Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cognitive Psychology 53 (2006) 27–58
Cognitive Psychology
www.elsevier.com/locate/cogpsych

Categories and causality: The neglected direction ☆

Michael R. Waldmann *, York Hagmayer

Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany

Accepted 3 January 2006
Available online 23 February 2006

Abstract

The standard approach guiding research on the relationship between categories and causality views categories as reflecting causal relations in the world. We provide evidence that the opposite direction also holds: categories that have been acquired in previous learning contexts may influence subsequent causal learning. In three experiments we show that identical causal learning input yields different attributions of causal capacity depending on the pre-existing categories to which the learning exemplars are assigned. There is a strong tendency to continue to use old conceptual schemes rather than switch to new ones even when the old categories are not optimal for predicting the new effect, and when they were motivated by goals that differed from the present context of causal discovery. However, we also found that the use of prior categories is dependent on the match between categories and causal effect. Whenever the category labels suggest natural kinds which can be plausibly related to the causal effects, transfer was observed. When the categories were arbitrary, or could not be plausibly related to the causal effect learners abandoned the categories, and used different categories to predict the causal effect.

Keywords: Casual learning; Casual reasoning; Categorization; Natural kinds

☆ Portions of this research were conducted when the authors were affiliated with the University of Tübingen (M.W. and Y.H.) and the Max Planck Institute for Psychological Research, Munich (M.W.). We thank B. Meder and M. v. Sydow for many helpful comments and their help with Experiments 2 and 4. Experiment 1 was presented at the 1999 meetings of the Cognitive Science Society, Vancouver, of the European Society of Cognitive Psychology, Ghent, and at the Experimental Psychology (TEAP) conference, Leipzig.
* Corresponding author. Fax: +49 551 393656.
E-mail addresses: michael.waldmann@bio.uni-goettingen.de (M.R. Waldmann), york.hagmayer@bio.uni-goettingen. de (Y. Hagmayer).

0010-0285/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2006.01.001

1. Introduction

Traditionally research about the representation of causal relations and research about the representation of categories were separated. This research strategy rests on the assumption that categories summarize objects or events on the basis of their similarity structure, whereas causality refers to relations between causal objects or events. Our goal in the present research is to show that the relationship between causality and categorization is more dynamic than previously thought.

1.1. The standard view: Causality rests on fixed categories

The standard view guiding research on causality presupposes the existence of objective networks of causes and effects, which cognitive systems try to mirror. Regardless of whether causal learning is viewed as the attempt to induce causality on the basis of statistical information or on the basis of mechanism information, it is generally assumed that the goal of causal learning is to form adequate representations of the texture of the causal world.

Studies on causal learning typically investigate trial-by-trial learning tasks which involve learning the contingencies between causes and effects. For example, Waldmann (2000) gave participants the task to learn about the strength of causal relations between different substances (e.g., substance 1) in fictitious patients' blood and a new blood disease, Midosis, which is caused by these substances. This task is a representative example of a large number of studies which focus on causal contingency learning (see De Houwer & Beckers, 2002; Shanks, Holyoak, & Medin, 1996, for overviews). A characteristic feature of these tasks is that they present categorized events representing causes (e.g., "substance 1") and effects (e.g., "Midosis") which are statistically related. Cause and effect categories are viewed as fixed entities that are already present prior to the learning task. The goal of learning is to estimate causal strength of individual causal links or to induce causal models on the basis of observed covariations. The role of cause and effect categories in the learning process is not the focus of interest in these approaches; they are simply viewed as given.

A similar approach underlies research on the relationship between categories and causality. According to the view that categorization is theory-based, traditional similarity-based accounts of categorization are deficient because they ignore the fact that many categories are grounded in knowledge about causal structures (Murphy & Medin, 1985; see also Murphy, 2002). In natural concepts features often represent causes or effects with the category label referring to a complex causal model. For example, disease categories frequently refer to common-cause models of diseases with the category features representing causes (e.g., virus) and effects (e.g., symptoms) within this causal model. A number of 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 generalization (Rehder, 2003a, 2003b; Rehder & Hastie, 2001, 2004; Waldmann, Holyoak, & Fratianne, 1995; Waldmann, 1996, 2000, 2001). The main goal of these studies was to investigate the effect of different causal *relations* connecting the causal features. As in contingency learning studies, the cause and effect features within the causal models were treated as fixed, categorized entities, which already existed prior to the learning context.

1.2. The neglected direction: Categories shape causality

It is certainly true that many interesting insights can be gained from investigating how people learn about causal models on the basis of pre-existing cause and effect categories. However, there is also a link between categories and causality in the opposite direction: The categories that have been acquired in previous learning contexts may have a crucial influence on subsequent causal learning. This direction has typically been neglected in research on the relationship between categories and causality.

The basis of the potential influence of categories on causal induction lies in the fact that the acquisition and use of causal knowledge is based on categorized events. Regardless of whether causal relations are viewed as statistical relations (probabilistic causality view) or as mechanisms (mechanism view), both accounts postulate causal regularities that refer to *types* of events. Causal laws, such as the fact that smoking causes heart disease, can only be noticed on the basis of events that are categorized (e.g., events of smoking and cases of heart disease). Without such categories causal laws neither could be detected nor could causal knowledge be applied to new cases. Thus, causal knowledge not only affects the creation of categories, it also presupposes already existing categories for the description of causes and effects.

Given that the induction of new causal knowledge is based on already existing categories, the question arises whether the outcome of causal learning may be influenced by the categories that are being used. The potential influence of categories is due to the fact that one of the most important cues to causality is statistical covariation between causes and effects. Many (otherwise conflicting) views agree that causal induction is based on the observation of causes altering the probability of effects (e.g., contingency view; associationist theories)(see Shanks et al., 1996).

However, statistical regularities are not invariant across different categorical segmentations of domains. This can easily be shown with a simple example. Let us assume, for example, a world with four different (uncategorized) event tokens, A, B, C, and D, that represent potential causes. It has been observed that A and C are followed by a specific effect, but B and D are not. Now the statistical regularities that are observed in this mini-world are crucially dependent on how these four events are categorized. If A and B are exemplars of Category 1, and C and D exemplars of Category 2, no causal regularity would be observed. Within this conceptual framework the effect has a base rate of 0.5, which is invariant across the two categories. By contrast, categorizing A and C (Category 3), and B and D (Category 4) together would lead to the induction of a deterministic causal law. Events that belong to Category 3 always produce the effect, whereas Category 4 is never associated with the effect. Thus, the causal regularities observed in a domain are dependent on the way the domain is categorized. In fact, as pointed out by Clark and Thornton (1997) in an example with (non-causal) continuous features, there is an infinite number of descriptions of the world with a potentially infinite number of statistical regularities entailed by these descriptions.

A number of philosophers have raised similar arguments against the traditional view of "metaphysical realism" which assumes that there is a ready-made world of objects and processes that exist independent of the concepts we are using to represent them. Many philosophers have argued that there are alternative conceptual schemes that can be used to describe reality and that truth is a joint function of reality and the conceptual scheme being used to describe states of affairs in the world (see Dupré, 1993; Goodman, 1978; Hacking, 2000; Nozick, 2001; Putnam, 1987). This view implies that the causal relations we see in the world will depend on the categorical schemes we use to describe causes and effects.

Philosophers and historians of science provided evidence that alternative categorizations are not only a theoretical possibility, but also a practical reality. Kuhn's (1962) concept of scientific paradigms can be reconstructed as denoting categorical schemes (see Hacking, 1993; Thagard, 1999). The concept of paradigm and the analyses of theory change in science have also been used to explain knowledge development in childhood (see Carey, 1985; Gopnik & Meltzoff, 1997).

1.3. Alternative ways of categorizing the world: Evidence from psychology

At this point it could be argued that the dependence of causal knowledge on pre-existing categories is a philosophical rather than a psychological problem as long as it has not been shown that there is evidence for multiple categorizations of the same domains outside of the scientific realm. Following the work of Eleanor Rosch, many psychologists have assumed that natural categories are relatively stable since they are reflecting the correlational structure of features in the world (see Rosch, 1978). Thus, even if multiple categorizations might be theoretically possible, and could be artificially generated, the existence of alternative categorizations of the same domains might not seem very plausible to these researchers. However, in the past years psychologists have started to question the assumption that categories merely reflect objective correlations. Theoretical analyses and empirical research began to draw attention to pragmatic factors that affect the particular choice of a categorical framework. We only have space to give some pointers to this research.

