In order to study how learners induce categories and transfer them, we confronted participants in consecutive learning phases with causal models consisting of two causal links: causal chains (Experiment 1) and common cause models (Experiments 2a, b). No information about category labels was presented to participants; they only received information about exemplars and their causes or effects. Based on the research by Lien and Cheng (2000) we expected learners to induce maximally predictive categories in the first learning phase in which only a single causal relation was presented. Our main novel goal was to study whether learners would transfer these categories to a second learning phase in which a partially overlapping causal relation was presented. Importantly, the categories entailed by the first causal learning phase were somewhat but not optimally predictive for the novel event presented in the second causal relation. This setup allowed us to find out whether participants in fact induced categories during the first causal learning phase and transferred them to the second learning phase. If they do, we should see an effect of the categories entailed by the first causal relation upon causal judgments referring to the second relation. By contrast, if participants prefer to induce novel categories for each causal relation, we should not see an effect of the first causal learning phase on judgments about the second relation.

Experiment 1

Experiment 1 investigated learning about a causal chain whose two links were learned consecutively (see Fig. 1). The middle event of the chain was an uncategorized set of virus exemplars which were causally linked to two binary pre-categorized causal events. The experimental paradigm consisted of two consecutive learning phases corresponding to the two causal relations: In the first phase (Phase 1) participants had to learn about the causal relation between the initial cause (two types of radiation) and the intermediate events (a set of uncategorized virus exemplars). No reference to categories or classes was made, learners simply observed causes and effect exemplars. Then participants proceeded to the second phase in which they had to learn about the causal relation between the intermediate exemplars and a final dichotomous effect (Phase 2). After Phase 2, a test phase was administered in which participants were presented with exemplars belonging to the intermediate event in the chain ("viruses") and then asked to assess their causal efficacy with respect to the final effect ("splenomegaly", which is a swelling of the spleen). The conditions were designed to reveal whether participants used the categories induced during their learning of the first causal relation (between the initial and the intermediate event) when learning about and assessing the second relation (between the intermediate and the final event) (see below for details).

In order to test whether the categories entailed by the first learning phase have a lasting or only a transient effect on causal judgments with respect to the second relation, we repeated Phase 2 and the subsequent test phase. If additional learning led to the induction of novel categories optimal for predicting the effect in the Phase 2, we should see a smaller effect of the categories induced in Phase 1 after the second iteration of Phase 2.

The initial cause of the causal chain consisted of two types of radiation affecting the DNA of viruses, thereby creating new viruses (the intermediate event in the chain). The viruses were depicted schematically and had four binary features: brightness (light vs. dark), size (large vs. small), number of corners (five vs. seven), and number of surface molecules (two vs. four) (see Fig. 2 for the resulting 16 exemplars). The relation between these virus exemplars and the disease ("splenomegaly") was used as the second causal link. An example of the materials and the statistical structure of the learning data are given in Figs. 2 and 3.

The first causal relation involved a linearly separable family resemblance structure. Radiation of type Alpha caused prototypically small, light viruses with few corners and few surface molecules, while Beta radiation typically caused large, dark viruses with many corners and many surface molecules. Two experimental conditions and one control condition were used (see Fig. 2 for the experimental conditions). In Condition A, Alpha radiation caused exemplars with at least three out of four features of the small light prototype (Item 0000, cf. Fig. 4), whereas Beta radiation caused exemplars with at least two out of four

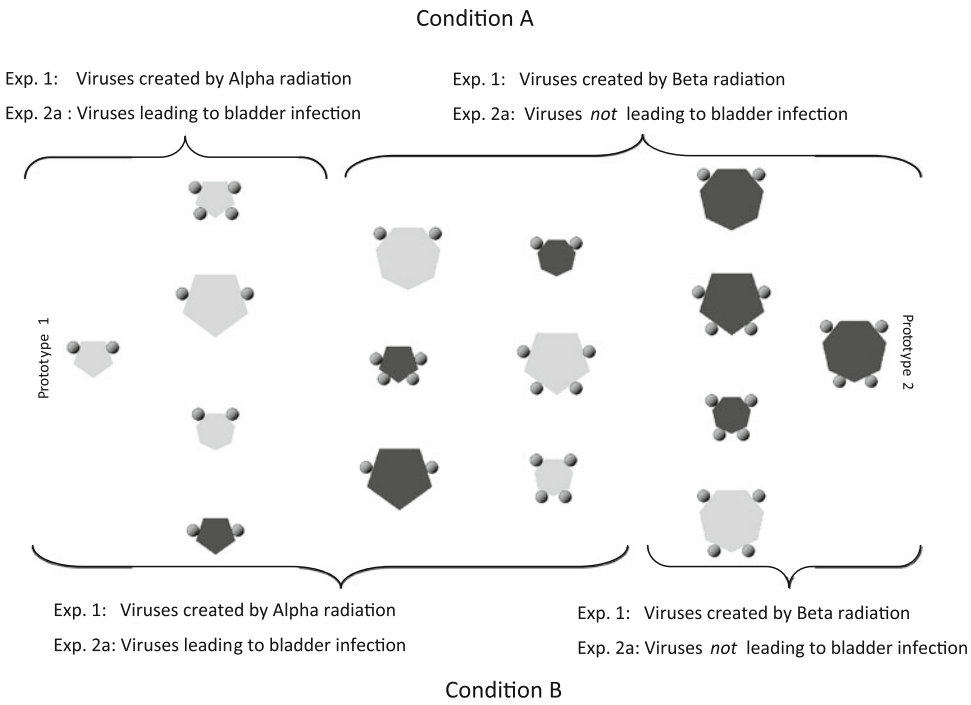

Fig. 2 Categories entailed by the first causal learning phase of Experiments 1 and 2a. In Experiment 1, cause-based categories are entailed by the first learning phase (i.e., virus exemplars can be categorized according to their causes, namely different types of radiation). In Experiment 2a, effect-based categories are entailed (i.e., viruses can be categorized according to an initially learned effect, bladder infection). See text for details

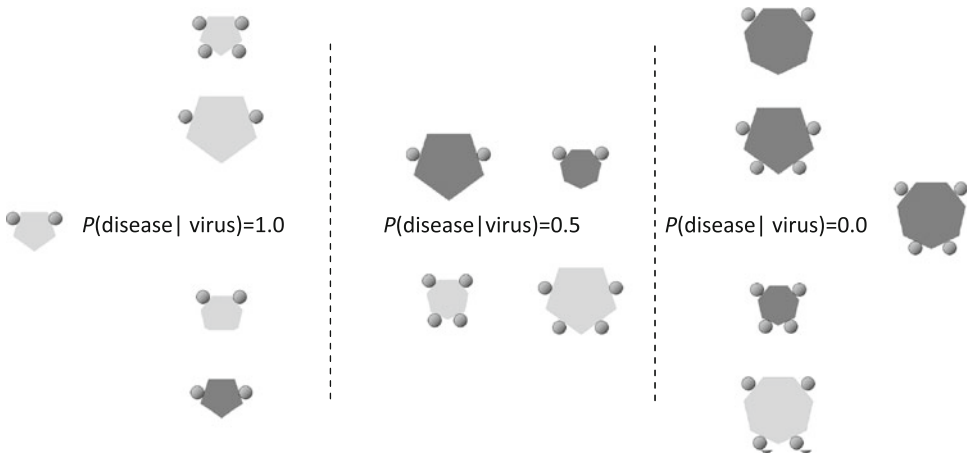

Fig. 3 Categories entailed by the causal relation in the second learning phase (Experiments 1 and 2a)

features of the big dark prototype (Item 1111). In Condition B, the boundary between these causal categories was moved. Now Alpha radiation caused exemplars with at least two out of four features of the small light prototype, and Beta radiation caused exemplars with at least three out of four features of the big dark prototype. As a consequence, Exemplars 6–11, which have two features in common with each of the two prototypes, either belonged to the small, light prototype category (Condition A) or the large, dark prototype category (Condition B) (cf. Fig. 2). In the control condition participants did not receive learning input about the first causal relation.