A number of different research areas have focused on the fact that exemplars may be *cross-classified* into different systems of categories depending on the goals of the categorizer. In social psychology, research on stereotypes has investigated alternative ways of classifying people. The same person may be categorized as male, a professor, a tennis player or a vegetarian. Depending on the chosen category, different associations are consciously and unconsciously evoked (see Kunda, 1999, for an overview). Ross and Murphy (1999) have studied cross-classifications in the context of food items. Food items can be classified according to taxonomic categories (e.g., drinks) or event-related categories (e.g., breakfast items). Ross and Murphy showed that we use both types of category systems in parallel, and activate them depending on the goals we currently pursue. Studies who contrasted expertise or different cultures have also provided evidence for alternative categorizations of domains (see Atran, 1998; Medin & Atran, 2004; Medin, Lynch, Coley, & Atran, 1997).

This research demonstrates that alternative categorizations of identical domains is not just a possibility but can be found in science as well as everyday cognition. However, very little is known about the effect of alternative category systems on how further causal knowledge is induced based on contingency information.

1.4. How categories shape causality: Alternative theoretical hypotheses

Section 1.2 presented arguments for the hypothesis that our causal knowledge may be dependent on the categories we use to describe a domain. Although a number of studies have provided empirical evidence for the reality of alternative categorizations of domains, the possible influence on later contingency-based causal induction has not been studied. Given that domains may be categorized differently depending on goals, expertise,

theoretical framework, or cultural background, the impact of categorization on causal induction warrants investigation.

To study the relation between categories and causal induction we have developed a new paradigm that consists of three phases. In all experiments we use fictitious viruses as learning exemplars. Prior to Phase 1, the *category learning phase*, we instructed participants, for example, that the virus exemplars had been, at some stage in our fictitious history of their discovery, classified by scientists on the basis of morphological features into different distinct and exhaustive categories (e.g., allovedic vs. hemovedic viruses). Then participants learned to classify the virus exemplars into the two categories in a trial-by-trial learning procedure. Prior to the second learning phase, the *causal learning phase*, we told participants that later other scientists had studied whether virus exemplars from these categories cause specific disease-related symptoms, for example a swelling of the spleen (splenomegaly). In this learning phase, participants passively observed whether individual virus exemplars were paired with the effect or not. In the third *test phase* we presented learners with another set of virus exemplars and asked them how likely it is in their opinion that the particular virus generates the effect. In both the causal learning and the test phase the categories from Phase 1 were not mentioned. Participants only received information on the exemplar level.

The main goal of the present research is to answer the question under what condition learners will tend to activate the categorical information from the earlier category learning phase when learning about causal contingencies in the later phase. To measure the influence of the prior categories, we manipulated the rules underlying the categories taught in Phase 1. For example, in one condition the viruses might be classified on the basis of their brightness (regardless of size) whereas in the contrasting orthogonal condition (varied between subjects) they might be classified on the basis of size (regardless of their brightness). On the assumption that learners tend to classify the test exemplars into the categories learned in Phase 1, different probability estimates are expected in our learning domains. This effect is entailed by the fact that the alternative categories form different reference classes. If, for example, light items tended to generate the effect but large items do not, then a light and large test item would be viewed as causally effective if it was classified by the learners as a member of the light class (as predicted in the brightness condition) but not if it was classified as a member of the large class (as predicted in the size condition).

Fig. 1 gives an outline of the learning options participants have. The left lower angle represents the set of learning exemplars (e.g., images of individual viruses), the upper angle the possible categories (e.g., allovedic vs. hemovedic viruses), and the right angle the causal effect from the second learning phase (e.g., splenomegaly). In Phase 1 of our learning paradigm, participants learn to classify the exemplars into the virus categories (upper left arrow). In Phase 2, participants are presented with exemplars again, which are paired with

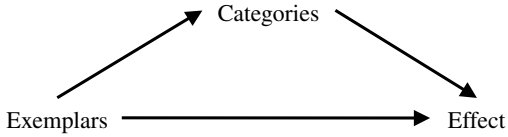

Fig. 1. Possible routes of learning in our paradigm (see text for explanations).

the presence or absence of an effect (splenomegaly)(lower arrow). In the subsequent test phase participants are shown other test exemplars and asked to rate the likelihood of the effect (splenomegaly). The crucial question is whether participants in Phases 2 and 3 would go through the upper route when answering this question and assign the test exemplars to the categories from Phase 1 or whether they would stick to the lower route and induce new categories within Phase 2. Because we manipulated the categories on the upper route across conditions, the responses to the test questions should reveal whether these categories were used.

What categories would underlie the responses in the transfer phase if learners opted for the lower route? One possible strategy may be to induce new categories that are maximally predictive of the effects. For example, all viruses that cause the target effect may be lumped together in one category, and the remaining exemplars may be grouped in the contrastive category. This strategy would obviously generate maximally predictive categories. Most likely these categories would not be labeled by participants, but they may still drive the predictive inferences in the test phase.

Lien and Cheng (2000) reported research consistent with this hypothesis. In their experiments, Lien and Cheng presented exemplars to learners, which could be classified by different features at different hierarchical levels of abstraction. Participants saw pictures of substances that varied in color and shape along with information about which of these substances make flowers bloom and which not. The results showed that learners categorized the substances according to the feature and to the hierarchical level that were maximally predictive for the effect. Thus, the induced substance category 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.

In sum, our research addresses the question which route learners will go. Will they routinely go through the upper road and activate category knowledge when learning about novel effects, or will they go the lower road and learn a new set of categories in Phase 2 that is maximally predictive within this learning phase?

1.4.1. The perceptual learning hypothesis

Before we outline our theoretical predictions which are derived from the view that categories are based on intuitive theories, it might be useful to look at what bottom-up, similarity-based theories might predict for our paradigm. We will use these predictions as a potential alternative account for our experimental results.

It is easy to see that standard similarity-based theories would not predict a transfer between the learning phases. Fig. 2 shows a simple connectionist one-layer network that may be used to understand the task we are going to explore in our experiments (see Gluck & Bower, 1988, for an example of these kinds of category learning models). The input layer represents a number of features and the output layer the outcomes that have to be predicted. One of the outcome nodes may represent the categories that have been learned in Phase 1 of a learning study. In Phase 2 learners are confronted with the same features but a second outcome, for example a causal effect of the exemplars. It can readily be seen that no transfer between the two tasks is predicted by this model. The associative weights between the features and the two outcome nodes are learned independently of each other. Similarly, most prototype or exemplar-based models (see Murphy, 2002) would not predict transfer. Thus, this type of models predicts that the two phases operate completely independent of each other.

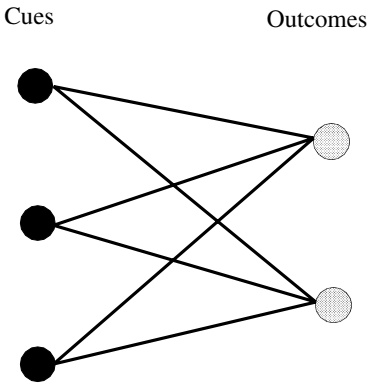

Fig. 2. Example of a one-layer connectionist network in which three cues (features) are linked to two outcomes. Learning to predict either outcome is independent of learning to predict the other outcome in these models.

However, there are alternative theories that would predict transfer. Goldstone, Steyvers, Spencer-Smith, and Kersten (2000) report a number of experiments that show that initial extensive training of novel categories might have numerous effects on the perception of stimuli (see also Goldstone & Steyvers, 2001; Schyns, Goldstone, & Thibaut, 1998). The general structure of the experimental paradigm in this area is very similar to the one we are using. For example, Goldstone (1994) trained participants to classify squares either according to a rule that was based on their size or their brightness. After extensive training, transfer tasks revealed that the category training affected how participants perceived the items. Sensitization to the relevant dimensions and desensitization to irrelevant dimensions were observed. Kruschke's (1992) ALCOVE model also predicts sensitizations and desensitizations to relevant dimensions as a result of learning.

Thus far, the impact of category learning on subsequent *causal contingency* learning has not been investigated within this paradigm. A plausible hypothesis might be that prior category learning affects the encoding of the stimuli whose similarity structure might change. This effect would be a function of factors affecting bottom-up learning, such as the featural structure of the learning items, the similarity structure of the categories, and properties of the learning procedure (e.g., type of feedback, extensiveness of training, etc.).

1.4.2. The dynamic theory modification hypothesis

We have argued that probability estimates necessarily require an assignment of the test exemplars to a reference category. According to the dynamic theory modification hypothesis, the choice of a reference class is a more flexible process than envisaged by the perceptual learning hypothesis, and is guided by both bottom-up and top-down factors. Learners in our paradigm generally have different options: They can continue to use categories from a previous stage (Phase 1), they can abandon these categories and use other (for example similarity-based) categories that they already bring into the learning session, or they can create new categories on the basis of causal information (e.g., maximally predictive categories; Lien & Cheng, 2000). Assuming that the learning exemplars are novel and that the set of exemplars does not already contain salient perceptual category boundaries (as in our tasks), participants primarily have a choice between inducing

new categories within the causal learning phase (Phase 2) or sticking to the categories learned in Phase 1 (see Fig. 1).

We hypothesize that the decision between old and new categories is driven by intuitive theories about the domain (top-down) and by a tendency to be parsimonious (bottom-up). As for the bottom-up component, we believe that in general people are reluctant to induce several competing category systems in parallel even though parsimony may come at the cost of sub-optimal predictability. If maximizing predictability was the main goal of category learning (see Anderson, 1991; Lien & Cheng, 2000) people should tend to induce a new category system for each target feature or target causal effect they are trying to predict (these category systems may overlap, of course). Old categories are in most cases not as predictive as new ones that could be induced from scratch. However, we believe that people try to minimize the number of alternative categorical schemes whenever possible. As long as the old categories allow us to make sufficiently satisfying predictions, people should have a tendency to continue to use them. Thus, we expect to see a general tendency to use old categorical schemes whenever they have at least some predictive value.

The main focus of the present research is the top-down component, which is the key feature of our dynamic theory modification hypothesis. Our general hypothesis is that learners view the task of learning new causal effects of categorized exemplars as a task of *dynamic theory modification*. Category labels are not just features among other features such as size or brightness, they often provide pointers to underlying causal structures (see Yamauchi & Markman, 2000, for evidence for the special status of category labels; see also Gelman, 2003). Thus, we expect people to view the categories learned in a previous context as skeletal causal theories (e.g., of novel viruses). These theories are viewed as incomplete so that new knowledge acquired later may be added or may be used to modify the previous theories.

Whether or not learners attempt to modify the theories implied by the categories will depend on whether they view the causal models underlying the categories as plausible generators of the novel causal effects. If the new causal hypothesis targets a potential effect that appears like a possible, yet unexplored effect of the old categories, then there should be a tendency to continue to use these categories. For example, viruses seem to be perfect candidate causes for diseases even when the original classification was based only on their morphology. Thus, there should be a tendency to use virus categories when learning about a novel disease-related effect, such as splenomegaly. In contrast, when the categories and the target effect seem hard to interrelate, then people may decide to abandon the old categories and induce new ones that are better suited for the current context of discovery.

This hypothesis is consistent with recent research on different kinds of categories (see Medin, Lynch, & Solomon, 2000). Viruses and diseases, for example, belong to the class of *natural kind* concepts. Medin and Ortony (1989) have argued that for this kind of categories we have a tendency to assume a hidden essence underlying the visible, variable features (psychological essentialism). More recently, it has been proposed that essences play the role of an invisible common cause placeholder which is responsible for the visible features (see Ahn et al., 2001; Gelman, 2003; Hirschfeld, 1996; Medin & Atran, 2004; Rehder & Hastie, 2004; but see Strevens, 2000). The existence of a hidden cause does not only explain the correlation among the visible features, it also makes it plausible to expect further, yet unknown effects (e.g., splenomegaly in the case of viruses).

Indirect empirical support for this hypothesis comes from a number of studies using paradigms different from ours. In a seminal study, Gelman and Markman (1986) have

shown that young children would project a novel feature ("feeds its baby mashed up food") on the basis of a common category label rather than similarity. Interestingly, Gelman and Coley (1990) found that the children did not make the inductive inference for any property but preferred properties that were plausibly and stably connected to the categories (see Gelman, 2003, for a summary of related research). Heit and Rubinstein (1994) showed that category-based inductive inferences depend on the kind of shared feature. For example, a biological property (e.g., a liver with two chambers) was viewed as more likely to be shared by whales and bears than by whales and fish. The opposite was observed with behavioral features (e.g., travel in a zig-zag path). Apparently people have general, abstract assumptions about the kinds of features different categories might generate. A possible explanation for these patterns is that different kinds of categories share different common causes whose probable causal effects are dissimilar. Lassaline (1996) has supported this hypothesis by showing that undergraduates were more likely to project a new property when the categories share a common cause of the property (see also Rehder & Hastie, 2004; Sloman, 1994).

Although in some domains people have expertise about the causal relations underlying causal categories and their effects (see Proffitt, Coley, & Medin, 2000), more often people have only skeletal, incomplete knowledge (see Rozenblit & Keil, 2002). Thus, people may often only have a very general framework theory (Wellman & Gelman, 1992) about what goes with what. In our experiments we explicitly say in the introduction to the category learning phase that the categories are solely based on perceptual commonalities. Although this instruction may discourage the assumption of natural kinds, the category labels (e.g., allovedic viruses) might trigger a tendency to assume natural kind categories anyway and to accept further causal effects that are globally consistent with domain assumptions about the categories (e.g., viruses cause symptoms). However, learners should tend to not use prior categories when there is a mismatch between the semantics of the category label and the effect (e.g., viruses causing aesthetic judgments).

Whether category labels associated with natural kinds trigger generalized expectations about further potential effects can also be tested by running control conditions in which the category labels are introduced as arbitrary. For example, a set of virus exemplars might be sorted into two arbitrary piles A and B. Although in this condition the exemplars individually refer to natural kinds, the superordinate categories (piles A and B) do not suggest a common hidden causal structure that might be systematically related to novel, yet unexplored effects. Thus, in this condition it is expected that people would tend to induce new categories rather than recruiting the arbitrary categories from Phase 1.

In summary, the dynamic theory modification hypothesis states that people will view our task as involving theory modification. Categories are not represented as simple unstructured holistic entities; they rather have an internal, in part unknown causal structure that may affect further causal learning processes. If people assume hidden, unknown common causes of natural kinds, they should be willing to use the category information from Phase 1, when Phase 2 involves a causal effect that seems, at least remotely, relatable to the category. In this case causal learning is a process that modifies and augments knowledge about categories that already have an internal causal structure. This process of theory modification is analogous to working within a given paradigm or framework theory in science (see also Carey, 1985; Wellman & Gelman, 1992). In contrast, whenever the category seems arbitrary, when the causal link between the category and the causal effect is implausible, or when the assumed underlying causal structure of the category is inconsistent with

the discovery of the novel causal effect, we expect people to prefer to abandon the categories learned in Phase 1, and resort to inducing alternative categories in Phase 2.

Although both bottom-up and top-down factors affect the learning process according to our theory, the following experiments will mainly focus on the top-down aspect of our hypotheses. The hypotheses that address the bottom-up influences will be explored in future research (see also Section 5).

2. Experiment 1

The goal of the first experiment is to empirically demonstrate the potential impact of different ways of categorizing domains on subsequent causal induction. It will show that, despite identical causal learning input, causal inferences can dramatically differ depending on the way people categorize a domain. Thus, the main goal is to demonstrate the interdependence between the way causes are categorized and further causal learning involving these categories. This paradigm will in later experiments be modified to test between the alternative theories outlined in Section 1.