In Phase 2 participants were requested to learn a second causal relation that was identical in all three conditions (see Fig. 3). Overall, half of the viruses caused the disease splenomegaly, the other half did not. Viruses that shared at least three features with the small light prototype deterministically caused the disease ("splenomegaly"), viruses that shared two features with each prototype caused the disease with a probability of 0.5, and viruses that shared at least three out of four features with the big dark prototype never caused the disease ("splenomegaly"). The most important prediction for examining transfer effects involves the six items that were equally similar to either of the two prototypes (Items 6–11). In the test phase learners were requested to assess the causal efficacy of these items regarding the final effect in the causal chain (i.e., splenomegaly). If learners based these estimates on the causal categories induced in the first learning phase, a particular inference pattern should be obtained for the two contrasted

conditions (A vs. B). In Condition A, participants should reason that the intermediate exemplars belong to the class of viruses generated by Beta radiation, which is only weakly associated with the generation of the disease. In this category, only two out of nine viruses created by Beta radiation lead to splenomegaly. In Condition B, by contrast, these exemplars belong to the category of viruses caused by Alpha radiation. Seven of the nine viruses in this causal category did generate the disease. Thus, if estimates of causal efficacy are derived from the causal categories induced in Phase 1, these exemplars should be assigned a high probability of causing splenomegaly.

Instead of transferring the category scheme from Phase 1 to Phase 2, participants could alternatively induce new categories that are maximally predictive for the causal effect in the second phase (Lien and Cheng 2000). If participants chose to optimize local predictability in this way, they should base their judgment on the category structure entailed by the relation to the second effect, splenomegaly. As can be seen from Fig. 3, a tripartite category structure underlies the second causal relation: viruses that always cause the effect, viruses that never cause splenomegaly, and viruses which cause the disease half of the time. In addition, the items in the middle zone are equally similar to both prototypes. Therefore, participants should give a 50% rating for these exemplars if they optimize local predictability and ignored the initially induced categories. This is also what is predicted for the control condition, in which no first learning phase was administered.

To sum up, we expected participants in the two experimental conditions to transfer the categories induced in Phase 1 to causal learning in Phase 2, thus opting for global coherence rather than maximization of local predictability. Hence, participants should give low ratings for the critical middle items in Condition A and high ratings in Condition B. In the control condition, in which categories are based on the second causal relation only, ratings around 50% should be observed. These differences should hold despite the fact that all participants are provided with identical learning input in the second causal learning phase.

Method

Participants and design

Forty-eight students from the University of Göttingen, Germany, participated in the experiment for course credit. They were randomly assigned to one of the three conditions. The two experimental conditions (A and B) presented two consecutive learning phases about a causal chain, and varied the category boundary entailed by the initial cause of a causal chain. In the control condition participants received no first learning phase.

Procedure and materials

The two experimental conditions consisted of two consecutive causal learning phases, each followed by a test phase. In Phase 1, participants were told that scientists had discovered that the exposure of DNA to radiation creates new viruses and that the new viruses vary in brightness, size, shape, and number of molecules on the surface. Furthermore it was mentioned that it has been discovered that Alpha and Beta rays generate different viruses. Then a figure was presented depicting the viruses caused by both kinds of radiation. This figure varied across conditions. In Condition A, all viruses sharing at least three features with the small, light prototypical virus with few corners and few surface molecules (Prototype 1) were caused by Alpha radiation. All other viruses were caused by Beta radiation. The viruses caused by Beta radiation had at least two features in common with the large, dark prototype having many corners and many surface molecules (Prototype 2). In Condition B, the mapping between radiation and appearance of the viruses was altered. Here all viruses sharing at least two features with Prototype 1 were generated by Alpha radiation and all other viruses by Beta radiation. Figure 2 illustrates the two conditions and the respective category boundaries. The assignment of the radiation labels (Alpha, Beta) and viruses were counterbalanced across the two conditions.

After participants familiarized themselves with the viruses and their relation to the radiation, the figures were removed and participants were shown a pile of index cards one after another. On the front side of each card participants were informed about the type of radiation and then on the card's backside about the resulting virus. Participants received eight randomized blocks with 16 trials each, resulting in a total of 128 cards. As a check of learning, participants were presented with the 16 different viruses at the end of Phase 1, and asked to diagnose the type of radiation which created them. If participants were not able to correctly say for each exemplar by which type of radiation it was caused, they were told that their judgments concerning the causal relations were not entirely correct, and an additional 64 trials were administered followed by the same test. Participants were only allowed to proceed to the second phase if they made no mistakes this time. Unlike in the two experimental conditions, there was no Phase 1 learning in the control condition. In the first phase of the control condition, participants were only informed that researchers had discovered new viruses and were shown a figure depicting the 16 viruses in random order.

While the first causal learning phase varied across conditions, the second causal learning phase was identical in all three conditions. Participants were instructed that veterinarians had investigated whether the newly created

viruses were causing splenomegaly. It was pointed out that every result was possible: the viruses might always cause splenomegaly, they might only sometimes cause splenomegaly, or they might be completely unrelated to the disease. Data about the relation between the viruses and splenomegaly were again shown on index cards (in random order). On each trial participants were first shown a virus and then were informed whether splenomegaly had occurred or not. An example of the data structure is shown in Fig. 3 (cf. Fig. 4). In the particular counterbalancing condition shown in the figure all viruses sharing at least three of four features with Prototype 1 always caused the disease, and all viruses sharing at least three out of four features with Prototype 2 never caused splenomegaly. Finally, viruses sharing two features with each prototype caused splenomegaly with a probability of 0.5. To ensure that all features were equally correlated with the disease it was necessary to omit two exemplars from this learning phase (Items 8 and 9, cf. Fig. 4). Hence, two of the items in this category always caused the effect and two items never caused the effect. The two remaining items were not presented to participants during learning. All six items were used as test exemplars for the middle category. The assignment of the disease to the viruses was counterbalanced by reversing the mapping between viruses causing the disease and not causing it.