To demonstrate the influence of categories on causal induction, we used a three-phase paradigm: In Phase 1, the *category-learning* phase, participants learned to categorize a novel domain. They were told that scientists had discovered new types of viruses, which they had classified on the basis of their *appearance* into two categories, *allovedic* and *hemovedic* viruses. In the learning task, participants saw pictures of viruses and learned to classify them into the two categories. Fig. 3 depicts the exemplars shown in this and the other two phases.

While Phase 1 differs between conditions, the subsequent Phases 2 and 3 were *identical* across conditions. In Phase 2, the *causal-learning* phase, participants were told that later physicians became interested in exploring the relationship between the newly discovered viruses and diseases in animals. In particular, they wanted to find out whether the viruses

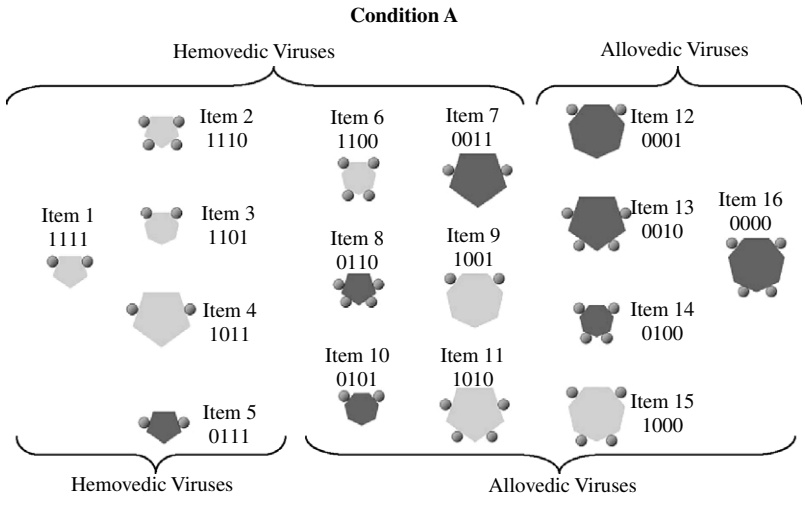

Fig. 3. Learning items and category boundaries in Experiments 1 and 2 (see text for details).

cause splenomegaly, which is a swelling of the spleen. Therefore, the scientists studied animals that were infected with the new viruses. We pointed out that any outcome of this study was possible, including the possibility that there was no causal relationship between the viruses and splenomegaly. After the instructions, participants were shown a new set of virus exemplars one after another. First, they were presented the picture of a virus, and then information was given about whether this particular virus causes splenomegaly. The categories "allovedic" or "hemovedic" were not mentioned within this phase. In this phase, participants observed the potential cause and the effect but unlike in Phase 1 they were not requested to make predictions. In the real world we also simply observe cause-effect relations, and accept strong or weak contingencies as facts. Asking for predictions and providing corrective feedback would have suggested to participants that the task was to find a perfect rule as in Phase 1.

Phase 2 was followed by Phase 3, the test phase. In this phase, exemplars were shown again along with the request to estimate the likelihood that the particular exemplar causes the effect. Again the categories from Phase 1 were not mentioned. Since only Phase 1 varies, this design allows us to test whether the ratings in the test phase are influenced only by the causal learning phase, which provided identical information about the statistical relation between cause and effect, or whether participants recruit category knowledge from Phase 1 that should differentially affect their ratings.

In the contrasting categorization conditions (varied between subjects) we used linearly separable, family resemblance categories that were based on four binary features (see Figs. 3 and 4). None of these features was individually sufficient for achieving correct classifications. However, correct classifications could be learned by an additive integration of the four features. Since all features were equally relevant, transfer effects cannot be accounted for by theories that postulate a carry-over of selective attention or sensitizations to individual features.

We manipulated the size and variance of the contrasting categories. Fig. 3 shows the categorical boundaries for the two conditions A and B. In one condition (Condition A), one type of viruses (e.g., hemovedic) was characterized by having a light color, small size,

Items	Features				Effect	Test	Categorization	
	Light.	Size	Corners	Mol.			A	B
1	1	1	1	1	E	T	Hemovedic viruses	Hemovedic viruses
2	1	1	1	0	E			
3	1	1	0	1	E	T		
4	1	0	1	1	E			
5	0	1	1	1	E			
6	1	1	0	0	E	T		Allovedic viruses
7	0	0	1	1	E	T		
8	0	1	1	0	-	T		
9	1	0	0	1	-	T		
10	1	0	1	0	~E	T		
11	0	1	0	1	~E	T		
12	0	0	0	1	~E		Allovedic viruses	
13	0	0	1	0	~E			
14	0	1	0	0	~E	T		
15	1	0	0	0	~E			
16	0	0	0	0	~E	T		

Fig. 4. Structure of learning exemplars, categories (category learning phase), effects (causal learning phase), and test exemplars in Experiments 1 and 2.

few corners, and few surface molecules. All viruses that had at least two out of these four features belonged to the category, with all the remaining exemplars belonging to the contrast category (e.g., allovedic). In the other condition (Condition B), the hemovedic viruses had at least three out of the four features; again the remaining exemplars belonged to the contrast category. Thus, we compared two categorization schemes that shared the same prototypes (1 1 1 1 or 0 0 0 0) but whose variances differed.

Fig. 4 shows how the causal effect (i.e., splenomegaly) was distributed. Half of the exemplars generated the effect; the others did not. In the test phase (Phase 3), we presented 10 of the 16 items again (see Fig. 4), and had participants rate the likelihood of the effect. The most important predictions involve the six items (6–11) lying between the category boundaries of the two conditions. In Condition A, these items should be viewed as being members of the hemovedic virus type. Because within this group seven out of nine viruses caused splenomegaly, high ratings are to be expected. By contrast, the very same items should yield low ratings in Condition B. In this condition the six items belong to the allovedic viruses, which cause the effect in only two out of nine cases.

Alternatively, participants could opt for neglecting the categories from Phase 1, and induce new categories. In this case, the most plausible categories are the two groups of exemplars closely similar to the two prototypes which either cause or do not cause the effect, and a category in between whose items are equally similar to both prototypes in terms of feature overlap (i.e., Items 6–11). These exemplars in the middle zone have a 50% chance of causing the effect. Therefore, the ratings for these items should be on average identical across the two conditions and hover around the 50% value. The resulting categories would be maximally predictive (Lien & Cheng, 2000).

2.1. Method

2.1.1. Participants and design

Thirty-two students from the University of Tübingen were randomly assigned to one of the two learning conditions (category boundary A vs. B). In all four experiments, only participants were included in the statistical analyses who met the learning criterion. Participants who did not learn the categories (fewer than 5%) were replaced to preserve the counterbalancing schemes. This was decided because it does not make sense to investigate the potential use of initial categories if these categories are not being learned in Phase 1. We adopted this policy in all three experiments.

2.1.2. Procedure and material

The exemplars (fictitious viruses) varied in *four binary* dimensions: brightness (20 vs. 60%), size (diameter of 30 vs. 42 mm), number of corners (5 vs. 7), and number of molecules on the surface (2 vs. 4). Fig. 3 shows the items and Fig. 4 displays the structures of the learning items with the feature value 1 representing low values and the value 0 high values.

The experiment consisted of three phases: In Phase 1, the *category-learning phase*, participants were told that scientists had discovered new types of viruses that vary in the dimensions brightness, size, shape, and number of molecules on the surface. Cytophysiological investigations had revealed two types of viruses, which can be distinguished on the basis of their appearance, *allovedic* and *hemovedic* viruses. After the instructions we requested a summary of the participants, which should include the four dimensions of the materials that were mentioned. Then participants were shown index cards with pictures of

viruses one after another, and they had to judge whether the respective exemplar represented a hemovedic or an allovedic virus. After each judgment corrective feedback was given.

Two categorization conditions were compared that manipulated the location of the category boundaries (see Fig. 3). In *Condition A*, hemovedic viruses had at least two low values on the four dimensions (Items 1–11), whereas allovedic viruses (Items 12–16) had only one or no low value. By contrast, in *Condition B* hemovedic viruses (Items 1–5) had three or more low values, whereas allovedic viruses (Items 6–16) had at least two high values. We used a learning criterion in Phase 1. Learning proceeded until participants managed to correctly classify one block of 16 items. A maximum of 8 blocks was administered. The items were presented in random order within blocks. The category labels were counterbalanced.

Whereas Phase 1 differed across the two conditions, the subsequent causal-learning phase and the test phase were *identical* for all participants. In Phase 2, the *causal-learning phase*, participants were told that physicians were interested in exploring the relationship between the newly discovered viruses and diseases in animals. In particular, they wanted to find out whether the viruses cause splenomegaly. Therefore, they studied animals that were infected with the new viruses. It was pointed out that any outcome of this study was possible including the possibility that there was no causal relationship between the viruses and splenomegaly. Participants received index cards that depicted exemplars of the viruses with information on the backside on whether the respective virus causes splenomegaly (E) or not ($\sim E$). Each side was presented for roughly three seconds without further corrective feedback. To avoid an unequal association of individual features with the effect, Items 8 and 9 were not presented in this phase. Eliminating these two items ensured that no individual feature was correlated with the effect. In the particular counterbalancing condition shown in Fig. 4, Items 1–7 caused the effect, whereas Items 10 through 16 did not cause it. In a second counterbalancing condition effects and non-effects were exchanged. The causal-learning phase consisted of three blocks of the 14 cases, which were presented in a random order.

In Phase 3, the test phase, participants received 10 exemplars (1, 3, 6–11, 14, 16). Learners' task was to express their assessment of the likelihood that the respective virus causes splenomegaly by using a rating scale that ranged from 0 ("never") to 100 ("always"). After these ratings, participants also gave a general assessment of the likelihood that the two virus types, allovedic and hemovedic viruses, caused the effect. These ratings allowed us to check whether participants had encoded the causal relation on the category level.

2.2. Results and discussion

Table 1 shows the results. The most important analysis involves the test items between the two category boundaries (Items 6–11). The mean ratings for these six items clearly differed across the two category boundary conditions A and B, $F(1, 30) = 14.7$, $p < .01$,

Table 1
Mean ratings of the likelihood of the causal effect for critical and uncritical items (Experiment 1)

Categories	Uncritical high items 1, 3	Critical items 6–11	Uncritical low items 14, 16
A	90.3	65.9	25.0
B	86.2	45.6	21.1

$MSE = 224.3$. By contrast, the items that should be rated high (e.g., 1, 3) or low (e.g., 14, 16) regardless of learning condition did not differ significantly in the contrasting conditions, all Fs < 1.

The final ratings showed that participants generally encoded the relationship between categories and the effect. They rated the causal efficacy of the two categories clearly differently regardless of the location of the category boundary, $F(1, 31) = 80.5$, $p < .01$, $MSE = 509.5$ ($M = 74.8$ vs. $M = 24.2$). All but five participants gave ratings consistent with this trend.

Given that we used family resemblance structures we also looked for potential influences of typicality. Rehder and Hastie (2004) have shown that less typical exemplars yield weaker inductive generalizations than exemplars that are more typical for the category. We compared the causal ratings (test phase) for the two prototypes (Items 1 and 16) with the ratings for test items that deviated in only one feature from the prototypes (Items 3 and 14). We focused on these items because in Phase 2 they had the same probability of causing splenomegaly (100% or 0%). The ratings for Items 14 and 16 were recoded to match the ratings for Items 1 and 3 by subtracting the ratings from 100. An analysis of variance with the factor typicality (prototype vs. less typical item) as a within-subjects factor and category boundary condition (A vs. B) as a between-subjects factor yielded a significant effect of typicality, $F(1, 30) = 13.6$, $p < .01$, $MSE = 131.1$, all other Fs < 1. This result shows that participants were indeed sensitive to the differences in typicality. Less typical exemplars received less extreme ratings than the prototypes. However, it is important to note that typicality differences cannot account for our findings as the critical test items from the middle zone shared the same number of features (2) with either of the two prototypes. Thus, the typicality of the critical items was the same with respect to both categories. To account for differences in the ratings of the critical items, it is necessary to assume that learners encoded the category boundaries that implied different category variances.

In sum, Experiment 1 demonstrates that causal induction is influenced by the way the cause exemplars are categorized. Despite the fact that all participants received identical cause-effect information in the causal-learning phase, the ratings of the causal efficacy of the exemplars were moderated by the categories to which they belonged.

Although the experiment was not designed to test between the perceptual learning and the dynamic theory modification hypotheses, it rules out simple bottom-up learning models that might postulate transfer of selective attention or sensitizations as the basis of our transfer effects. The family resemblance categories contain features that are all equally relevant. The results show that learners were apparently sensitive to both the similarity structure of the categories and their variability. This finding is consistent with previous research (e.g., Flannagan, Fried, & Holyoak, 1986; Fried & Holyoak, 1984) for categorization tasks, but it additionally demonstrates how this kind of category knowledge affects novel learning about causal relations involving the categories.

3. Experiment 2

Experiment 2 goes one step further and provides a first test between the competing theories. In Section 1, we hypothesized that learners may categorize the viruses on the basis of superficial features such as brightness, size, number of corners, and molecules, but that the driving force behind transfer is learners' assumption that they are actually learning something about real viruses (i.e., natural kinds), which have specific causal

powers, including the power to cause disease-related symptoms. In contrast, similarity-based theories (including the perceptual learning hypothesis) view categories as economic ways to represent collections of features with category labels as arbitrary additional features (Anderson, 1991). Even when category labels are given a special status which sets them apart from mere features (Yamauchi & Markman, 2000), the general role as category labels and not their semantic contents are viewed as crucial for the distinction between category label and feature. Thus, according to the perceptual learning hypothesis it should make no difference how novel categories are labeled. Whether an exemplar is labeled as referring to allovedic viruses or to an arbitrary category A, should not affect transfer as long as the learning procedure and the learning items are otherwise identical.

In contrast, according to the dynamic theory modification hypothesis the semantic content of the label is crucial. Labeling the superordinate category "virus" should lead to a category representation that implies that the visible features are caused by a hidden causal structure that is common to the members of the categories, and that this hidden causal structure has the potential of causing further semantically related effects. In contrast, arbitrary categories which are described with labels that are not associated with specific domain assumptions (e.g., A, B) should not lead to such natural kind representations. There is no reason to assume that an arbitrary collection of viruses shares a common hidden causal structure. Consequently, it should be less likely that learners expect a novel effect to be associated with the arbitrary category.

To test these predictions, Experiment 2 adds a control condition to the design of Experiment 1. Again we are using the family-resemblance categories from the previous experiment. In the regular natural kind category-learning condition one group of participants learned about the two types of viruses. As in the first two experiments, these two types were labeled hemovedic and allovedic. The previous experiment has shown that people tend to continue to use these categories when learning about a disease-related symptom in the second causal-learning phase.

In the present experiment we added an *arbitrary category-learning* condition, which was run on a different sample of participants. In this second condition, participants were also told that they were going to see viruses. The four dimensions were mentioned as well. But then participants were told that the first phase serves the purpose of familiarizing them with the different viruses. To accomplish this, they would learn to categorize them into two classes. It was pointed out that the categories were based on an arbitrary rule. Since there are many possible rules, we mentioned six rules, and asked participants to roll a die to select the rule that would be used. This part was meant to emphasize that the categories were indeed arbitrary. In fact, all participants learned on the basis of the same family-resemblance rules displayed in Figs. 3 and 4. Otherwise, learning was identical in the two conditions. Participants in both conditions received learning exemplars in a trial-by-trial learning procedure with corrective feedback in the category-learning phase and proceeded until they reached a learning criterion of one correct block of 16 exemplars. The following causal-learning and test phases were again identical to the ones of Experiment 1.