After 42 trials (3 blocks of 14 viruses) a first test phase was administered. Participants' estimates of the causal efficacy of each of a set of viruses were collected. 12 of the 16 possible viruses were presented in random order: the two prototypes, one virus sharing three features with Prototype 1, one virus sharing three features with Prototype 2, and all six critical viruses that shared two features with each prototype. For each item participants were asked to rate the probability that the particular virus causes splenomegaly using a scale ranging from 0 ("the virus never causes splenomegaly") to 100 ("the virus always causes splenomegaly"). Subsequent to collecting these ratings, learners were requested to rate the causal efficacy of viruses caused by Alpha radiation and viruses caused by Beta radiation. This question was administered to directly tap onto participants' knowledge about the causal relation between the causal categories entailed by the first learning phase and the final effect of the causal chain. The same rating scale as before was used. Participants received no feedback regarding the accuracy of their estimates.

After giving these estimates, participants were shown an additional 42 learning trials that exhibited the same data structure as the previous trials. Subsequently, estimates about the causal impact of the viruses were collected a second time using the same procedure as before. The rationale underlying the repeated measure was to explore what would happen when more learning data about the second causal relation becomes available. Basically, there are two hypotheses regarding the transfer of the category structure from Phase 1. First, it may be the case that these categories are only initially used but tend to be abandoned when more learning data becomes available in Phase 2.

Items					Phase 1			Phase 2	Test Phase
Item	Size	Color	Shape	Molec.	Cond. A	Cond. B	Control		
1	1	1	1	1	Exp. 1: Caused by Beta-Rdiation Exp. 2a: Bladder Infection	Beta-Rdiation B. Infection	(No first learning phase)	—E	T
2	1	1	1	0				—E	
3	1	1	0	1				—E	T
4	0	1	1	1				—E	
5	1	0	1	1				—E	
6	1	1	0	0		Exp. 1: Caused by Alpha-Rdiation Exp. 2a: No Bladder Infection		—E	T
7	0	0	1	1				—E	T
8	1	0	1	0				?	T
9	0	1	0	1				?	T
10	0	1	1	0				E	T
11	1	0	0	1				E	T
12	0	0	0	1	Alpha-Rdiation No B. Infection			E	
13	0	0	1	0				E	
14	1	0	0	0				E	T
15	0	1	0	0				E	
16	0	0	0	0				E	T

Fig. 4 Learning input in Experiments 1 and 2a. The four binary features are coded as follows: size: *1* large, *0* small; color: *0* light, *1* dark; shape: *0* five corners, *1* seven corners; molecules: *0* two molecules, *1* four molecules. Regarding Phase 2, "E" and "−E" denote the presence and absence of the causal effect. "T" denotes the items presented in the test phase

Thus, in the course of Phase 2 participants may induce the tripartite, maximally predictive category structure, or make use of exemplar knowledge. In both cases, this would entail that possible transfer effects would vanish. However, it may alternatively be the case that participants recognize over the course of learning that the initially induced categories are sufficiently good for making predictions. In this case, a potential category effect may remain stable or become even more pronounced after more trials (i.e., at the second measurement).

Of the 12 viruses that participants rated, six were critical for testing category transfer from Phase 1. These "critical items" shared two features with each prototype. Depending on the condition, participants were expected to rate the likelihood that these viruses caused the effect differently. If participants induced causal categories and used them to derive inferences in the test phase they should give ratings lower than 50 in Condition A and ratings higher than 50 in Condition B. Ratings around 50 should result if participants based their judgments only on the data from the second learning phase (i.e., in the control condition). Estimates for the other viruses should not differ between conditions. Prototype 1 and the virus sharing three features with this prototype should be assigned high causal efficacy ("uncritical-high items"). Conversely, Prototype 2 and its nearest neighbors should be seen as not causing the disease ("uncritical-low items"). As these exemplars always belonged to the categories having a high and low probability of generating the effect, respectively, estimates regarding these items should not differ across conditions.

Results and discussion

Regarding the first causal learning phase (radiation → viruses), 14 out of 32 participants in the experimental conditions made errors when first tested, and were therefore shown four additional blocks of learning trials. After this additional training all participants were able to correctly diagnose the type of radiation that created the viruses. Therefore, all participants proceeded to the second learning phase. For the analyses, the ratings in the counterbalancing conditions were recoded to fit the example described in the Introduction and the Procedure sections. Estimates were averaged over the two uncritical-high items, the two uncritical-low items, and the six critical items. The mean ratings of causal efficacy are shown in Table 1, separately for the two measurements after 42 and 84 trials, respectively. A first inspection of these data indicates that learners' estimates for the critical items seem to be affected by the category structure underlying the initial learning phase. Learners tended to give ratings below 50 in Condition A and ratings above 50 in Condition B, with the control condition being very close to 50. Interestingly, the difference between the two experimental conditions increased in the course of learning, which can be seen when comparing participants' estimates of causal efficacy of the critical items at the first and second measurement (Table 1). The judgment pattern obtained for the uncritical items also qualitatively confirms our predictions: the viruses belonging to the categories having a high or low probability of generating the effect received high and low ratings, respectively, regardless of condition.

Separate analyses were conducted for critical and uncritical items. First, we analyzed the uncritical items using an analysis of variance with the variables exemplar type (uncritical-high vs. uncritical-low) and measurement (first vs. second) as within-subjects variables and condition (A, B, control) as a between-subjects variable. As expected, a strong effect of exemplar type resulted, $F(1, 45) = 317.8$, $P < 0.01$, $\mathrm{MSE} = 857.8$, indicating that participants in all three conditions correctly identified the viruses causing and not causing splenomegaly. There was no significant main effect of measurement, but the interaction between exemplar type and measurement was significant, $F(1,45) = 10.7$, $P < 0.01$, $\mathrm{MSE} = 123.6$. As the means displayed in Table 1 show, estimates became more extreme at the second measurement. Two further effects turned out to be significant. There was a main effect of condition, $F(2, 45) = 3.55$, $P < 0.05$, $\mathrm{MSE} = 272.0$, and an interaction between exemplar type and condition, $F(2, 45) = 3.31$, $P < 0.05$, $\mathrm{MSE} = 857.8$. Both effects can be traced back to the control condition, especially to the estimates for the exemplars not causing splenomegaly. Participants in the control condition gave less extreme

Table 1 Mean ratings (±SE) of the likelihood of the causal effect (splenomegaly) of the second causal relation for critical and uncritical items in Experiment 1 ($N = 48$)

	First measurement (after 42 trials)			Second measurement (after 84 trials)		
	Uncritical-high items	Critical items	Uncritical-low items	Uncritical-high items	Critical items	Uncritical-low items
Condition A	91.9 (3.38)	46.7 (3.15)	11.6 (3.35)	98.8 (0.97)	39.3 (4.4)	4.7 (2.16)
Control condition	81.6 (6.39)	50.6 (3.12)	24.7 (6.57)	89.4 (5.74)	48.8 (3.1)	24.4 (4.65)
Condition B	81.3 (6.05)	62.8 (5.43)	8.1 (5.18)	91.3 (3.37)	61.3 (3.1)	8.4 (4.93)

ratings to these viruses than in the other conditions. We speculate that participants in this condition may have focused on the positive set of viruses causing splenomegaly to identify the critical features and thereby may have neglected the negative set. By contrast, participants in the experimental conditions could use the two categories induced in Phase 1, thus simplifying the representation, which may have led to more capacity to process all exemplars. Another possible explanation of the found tendency might be that the learners' overall exposure time to the virus items was reduced in the control condition relative to the other conditions because of the missing first learning phase. Reduced exposure may have led to higher uncertainty which may have manifested itself in a tendency to gravitate toward the mean.