Since in the arbitrary category-learning condition, participants acquired the same categories as in the regular category-learning condition, all theories that use a bottom-up learning mechanism (i.e., all similarity-based theories including the perceptual learning

hypothesis) would predict identical transfer. In fact, except for the initial instructions and the category labels, which in both conditions were novel and unfamiliar (A, B or allovedic, hemovedic) the two contrasting conditions were identical.

3.1. Method

3.1.1. Participants and design

Forty-eight students from the University of Göttingen were randomly assigned to one of the four conditions spanned by the factor type of category (natural kind vs. arbitrary category), and the factor category boundary (A vs. B) (see Fig. 4).

3.1.2. Material and procedure

We used the same materials and the same category structures as in Experiment 1. Again we manipulated the location of the category boundary across conditions (see Fig. 4). Moreover, the natural kind category-learning condition was an identical replication of Experiment 1 with largely the same instructions, learning input, feedback, and tests as in this experiment. The only difference was that in the category-learning phase (Phase 1) participants were first shown all 16 exemplars and the two categories to which they belonged. Participants were asked to study the two groups of exemplars before learning began. We added this part to simplify the learning of the categories. Afterwards participants went through a trial-by-trial learning phase like in the previous experiment. They were asked to assign each virus either to a pile labeled 'hemovedic viruses' or to a pile labeled 'allovedic viruses'. As before, they received corrective feedback on each trial.

The most important extension in the present experiment is the addition of an *arbitrary category-learning* condition. As in the natural kind category-learning condition, participants were told that they were going to learn about viruses. Again the four critical features were pointed out. In contrast to the category-learning condition, however, participants were then told that in the first phase of the experiment they were going to learn to sort the viruses into two piles in order to familiarize them with the learning material. It was pointed out that many rules are possible, and that one out of six would be randomly selected. To emphasize the arbitrary character of the two resulting categories, participants had to roll a die to determine the rule for the allocation of the viruses. Regardless of the outcome of the die, the two category structures displayed in Fig. 4 were used. As in the natural kind category-learning condition participants were first shown to which categories the individual exemplars belonged, and then went through a trial-by-trial learning phase with corrective feedback in which they classified individual exemplars one after another. The only difference to the natural kind category learning condition was that participants classified the individual exemplars as belonging to pile A or B instead of belonging to the piles labeled hemovedic or allovedic. The learning criterion in the arbitrary category learning condition was the same as in the contrast condition (see also Experiment 1).

Phases 2 and 3 were identical in both the arbitrary and the natural kind category-learning conditions. We used the same procedure as in Experiment 1. Thus, again, participants' task was to learn which viruses caused splenomegaly, and then to rate the likelihood that the test exemplars produced this symptom. Finally, they rated the likelihood that the contrasting categories cause the effect (as in Experiment 1).

Table 2
Mean ratings of the likelihood of the causal effect for critical and uncritical items (Experiment 2)

Categories	Natural kind category-learning condition			Arbitrary category-learning condition		
	Uncritical high items 1, 3	Critical items 6–11	Uncritical low items 14, 16	Uncritical high items 1, 3	Critical items 6–11	Uncritical low items 14, 16
A	81.7	55.3	15.4	88.3	42.1	12.9
B	75.8	40.8	23.3	81.3	45.7	25.8

3.2. Results and discussion

Table 2 shows the results for critical and uncritical items in all conditions. The most interesting result concerns the critical items from the middle area of Figs. 3 and 4. As Table 2 shows, the categories learned in Phase 1 affected the causal ratings for these exemplars only in the category-learning condition, but not in the arbitrary category-learning condition. A 2 (natural kind vs. arbitrary category) × 2 (category boundary A vs. B) analysis of variance with the mean ratings of the 10 test exemplars as dependent variable yielded the predicted significant interaction, $F(1, 44) = 4.61$, $p < .05$, $MSE = 211.9$. To pinpoint the pattern underlying the interaction, we analyzed the natural kind category-learning and arbitrary category-learning conditions separately. As can be seen in Table 2, the critical items were rated differently depending on the learned category boundaries in the natural kind category-learning condition, $F(1, 22) = 7.73$, $p < .05$, $MSE = 161.9$, but not in the arbitrary category-learning condition ($F < 1$). Moreover, the items that should be either rated high or low in both conditions did not differ significantly across the two conditions (all $Fs < 1.02$).

Again we checked for the potential influence of the typicality of the exemplars on the causal ratings by comparing the ratings of the two prototypes with the exemplars that deviated in only one feature from the prototype while having the same probability of causing splenomegaly. As in the first experiment we recoded the ratings to make the analyzed items comparable. An analysis of variance with the factor typicality (prototype vs. less typical exemplar) as a within-subjects factor, and the factors category boundary (A vs. B) and category type (natural kind vs. arbitrary category) as between-subjects factors was conducted. The results yielded a significant effect of typicality, $F(1, 44) = 4.46$, $p < .05$, $MSE = 262.4.4$. All other factors failed to reach significance. Thus, the finding of Experiment 1 was replicated. There was an influence of typicality upon participants' estimates. Nevertheless, the crucial difference between the critical items cannot be traced back to typicality because the typicality of the critical items with respect to either of the two prototypes was the same in all conditions.

We also analyzed the ratings of the relationship between the categories (i.e., viruses or piles) and the causal effect. The final ratings showed that participants generally encoded the relationship between categories and effect. The ratings of causal efficacy of the two categories or piles clearly differed regardless of the location of the category boundary, $F(1, 47) = 37.7$, $p < .01$, $MSE = 1196.5$ ($M = 74.3$ vs. $M = 31.0$). Eight participants gave ratings contrary to this trend, six of them in the natural kind category-learning condition. Interestingly, there were no differences between the natural kind category-learning and the arbitrary category-learning conditions, the mean differences were even slightly higher in the arbitrary category-learning condition. Thus, all participants, including the ones from the arbitrary category-learning condition, knew about the relationship between categories and

causal effect but participants in the arbitrary category-learning condition actively ignored this information in the test phase. This finding further weakens the possible alternative theory of a perceptual learning account that learners might have focused less on the arbitrary category labels than the natural kind labels. Learners in all conditions clearly learned the categories (all participants passed the learning criterion) and encoded the statistical relation between categories and effect. Thus, their decision to ignore the arbitrary categories must have been driven by their knowledge about causal relevance, a factor that transcends bottom-up perceptual learning accounts.

Finally, we looked at the individual items in the arbitrary category-learning condition in which the categories were ignored. The results indicate that participants divided the viruses into three groups; one which always causes the disease, one which never causes it, and one which falls in between, and sometimes causes the disease (see Table 1). The items that were equally similar to the two prototypes formed this third group. This hypothesis is supported by the fact that all items in this group received ratings distinctively different from the estimates for all other exemplars. Thus, there seemed to be a tendency to induce fuzzy categories on the basis of their association with the effect with items near the boundaries assigned a more probabilistic intermediate status. This strategy is in line with the hypothesis that learners who neglected the initial categories induced a new, although fuzzy category boundary in the middle region to distinguish between items that cause the effect from the ones that do not. These categories do not have labels but nevertheless distinctly affected the likelihood ratings in the test phase.

In sum, Experiment 2 clearly supports the dynamic theory modification hypothesis. Whenever participants believed that the virus categories were real, they used category-level knowledge for their predictions in the test phase, whenever they thought the categories were arbitrary they suppressed this knowledge. This finding weakens the perceptual learning hypothesis because apart from the use of different instructions and labels (which were novel and unfamiliar in both conditions) the learning procedure was identical.

Three other aspects of our findings deserve to be emphasized. In both Experiments 1 and 2 people activated category-level knowledge although the instructions clearly stated that the virus categories were based on superficial, morphological similarities. Nevertheless, the label "virus" along with the instruction that the virus categories were real and had been discovered by scientists seemed to promote a strong bias that the members of the exemplar share a common causal structure. University students certainly do not believe that size or brightness are potential causes of splenomegaly, it seems more likely that they viewed the perceptual features as indicators of an invisible causal power, which is consistent with a causal reinterpretation of psychological essentialism (e.g., Gelman, 2003). This essentialist bias seems to be strong even when instructions are given that should discourage this bias.