We next analyzed the critical items which, according to our hypothesis, should receive different ratings across conditions. First a repeated measurement ANOVA with the variables measurement (1 vs. 2) as a within-subjects variable and condition (A, B, control) as a between-subjects variable was computed for the critical items. The analysis yielded a strong effect of condition, $F(2, 45) = 9.89$, $P < 0.01$, MSE = 302.5. No other effects proved significant. This result provides initial evidence that participants' learning of the first causal relation affects their assessment of the second causal relation. Hence, the results suggest that participants induced categories of viruses in the first causal learning phase and used them to make inferences regarding the effect relation in the second learning phase.

To analyze these findings in more detail, we conducted a number of pair-wise planned comparisons. First, we computed contrasts of learners' estimates of causal efficacy for the first measurement (after 42 trials). Since we specified the predicted direction of the effect in advance, we generally conducted one-tailed contrast tests. Consistent with our hypothesis, a reliable difference was obtained between the two experimental conditions ($M_A = 46.7$ vs. $M_B = 62.8$, SE = 5.72), $t(57) = 2.82$, $P < 0.01$. This finding corroborates the overall conclusion and shows that participants' judgments differed depending on the category structure entailed by the first causal relation. We then tested whether the two experimental conditions also differed from the control condition with no Phase 1 training. The difference between Condition A and the control condition ($M_A = 46.7$ vs. $M_{control} = 50.6$, SE = 3.95) did not turn out to be significant [$t(57) = 0.69$, $P = 0.25$], but a reliable difference was obtained when contrasting Condition B with the control condition ($M_B = 62.8$ vs. $M_{control} = 50.6$, SE = 5.72), $t(57) = 2.12$, $P < 0.05$.

Next we conducted the same comparisons for the second measurement (after 84 trials). As can be seen from an inspection of Table 1, the difference between the three conditions has become even larger than at the first measurement. The two experimental conditions clearly differed ($M_A = 39.3$ vs. $M_B = 61.3$, SE = 5.08; $t(57) = 4.32$, $P < 0.001$), but reliable differences were also obtained when contrasting the two experimental conditions with the control condition. In line with our predictions, participants' estimates in Condition A were lower than in the control condition ($M_A = 39.3$ vs. $M_{control} = 48.8$, SE = 5.08; $t(57) = 1.86$, $P < 0.05$, while learners in Condition B gave higher estimates than in the control condition ($M_B = 61.3$ vs. $M_{control} = 48.8$, SE = 5.08), $t(57) = 2.46$, $P < 0.01$).

Finally, we analyzed the data which we collected to tap directly onto learners' category-level knowledge (Table 2). Participants were asked for the probability that viruses that had been produced by Alpha [Beta] radiation caused a swelling of the spleen. It is important to recall that participants were never asked to learn this relationship. Table 2 shows the results. An ANOVA with the variables measurement (1 vs. 2) and category (viruses that were created by Alpha vs. Beta radiation) as within-subjects variables, and condition (A vs. B) as a between-subjects variable was computed. Two effects turned out to be significant. First, there was a strong effect of category, $F(1, 27) = 117.4$, $P < 0.01$, MSE = 971.1, which indicates that participants not only induced cause-based categories in the first learning phase, but also encoded the relation between these categories and the effect in the second learning phase. The second significant effect was the interaction between category and measurement, $F(1, 27) = 4.44$, $P < 0.05$, MSE = 167.6. In accordance with the causal efficacy estimates, the difference between the estimated likelihoods for the two categories of viruses became slightly larger at the second measurement.

Table 2 Mean ratings (±SE) of the likelihood of splenomegaly for viruses created by Alpha radiation or Beta radiation in Experiment 1 ($N = 48$)

	First measurement (after 42 trials)		Second measurement (after 84 trials)	
	Category 1: viruses created by α-radiation	Category 2: viruses created by β-radiation	Category 1: viruses created by α-radiation	Category 2: viruses created by β-radiation
Condition A	78.7 (4.87)	23.3 (4.75)	86.3 (3.40)	15.6 (3.16)
Condition B	82.1 (5.66)	22.1 (5.26)	80.6 (6.68)	22.5 (5.66)

Overall the results provide strong evidence for the hypothesis that participants spontaneously induced cause-based categories in the first causal learning phase and used them in further learning about the second causal relation involving the same exemplars. Interestingly, this category-based transfer effect became even stronger in the course of learning. Moreover, the relation between the categories entailed by the first relation and the terminal effect in the chain was encoded on the category level. Hence, the learning of the second causal relation in the experimental conditions clearly did not fully optimize the predictability on the local level (as in the control condition), but paid tribute to the advantages of consistently using a single set of categories for the whole causal chain.

Experiment 2

Experiment 1 showed that people spontaneously induced cause-based categories of effect exemplars which, in turn, influenced subsequent causal learning in a causal chain. Causal chains, however, might be special with respect to category transfer. The initial cause in Experiment 1 *created* different new natural kinds. Because of the deep causal impact of the initial cause on the intermediate exemplars, learners may have had a strong tendency to use and transfer the newly induced categories. But what about categories based on effects rather than causes? An effect entailing a categorical scheme cannot causally influence another effect of the same exemplars. Nevertheless, we hypothesize that people may also transfer effect-based categories. Although effects cannot cause new natural kinds, they can be viewed as indicators of natural kinds (see Waldmann and Hagmayer 2006). In this case transfer should be observed as well. To investigate transfer with effect-based categories we used a conceptually different causal structure in Experiments 2a and b, a common-cause model (cf. Fig. 1). In this model a single cause, which is presented as a set of different exemplars, is linked to two binary effects, which are presented in consecutive learning phases. Again our goal was to investigate whether learners induce categories based on the effect observed first, and transfer them to a causal relation involving the second effect.

Experiment 2a

Figure 1 (second panel) depicts the common-cause model used in Experiment 2a. As in Experiment 1, we manipulated the boundaries of the category structure in the first causal learning phase while providing all participants with identical learning input in the second learning phase. In Experiment 2a we again used a family resemblance structure for the virus exemplars, but now these viruses played the role of a common cause. In the first learning phase participants learned about the causal relation between the cause exemplars, the viruses, and the first effect, bladder infection. Then learners proceeded to the second stage in which they were presented with data regarding the relation between the virus exemplars and the second effect, splenomegaly. Subsequent to the second learning phase, participants were requested to assess the causal efficacy of several cause events with respect to the second effect event. Again we were interested in exploring whether learners would transfer the categories entailed by the first causal relation when subsequently learning about a second causal relation. While in Experiment 1 the category scheme was induced on the basis of information about a binary *cause* ("cause-based categories"), in Experiment 2a the categories were based on binary *effect* information ("effect-based categories"). As in Experiment 1, no explicit category feedback was provided.