This interpretation is supported by the fact that learners in the arbitrary category-learning condition neglected the initially learned categories although the exemplar labels (viruses) were clearly semantically related to the causal effect in Phase 2. They apparently only assumed a common underlying causal structure when the categories were created on the basis of a scientific decision and not when the categories were arbitrary.

A third interesting finding is that in the arbitrary category-learning condition learners encoded the relation between the category and the causal effect. Thus, category-level information (the upper route in Fig. 1) seems to be routinely encoded although in the test phase learners apparently may opt to neglect it. This indicates an active role of top-down factors

that may override category knowledge that was acquired in the learning phase. This finding along with the fact that all participants learned the categories also weakens the possible route of the perceptual learning approach to postulate greater attentional weights for natural kind category labels.

4. Experiment 3

Thus far we have shown that participants tend to continue to use prior categories when the category labels suggest that they refer to natural kind categories. Such categories typically are represented as having deeper causal communalities, which might lead to the discovery of novel, yet unknown further causal effects. People make this inference on the basis of the category label even when they were explicitly told that the categories were motivated by superficial commonalities. Supporting this hypothesis, the last experiment has shown that arbitrary categories (e.g., two piles supposedly generated by a random rule) are neglected as categories even when the exemplar labels (viruses) are semantically related to the effect.

Experiment 3 goes one step further in testing our theory. According to our hypothesis, participants should ignore previously acquired categories even when they refer to natural kind categories when the causal effect is not semantically related to the category. Categories should only be used to predict a novel effect if a causal link between the effect and the hidden causal structure of the category appears at least remotely plausible within the framework theory of the learning domain. Consequently, our goal in this experiment was to test whether participants suppress previously acquired natural kind categories when these categories are unrelated to the target effect.

The present experiment adopts the paradigm and learning procedure from the previous experiments. In one condition participants learned again about hemovedic and allovedic viruses in Phase 1, and later learned about the relation between the virus exemplars and splenomegaly. In the test phase, participants judged the likelihood of splenomegaly for the test set of viruses. In the new condition of Experiment 3, a different sample of participants also first learned to categorize the viruses as hemovedic and allovedic. The crucial difference concerned the target effect in the causal-learning phase (Phase 2). We now told participants that designers had thought that pictures of the viruses can be used to generate attractive patterns. To test their hunch, the designers had conducted an empirical study investigating whether the viruses resulted in liking judgments or not. The causal-learning phase was similar to the one in the standard condition except that participants now observed which of the viruses observers from this study liked and which they did not like. Phase 3 was adapted to this condition: Now participants had to rate the probability with which new viruses resulted in liking judgments. Our prediction was that hidden causal structures underlying virus categories are normally not considered to be causally related to appearances of these viruses that are differentially liked by observers. Thus, we expected learners to be reluctant to use the categories from Phase 1 when estimating the probability of liking judgments in Phase 3.

In this experiment, we switched to a different category structure with orthogonal category boundaries. This structure was inspired by the categories Goldstone (1994) had used. Fig. 5 shows the structure of the categories and their relationship to the causal effects. Half of the exemplars were shown in the *category-learning phase* (indicated by A). Two conditions were compared: In the *size* condition participants learned, for example, that the

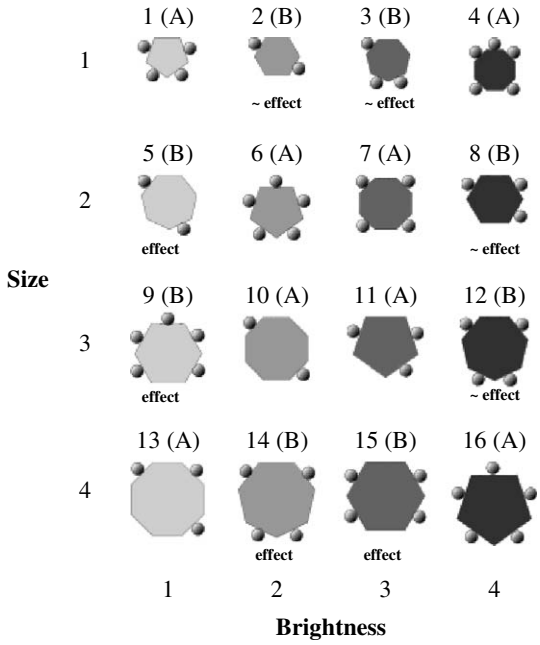

Fig. 5. Structure of learning exemplars, categories (category learning phase), effects (causal learning phase), and test exemplars in Experiment 3 (see text for explanations).

bigger viruses were allovedic, and the smaller ones hemovedic, in the orthogonal *brightness* condition they learned, for example, that the darker exemplars were allovedic and the lighter ones hemovedic. Fig. 5 displays the structure of the items with respect to the two relevant dimensions size and brightness. In the brightness condition the left half of the figure (Levels 1 and 2 of the brightness dimension) may represent allovedic viruses and the right half hemovedic viruses (Levels 3 and 4). By contrast, in the size condition the upper half represented one category (Levels 1 and 2 of the size dimension), and the lower half the other category (Levels 3 and 4).

While Phase 1 differed between conditions, the categories implied by the effects in the subsequent *causal learning phase* were *identical* across conditions. In this phase, participants were shown a new set of exemplars along with information about the presence or absence of splenomegaly or liking, depending on the condition. Items presented in this phase are indicated by B in Fig. 5. In the figure it can also be seen which exemplars caused splenomegaly or liking (*effect*), and which did not (∼*effect*).

In Phase 3, the *test phase*, we switched back to exemplars corresponding to the A-items from the category-learning phase, but with different values of the irrelevant features. Thus, these items had never been presented before. Participants' task was to express their assessment of the likelihood that the respective virus causes splenomegaly or liking.

In sum, the dynamic theory modification hypothesis predicts that initial natural kind categories should only be activated when the novel effect in Phase 2 is plausibly generated by the hidden causal structure underlying the natural kind. In contrast, according to the perceptual learning hypothesis the label of the effect (splenomegaly vs. liking) should make no difference as long as the learning procedure is the same.

4.1. Method

4.1.1. Participants and design

Forty-eight students from the University of Göttingen were randomly assigned to one of four conditions generated by crossing the factor category boundary (brightness vs. size) and type of effect (disease vs. liking).

4.1.2. Materials and procedure

The exemplars varied continuously in four dimensions: brightness, size, shape (number of corners), and number of molecules on the surface. The two relevant features in this experiment were size and brightness. The diameter of the viruses varied between 30 and 48 mm (size levels 1–4 in Fig. 5), and brightness was manipulated by using four equally spaced levels of grayness (20–80%, brightness levels 1–4 in Fig. 5). The two irrelevant features also came in four levels. The number of corners varied between 5 (pentagon) and 8 (octagon), and the number of molecules between 2 and 5. The four levels of all features were factorially combined, which yielded 256 different items. Our goal behind this large number of items was to discourage exemplar learning. Fig. 5 shows examples of the 16 crucial types of viruses that can be created by combining the four values of size and brightness.

We compared two conditions with orthogonal categories, which were either based on brightness or on size (see Fig. 5). Phase 1 was similar to the previous experiments. Participants learned to categorize different viruses into the two categories hemovedic and allovedic viruses, and were given corrective feedback after each judgment. From the class with 128 exemplars (marked by A in Fig. 5) a maximum of 120 were presented in random order in this phase. For the test phase, a subset of eight exemplars from this class was selected, in which all four feature values occurred with equal probability. Each exemplar was shown only once. Learning proceeded until participants met a learning criterion, 10 correct classifications in a row. Two conditions were compared: In the size condition, participants learned, for example, that the bigger viruses were allovedic, and the smaller ones hemovedic, in the orthogonal brightness condition they learned, for example, that the darker exemplars were allovedic and the lighter ones hemovedic. It was ensured that no other feature than the relevant one was correlated with the category and that the features were not intercorrelated. At the end of Phase 1 we asked participants to name the feature dimension they had used for the classifications.