Method

Participants and design

Sixty students from the University of Göttingen, Germany, participated in this experiment for course credit. They were randomly assigned to one of three conditions. In the two experimental conditions (Conditions A and B) participants first learned about a cause-effect relation in Phase 1. The category structures entailed by the initially presented causal relation was manipulated between conditions. In Phase 2, participants were requested to learn about a second cause-effect relation. In this phase all participants were provided with identical learning input. In the control condition the first learning phase was omitted.

Procedure and materials

For Experiment 2a we used the same set of stimuli ("viruses") as in Experiment 1. The procedure and the instruction closely resembled those used in Experiment 1. The main differences between the experiments arose from the fact that in Experiment 2 participants had to learn a common cause structure and not a causal chain. In the experimental conditions, again two subsequent causal learning phases were administered. The assignment of the two effects to the two causal learning phases was counterbalanced; we here use one of the counterbalancing conditions to describe the procedure.

Prior to the first causal learning phase, participants were told that scientists had discovered new viruses and are now investigating whether these viruses can cause an infection of the bladder in animals. Like in Experiment 1 participants

were shown an overview of all 16 viruses. Unlike in Experiment 1, the viruses were now described as causes, not as effects. The overview contained information about which viruses did cause a bladder infection and which did not cause a bladder infection. After inspecting this overview, participants proceeded to the first learning phase. As in the previous study, the learning data regarding the viruses and bladder infection were shown on index cards one after another (in random order). First, a virus exemplar was shown on the front side of an index card. Then, learners had to predict whether the shown exemplar would cause the disease. After making a prediction, they received feedback on whether this virus exemplar had in fact caused an infection or not. Figure 3 shows an example of the family resemblance structure of the learning data and of two different causal categories entailed by the causal relation in Phase 1 of Conditions A and B. For each of these conditions we counterbalanced which items were linked to effects. Learning continued until participants made correct predictions throughout a learning block, allowing only for one error in 16 items. Then, participants proceeded to the second learning phase. In the control condition, no first learning phase was administered.

While Phase 1 was manipulated between conditions, Phase 2 was identical for all participants. We used the same instructions and learning data as in Experiment 1 (cf. Figs. 3, 4). Participants were told that veterinarians investigated whether the viruses caused a swelling of the spleen. Again it was pointed out that every result was possible: the viruses might always cause a splenomegaly, only some of them might cause a splenomegaly, or they might be completely unrelated to the splenomegaly. As in the previous study, the learning data were presented on index cards one after another. In this phase, learners did not make any overt predictions but only passively observed the cause-effect pairs. As before, the assignment of the disease to the viruses was counterbalanced.

After 42 trials (3 blocks of 14 viruses) the test phase began. We used the same test items and rating scales as in Experiment 1. For each test exemplar participants were asked to judge the likelihood of the virus causing splenomegaly. The final questions aimed at learners' knowledge on the category level. Participants were asked to rate the likelihood of viruses causing (or not causing) a bladder infection to cause a swelling of the spleen. We used the same rating scales as in Experiment 1.

Results and discussion

All participants met the learning criterion in the first causal learning phase and therefore proceeded to the second learning phase. The ratings concerning the second causal relation in the counterbalancing conditions were recoded to fit the example shown in Fig. 2. Following the analyses of Experiment 1, participants' causal judgments for the test items were averaged for the two uncritical-high items, the two uncritical-low items, and the six critical items. Table 3 shows participants' mean estimates.

We first analyzed participants' judgments for the uncritical items using an analysis of variance with exemplar type (uncritical-high vs. uncritical-low) as within-subjects variable and condition (A, B, control) as a between-subjects variable. Consistent with our predictions, a strong main effect of exemplar type was obtained, $F(1, 57) = 366.36$, $P < 0.001$, MSE = 397.4, which shows that participants clearly distinguished between viruses causing and not causing the effect. In line with our predictions no interaction was obtained ($F < 1$). A weak effect of condition was found, $F(1, 57) = 2.6$, $P = 0.08$, which had not been predicted, but can be traced back to participants' estimates for the uncritical-low items in Condition B, which were slightly below the two other conditions. However, the overall pattern clearly supports the hypothesis that learners' estimates for the uncritical items should be invariant across category conditions.

The next analyses concerned the critical items. If participants' estimates for these exemplars were influenced by the categories entailed by the first causal learning phase, judgments should systematically differ across conditions. The data shown in Table 3 indicate that the ratings conform to the predicted qualitative trend. For the very same viruses, estimates in Condition B ($M = 53.7$) were higher than in Condition A ($M = 44.5$), with the control condition ($M = 49.7$) being in between the two estimates. In this experiment, only one factor was manipulated, which we analyzed using planned directed t tests. We again

Table 3 Mean ratings (±SE) of the likelihood of the second causal effect for critical and uncritical items and mean ratings (±SE) of the likelihood of the effect for causes generating the first effect in Experiment 2a ($N = 60$)

	Uncritical-high items	Critical items	Uncritical-low items	Category 1: viruses causing bladder infection	Category 2: viruses not causing bladder infection
Condition A	89.8 (3.49)	44.5 (4.51)	17.0 (4.71)	66.0 (5.73)	31.0 (6.11)
Control condition	83.5 (3.86)	49.7 (2.44)	20.3 (3.71)	–	–
Condition B	83.0 (3.86)	53.7 (3.88)	10.0 (3.40)	59.0 (6.11)	39.5 (5.87)

conducted one-tailed contrast tests, which show a significant difference between the experimental conditions [$t(57) = 1.74$, $P < 0.05$; SE = 5.24]. The two experimental conditions did not significantly differ from the control condition. However, the qualitative pattern of judgments was clearly in line with the predictions.

Finally, we analyzed learners' category estimates of the overall probability that viruses that caused the effects in Phase 1 (bladder infection) led to the effect learned in Phase 2 (splenomegaly). We used an analogous question concerning viruses that did not cause the effects in Phase 1. The mean estimates shown in Table 3 indicate that participants were indeed very sensitive to the relation between the effect-based categories entailed by the first causal relation and the second effect. An ANOVA with the categories entailed by the first effect as within-subjects variable and condition (A vs. B) as between-subjects variable yielded a strong effect of category, $F(1, 38) = 11.67$, $P = 0.001$, MSE = 1,273.6. No other effect was significant. This finding provides strong evidence that learners encoded the relation between the effect-based categories of the first learning phase and the second causal effect.

Taken together, these findings corroborate the results of Experiment 1 and generalize the findings to effect-based categories and a common-cause model linking two causal relations. Similar to the previous study, learners' causal judgments regarding the second causal relation differed systematically in accordance with the categories entailed by the first causal learning phase. Apparently the effect from the first causal relation was interpreted as an indicator of a stable natural kind category (virus) that is also relevant for predicting further collateral symptoms. This knowledge not only affected individual predictions but was also accessible on the category level.

Experiment 2b

Experiment 2b was designed as a follow-up of the previous study to replicate the findings with a different category structure and a different procedure. While we used a prototype category structure in Experiments 1 and 2a, we used orthogonal category boundaries in this experiment (Goldstone 1994; Waldmann and Hagmayer 2006). The exemplars had four features, two relevant and two irrelevant ones, with four levels each, which resulted in an exemplar space of 256 items. The large number of items allowed us to present two non-overlapping sets of cause exemplars in the two causal learning phases. A second novel feature of Experiment 2b is that we aligned the two causal learning phases to ensure equal exposure to both causal relations. In the previous studies, the first learning phase continued until participants had reached a learning criterion (e.g., 15 out of 16 correct predictions in a row), while the second phase comprised a fixed number of trials (which was usually smaller than the first phase). In the present study, both causal learning phases are equally long. In addition, while previously learners had to make explicit predictions in Phase 1 (but not in Phase 2), this time the learning data was passively observed in both phases.

Method

Participants and design

Forty-eight students from the University of Göttingen, Germany, participated for course credit. They were randomly assigned to one of two conditions, which differed with respect to the causal category structure entailed by the first causal relation (category structures A vs. B).

Procedure and materials

The experimental rationale followed the previous studies. We again administered two causal learning phases, with the first phase being manipulated between conditions. The second learning phase was identical for all participants. In the test phase the learners were requested to assess the causal efficacy of a number of cause exemplars to examine the influence of the first causal learning phase on the second one.

This time we used an orthogonal category structure (see Fig. 5). Either the size of the viruses or their brightness was the relevant feature in the first causal learning phase. Each of the four features (size, brightness, number of corners, and number of molecules) had four levels. The diameter of the viruses varied in four steps between 30 and 48 mm, and brightness was manipulated by using four equally spaced levels of grayness. The number of corners varied between five and eight, and the number of molecules between two and five. Factorially combing all four levels of all features yielded 256 different exemplars. Figure 5 shows examples of the 16 types of viruses that can be created by combining the four values of the relevant dimensions size and brightness.

Depending on condition either size or brightness was the relevant feature in the first causal learning phase. As before, the learning data were presented on index cards one after another. Each learning phase comprised a subset of 64 exemplars with each exemplar being shown only once. We ensured that no other feature than the relevant one was correlated with the categories and also made sure that the exemplars' features were not intercorrelated. At the end of Phase 1, we asked participants to speculate which feature was relevant for predicting the effect bladder infection. No feedback was given about whether the guess was correct.

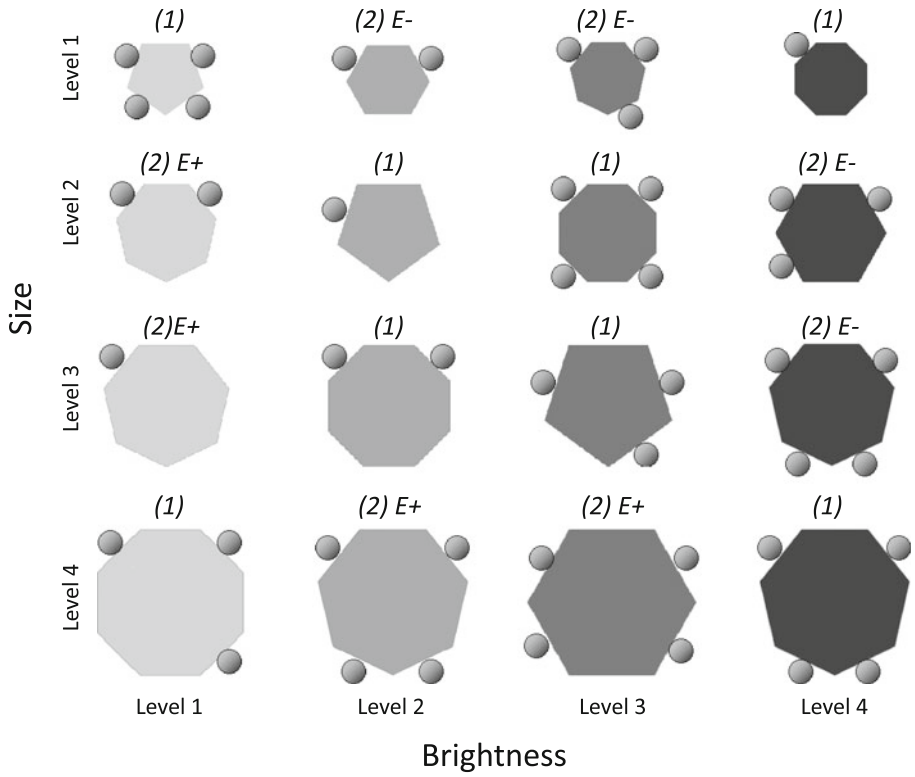

Fig. 5 Example of stimuli used in Experiment 2b. Exemplars denoted with a (*1*) were shown in the first causal learning phase (e.g., viruses → bladder infection). Depending on condition either brightness or size was the relevant feature. Items denoted with (*2*) were shown in the second causal learning phase (e.g., viruses → splenomegaly). *E*+ and *E*− denote the presence and absence, respectively, of the second causal effect presented in Phase 2. See text for details

However, only participants who correctly identified the relevant dimension proceeded to the next phase. In total, 22 participants failed to reach the criterion. To ensure sufficient sample size, we ran additional subjects to replace those who did not reach the criterion.

The second learning phase was identical for all participants. As in the previous studies, learners received data regarding the relation between the viruses and splenomegaly. This second learning phase also comprised 64 exemplars, which were different from the exemplars in the first phase (see Fig. 5). The learning input was again presented on index cards which participants observed passively.

In the test phase we switched to test exemplars similar but not identical to the ones presented in Phase 1 (i.e., they varied on the irrelevant dimensions). Participants were presented with eight test items one after another. For each item they had to assess the likelihood that the shown virus would cause splenomegaly (the effect of the second learning phase). We used the same rating scales as in the previous experiments.

Figure 5 depicts the statistical structure of the two causal learning phases. Exemplars labeled with (1) denote a sample of items from the first causal learning phase, viruses marked with a (2) denote exemplars used in the second learning phase. As noted above, in the test phase we switched back to items similar to the ones used in the first learning phase [i.e., items marked with a (1)], but these items varied on the irrelevant dimensions. In the initial causal learning phase (Phase 1), a single dimension of the stimuli (size or brightness) was deterministically related to the effect. For example, in the size condition small viruses (size levels 1 and 2) did not cause an infection, whereas large viruses (size levels 3 and 4) did cause a bladder infection. Conversely, in the brightness condition all light viruses (brightness levels 1 and 2) caused the bladder infection deterministically, while dark viruses (brightness levels 3 and 4) did never lead to an infection. None of the other features was predictive for the causal effect. In Phase 2, which was identical for all participants, a combination of both size and brightness was predictive for the effect (e.g., splenomegaly). Thus, the causal categories entailed by the first causal relation are probabilistically predictive for the second effect, although not perfectly. For example, three out of four large viruses (size levels 3 and 4) caused splenomegaly, whereas only one out of four small viruses (size levels 1 and 2) caused the effect. Similarly, the effect is present in 75% of the light viruses but only in 25% of the dark viruses. A combination of the individual feature levels, by contrast, allows for maximal predictability. All large and light viruses cause splenomegaly, whereas the small and dark ones never generate the effect. Viruses with other combinations of these features (i.e., small and light; large and dark) had a 50% chance of causing the effect. Thus, to achieve maximal predictability in Phase 2, learners would have to induce a two-dimensional category boundary.

If people spontaneously induced effect-based categories in the first learning phase and transferred them to the second stage, a specific pattern of inferences should result. Similar to the previous studies, estimates for a number of cause exemplars should strongly differ between conditions (i.e., critical items) while other exemplars (uncritical items) should yield similar predictions regardless of condition. Consider again Fig. 5. The critical items are the small and light viruses and the large and dark viruses, both of which have an overall chance of 50% to cause the effect. If learners based their causal judgments only on the data presented in the second learning phase, both types of viruses should receive similar ratings. By contrast, if learners used the categories entailed by the first phase, then different judgments should arise. According to a brightness-based categorization, the small and light viruses should receive relatively high ratings, since overall the light viruses have a 75% chance of causing splenomegaly. Conversely, the large and dark viruses should receive low ratings, since large viruses generate the disease in only one of four cases. The exact opposite pattern is entailed when the causal categories of the first causal learning phase are based on the size of the viruses. Then, the small and light viruses should receive relatively low ratings, whereas the large and dark exemplars should elicit high ratings.

For the uncritical items no difference is predicted. Regardless of whether size or brightness had been predictive for the first causal effect, the cause items that are simultaneously large and light should be considered highly effective, since they always belong to the class that is predictive for the effect, regardless of condition. By contrast, the small and dark cause exemplars should always yield low ratings, since they in both conditions fall into the category that is only weakly associated with the second causal effect.

The assignment of effects (bladder infection and splenomegaly) to the learning phases and the assignment of exemplars to the two causal learning phases were counterbalanced. In addition, we also balanced which items were critical or uncritical by rotating the effect structure (Fig. 5) clockwise.

Results and discussion

Table 4 shows the results of Experiment 2b. Participants' estimates for the critical and uncritical items are listed separately depending on the predictions entailed by the respective condition. For example, the column labeled $\text{brightness}_{\text{low}}/\text{size}_{\text{high}}$ contains the mean rating for exemplars that, presuming that learners sticked with the initially acquired causal categories, should receive low judgments of causal efficacy in the brightness condition but high ratings in the size condition. A first inspection of the data for the critical items shows that the obtained response pattern indeed reveals a strong influence of the experimental manipulation on learners' judgments of causal efficacy for the critical items.

We conducted separate analyses for the uncritical and critical items. We first analyzed learners' estimates for the uncritical items. An ANOVA with exemplar type ($\text{uncritical}_{\text{high}}$ vs. $\text{uncritical}_{\text{low}}$) as within-subjects variables and condition (size vs. brightness) as between-subjects variable yielded a strong effect of exemplar type, $F(1, 46) = 63.28$, $P < 0.001$, $\text{MSE} = 867.1$, and a weak, non-significant effect of condition $F(1, 46) = 2.56$, $P = 0.11$, $\text{MSE} = 353.9$. As expected, no significant interaction was obtained.

Next, we conducted the same analysis for the critical items. An ANOVA with exemplar type ($\text{brightness}_{\text{low}}/\text{size}_{\text{high}}$ vs. $\text{brightness}_{\text{high}}/\text{size}_{\text{low}}$) as within-subjects variable and condition (size vs. brightness) as between-subjects variable yielded the predicted interaction effect, $F(1, 46) = 6.7$, $P = 0.01$, $\text{MSE} = 1{,}347.5$. Consistent with the expected disordinal interaction is the lack of a main effect for the exemplar variable ($F < 1$). However, unexpectedly, the between-subjects variable condition turned out to be significant, $F(1, 46) = 5.90$, $P < 0.05$, $\text{MSE} = 406.5$, which may be due to differences in the features' salience.

In summary, these results corroborate the findings of the previous studies. Using a different category structure, equally long learning phases and no learning feedback (i.e., unsupervised learning), we obtained further evidence for category use in a common-cause model. Initially induced causal categories were transferred to learning of a further causal relation involving the same category exemplars.

General discussion

The main goal of the present set of studies was to demonstrate the tight coupling between category and causal

Table 4 Mean ratings (±SE) of the likelihood of the second effect in Experiment 2b ($N = 48$)

Condition	Uncritical items		Critical items	
	$\text{Brightness}_{\text{high}}/\text{Size}_{\text{high}}$	$\text{Brightness}_{\text{low}}/\text{Size}_{\text{low}}$	$\text{Brightness}_{\text{high}}/\text{Size}_{\text{low}}$	$\text{Brightness}_{\text{low}}/\text{Size}_{\text{high}}$
Brightness	67.1 (5.01)	21.9 (5.22)	67.7 (5.78)	44.8 (6.22)
Size	75.8 (4.62)	25.4 (5.29)	38.1 (6.99)	54.4 (5.24)

induction. Lien and Cheng (2000) demonstrated that learners induce categories that maximize predictability when cause exemplars are linked to a single binary effect (see also Marsh and Ahn 2009). While this research focused on effect-based categories, we studied both effect- and cause-based categories. Effect-based categories were studied in the context of common-cause models, whereas we investigated cause-based categories in the context of a causal chain. In both models causal information was used to establish optimally predictive categories in the initial causal learning phase.

The central goal of our studies was to investigate the stability of the induced categories across causal learning contexts. While previous research has investigated categories in the context of single links (Lien and Cheng 2000; Marsh and Ahn 2009), our more complex models allowed us to investigate a potential conflict between local predictability and global coherence. Unlike in Waldmann and Hagmayer's (2006) experiments, which investigated this question with a different paradigm, no explicit category label feedback was provided and participants were never requested to induce categories.

We used a two-phase learning procedure in which learners were presented with learning input referring to two components of a causal chain (Experiment 1) or a common-cause model (Experiments 2a, b) with the overlapping event being presented as a set of uncategorized exemplars. Phase 1 was modeled after Lien and Cheng (2000) and supported their conclusion that learners induce categories based on causal input in the absence of explicit category labels. The novel question investigated in this paper concerns the subsequent learning phase, in which the exemplars were presented in the context of a further causal relation. Now participants had the choice between transferring the causal categories from Phase 1 at the cost of suboptimal predictions in Phase 2, or they could induce a new set of optimally predictive categories within Phase 2, but at the cost of a proliferation of different category schemes for the same set of exemplars. In the three presented experiments learners clearly chose the first option. They preferred to stick to the categories induced in the initial causal learning phase and transferred them to the subsequent causal learning phase. In order to predict a second causal effect of an exemplar, participants obviously considered whether the exemplar belonged to a category of exemplars that was either generated by a specific cause variable (Experiment 1) or was generated by a further previously learned effect (Experiments 2a, b). Moreover, learners were capable of directly accessing knowledge about the relation between the more abstract categories and the effect in Phase 2, although they never were asked to induce or use categories. This finding lends further support to the hypothesis that participants used the categories entailed by the first learning phase while learning about the second causal relation.

Directions for future research

An open question we did not examine in the present set of studies concerns the way the newly induced categories are represented. For example, are people forming an abstract category representation, or do they represent categories in terms of exemplars? Both Lien and Cheng (2000) and Marsh and Ahn (2009) have provided evidence for category formation in causal contexts. For example, Marsh and Ahn have shown that effect information alters the subjective similarity structure of the cause exemplars. Our category-level ratings also show that people do not only store exemplar knowledge but also learn on the category level. However, exemplar information is certainly not lost. Waldmann and Hagmayer's (2006) data suggest that learners gather knowledge about both categories and exemplars.

The most important finding of our studies is that learners generally preferred to stick to the initially learned categories over inducing new ones optimized for each causal relation. A plausible explanation for this preference, in line with the results of Waldmann and Hagmayer (2006), is that with categories referring to natural kinds, such as viruses, people believe in a common essentialist core, which must not arbitrarily be altered based on new information. This assumption is consistent with *psychological essentialism* (Medin and Ortony 1989), which is claimed to underlie the naïve representation of natural kinds from childhood on (Gelman 2003). A related view ties psychological essentialism to causal-model theory. A number of researchers have argued that natural kinds may be represented as common-cause models or chains with the essentialist features playing the role of a hidden common cause or initial event (Gelman 2003; Rehder 2003a, b; Rehder and Hastie 2004; but see Strevens 2000 for a slightly different view).

How does this causal approach explain our findings? Possibly, people have a strong intuition that viruses, an example of a natural kind, are defined by hidden causal features which are causally linked to the two effects in the common-cause model or to the initial cause and final effect in the causal chain. Thus, they may represent the relations between the three events as being connected by a continuous, unbroken causal mechanism, which would be disrupted if the events were re-categorized. This hypothesis could be tested by manipulating learners' beliefs about the underlying causal mechanism. Only if a common mechanism links the causal effects, transfer of categories should be observed, whereas in situations in which different aspects of the exemplars are involved in different relations, learners might opt for inducing new, more predictable

categories in Phase 2. Initial evidence consistent with this theory comes from Waldmann and Hagmayer (2006). They have shown that people abandon previously learned categories about viruses, when in the second learning phase a causal relation was presented that did not refer to the causal power of viruses but to their superficial appearance. In their Experiment 3, Waldmann and Hagmayer required participants to learn about two types of viruses in Phase 1, but in Phase 2 these viruses were introduced as candidates for esthetic patterns which could be used in interior design. The task in the second phase was to predict whether fictitious subjects in a study testing the attractiveness of the patterns would like the appearance of the virus or not. Thus, for the new causal relation the hidden causal power of viruses to cause diseases was irrelevant. As a consequence, learners neglected the categories from Phase 1 and induced new ones in Phase 2. It would be interesting to explore whether similar effects can also be shown with chains or common-cause models and a procedure in which no category label feedback was provided. Such a study would provide boundary conditions for the effect discovered in the present set of studies that learners tend to induce categories based on causal feedback and transfer them to new learning episodes.

Acknowledgments The research was funded by the DFG (Wa621/17-1, 2).

References

De Houwer J, Beckers T (2002) A review of recent developments in research and theories on human contingency learning. Q J Exp Psychol B 55:289–310

Gelman SA (2003) The essential child: origins of essentialism in everyday thought. Oxford University Press, New York

Goldstone RL (1994) Influences of categorization on perceptual discrimination. J Exp Psychol Gen 123:178–200

Hagmayer Y, Sloman SA (2009) Decision makers conceive of their choices as interventions. J Exp Psychol Gen 138:22–38

Kemp C, Goodman ND, Tenenbaum JB (2007) Learning causal schemata. In: McNamara DS, Trafton JG (eds) Proceedings of the 29th Annual Cognitive Science Society. Cognitive Science Society, Austin, pp 389–394

Lagnado DA, Waldmann MR, Hagmayer Y, Sloman SA (2007) Beyond covariation: cues to causal structure. In: Gopnik A, Schulz L (eds) Causal learning: psychology, philosophy, and computation. Oxford University Press, Oxford, pp 154–172

Lien Y, Cheng PW (2000) Distinguishing genuine from spurious causes: a coherence hypothesis. Cogn Psychol 40:87–137

Marsh JK, Ahn W (2009) Spontaneous assimilation of continuous values and temporal information in causal induction. J Exp Psychol Learn Mem Cogn 35:334–352

Meder B, Hagmayer Y (2009) Causal induction enables adaptive decision making. In: Proceedings of the 31st annual conference of the Cognitive Science Society

Medin DL, Ortony A (1989) Psychological essentialism. In: Vosniadou S, Ortony A (eds) Similarity and analogical reasoning. Cambridge University Press, Cambridge, pp 179–195

Murphy GL (2002) The big book of concepts. MIT Press, Cambridge

Murphy GL, Medin DL (1985) The role of theories in conceptual coherence. Psychol Rev 92:289–316

Rehder B (2003a) A causal-model theory of conceptual representation and categorization. J Exp Psychol Learn Mem Cogn 29:1141–1159

Rehder B (2003b) Categorization as causal reasoning. Cogn Sci 27:709–748

Rehder B, Hastie R (2001) Causal knowledge and categories: the effects of causal beliefs on categorization, induction, and similarity. J Exp Psychol Gen 130:323–360

Rehder B, Hastie R (2004) Category coherence and category-based property induction. Cognition 91:113–153

Shanks DR, Holyoak KJ, Medin DL (1996) The psychology of learning and motivation. In: Causal learning, vol 34. Academic Press, San Diego

Strevens M (2000) The essentialist aspect of naïve theories. Cognition 74:149–175

von Sydow M, Meder B, Hagmayer Y (2009) A transitivity heuristic of probabilistic causal reasoning. In: Proceedings of the 31st annual conference of the Cognitive Science Society

Waldmann MR (1996) Knowledge-based causal induction. In: Shanks DR, Holyoak KJ, Medin DL (eds) The psychology of learning and motivation. Causal learning, vol 34. Academic Press, San Diego, pp 47–88

Waldmann MR (2000) Competition among causes but not effects in predictive and diagnostic learning. J Exp Psychol Learn Mem Cogn 26:53–76

Waldmann MR (2001) Predictive versus diagnostic learning: evidence from an overshadowing paradigm. Psychon Bull Rev 8:600–608

Waldmann MR, Hagmayer Y (2006) Categories and causality: the neglected direction. Cogn Psychol 53:27–58

Waldmann MR, Holyoak KJ, Fratianne A (1995) Causal models and the acquisition of category structure. J Exp Psychol Gen 124:181–206

Waldmann MR, Hagmayer Y, Blaisdell AP (2006) Beyond the information given: causal models in learning and reasoning. Curr Dir Psychol Sci 15:307–311

Waldmann MR, Cheng PW, Hagmayer Y, Blaisdell AP (2008) Causal learning in rats and humans: a minimal rational model. In: Chater N, Oaksford M (eds) The probabilistic mind. Prospects for Bayesian cognitive science. University Press, Oxford, pp 453–484