The new factor was manipulated in Phase 2. One condition was taken from previous experiments. Participants had to learn whether the viruses caused splenomegaly or not. In the contrasting new liking condition different participants were told that designers had come across pictures of the viruses. The designers had thought that they represented novel patterns people might like. Therefore, they had conducted a study in which they had tested whether the viruses were liked or not. As in the disease condition it was pointed out that any outcome was possible including the possibility that the viruses never resulted in liking. Participants' task was to assess whether the pattern of the virus resulted in liking. The subsequent learning procedure was identical in both conditions except for the effect: In the disease condition participants observed whether a certain virus led to splenomegaly, and in the liking condition whether an exemplar was liked or not. In the causal learning phase participants saw a new set of 32 viruses one after another representing single instances of the viruses. On the backside of each card, information was given on whether the respective

virus had caused the particular effect. Each side (first the front side with the virus) was shown for roughly three seconds. In all conditions, the same items with identical associations with the effect were presented to participants. In the exemplars none of the irrelevant features was related to the presence of the effect. In contrast, the two relevant features were both equally related to the effect (see below for more details).

In Phase 3, the test phase, we switched back to exemplars corresponding to the items from the category-learning phase. These items had not been presented in the category-learning and the causal-learning phases. Participants received eight exemplars. Their task was to express their assessment of the likelihood that the respective virus would cause splenomegaly or liking by using a rating scale that ranged from 0 ("never") to 100 ("always"). After these ratings, participants also gave a general assessment of the likelihood that the two virus types, allovedic and hemovedic viruses, caused the respective effect.

Fig. 5 displays the statistical structure of the task. Letter A labels items presented in Phases 1 and 3, and B marks items from the causal-learning phase. To give the A items in Phases 1 and 3 a different physical appearance we varied the two irrelevant features. The figure gives an example of how these items were assigned to the two categories in Phase 1 and to the effect in Phase 2. In Phase 1, items could be correctly classified on the basis of a single feature with the feature levels 1 and 2 indicating one category, and 3 and 4 the contrasting category for either dimension (brightness or size). None of the other features or of the feature combinations were related to category membership. In Phase 2, both dimensions brightness and size were predictive for the effect. Each of these two dimensions correctly predicted the presence or the absence of the effect in 3 out of 4 cases (see Fig. 5). Combinations of specific levels of these features allowed even better predictions. For example, in Fig. 5 the large and light viruses caused splenomegaly and the small and dark ones did not. Viruses with other combinations of these two features had a 50% chance of causing the disease.

In our statistical analyses we distinguish between *critical items* that should yield different predictions in the contrasting category conditions, and *uncritical items* that should yield the same predictions in both conditions. In the example shown in Fig. 5, Items 1 and 6, and 11 and 16 are critical. Categorized according to size, Items 1 and 6 should be rated as weakly causally effective, as small viruses had only a probability of .25 (1 out of 4) to cause the disease. However, they should be considered highly causally effective when categorized according to brightness, because light viruses had an overall probability of .75 (3 out of 4) to cause splenomegaly. The opposite predictions hold for Items 11 and 16. They should attract high ratings when categorized according to size, and low ratings when categorized according to brightness.

Items 4 and 7, and 10 and 13 are uncritical ones. The first two items should yield low ratings of causal efficacy regardless of the categorization in Phase 1. If they were categorized according to size, they would belong to the small viruses, and if they were categorized according to brightness they would belong to the dark viruses. Either category has a probability of .25 to cause a swelling of the spleen. Therefore, similar estimates should be expected. Along the same lines, Items 10 and 13 should receive high ratings in both conditions.

We counterbalanced the assignment of labels to categories and the assignment of exemplars to the two learning phases (A, B). Moreover, we balanced which items were critical and which uncritical. To accomplish this, we rotated the effects in Fig. 5 clockwise by 90° so that Items 2, 3, 5, and 9 now showed the effect. Therefore, in this condition Items 4, 7, 10, and 13 became critical items.

4.2. Results and discussion

Table 3 shows the results. The most interesting finding concerns the critical items, which should be rated high or low depending on the category and the content of the effect. We expected to find that the categories from Phase 1 are used if the effect of the viruses was splenomegaly but not if the effect was a liking judgment. A 2 (brightness vs. size) × 2 (disease vs. liking) × 2 (Items 1 and 6 vs. Items 11 and 16) analysis of variance with the last factor being compared within subjects yielded a significant three-way interaction, $F(1, 44) = 4.52$, $p < .05$, $MSE = 623.0$. All other effects did not reach significance. In contrast, the same analysis with the within-subjects factor representing uncritical items (Items 7 and 4 vs. Items 10 and 13) that should in all category conditions be rated high or low showed no interaction but the expected highly significant main effect, $F(1, 44) = 157.6$, $p < .01$, $MSE = 390.2$. Moreover, a significant difference between the disease and liking condition was observed with the disease condition generally leading to higher ratings of causal efficacy than the liking condition, $F(1, 44) = 5.28$, $p < .05$, $MSE = 228.2$. This result can be interpreted as reflecting participants' intuition that disease-related symptoms are more natural effects of viruses than aesthetic impressions.

To further analyze the three-way interaction involving the critical items, we conducted separate analyses of variance for the disease and the liking conditions. The analysis of variance for the disease condition yielded a significant interaction between the factor representing the critical items (as a within-subjects factor) and the category boundary condition (as a between-subjects factor), $F(1, 22) = 5.57$, $p < .05$, $MSE = 691.6$. Both main effects failed to reach significance. In contrast, a separate analysis for the liking condition yielded no significant effects (all $Fs < 1$). These patterns reflect that the categories were only activated in the disease condition but not in the liking condition, which supports our prediction that category use is in part regulated by assumptions about the potential causal relevance of the categories with respect to the predicted effect.

The final ratings of the relationship between categories and causal effects show again that learners in both conditions encoded the relationship between categories and causal effects. Participants gave clearly different ratings for the two categories, $F(1, 47) = 248.3$, $p < .01$, $MSE = 218.1$ ($M = 73.3$ vs. $M = 25.8$). The differences were slightly higher for the virus condition ($M = 75.8$ vs. $M = 22.1$) than for the liking condition ($M = 70.8$ vs. $M = 29.6$) but the interaction was not nearly significant. Thus, replacing splenomegaly with liking did not result in differences of the difficulty of learning the category-effect relations. These findings provide further support for the hypothesis that the use of initial categories is an active decision in the test phase driven by assumptions about the plausibility of a causal relation between categories and effect.

Table 3
Mean ratings of the likelihood of the causal effect for critical and uncritical items (Experiment 3)

Categories	Disease condition				Liking condition			
	Critical items		Uncritical items		Critical items		Uncritical items	
	11/16	1/6	4/7	10/13	11/16	1/6	4/7	10/13
Size	51.3	44.2	29.6	80.4	48.3	55.4	24.6	63.8
Brightness	39.6	68.3	19.2	75.0	49.6	49.2	15.4	72.1

Whereas Experiments 1 and 2 used family resemblance categories which implied equal relevance of all features, Experiments 3 used one-dimensional categories that may lead to sensitizations to the relevant dimension. Although this perceptual learning process might explain transfer in the disease condition, it does not explain the absence of transfer in the contrasting liking condition. To explain the obtained effect, top-down factors need to be invoked.

Unlike in the previous experiments in which both the category boundaries in Phase 1 and the boundary separating effect-related items from non effect-related ones were equally complex, in the present structure the difficulty of the categories changed between Phases 1 and 2. Whereas in Phase 1 we used a one-dimensional category (based on size or brightness), the induction of categories that maximize predictability in Phase 2 would require learners to induce a more complex rule involving two dimensions. It is known that multidimensional rules are harder to learn than one-dimensional rules and possibly impossible to learn without corrective feedback (see Ashby, Queller, & Berretty, 1999; Ashby, Waldron, Lee, & Berkman, 2001). However, this gradient of difficulty may explain why learners stuck to the suboptimal categories in the condition in which categories and effect were related (disease) but makes the finding even more impressive that they neglected the categories in the contrasting condition (liking). Unfortunately, the many counterbalancing conditions do not allow us to deeply analyze what learners actually did when they neglected the initial categories. An alternative strategy to inducing a maximally predictive multidimensional boundary may be to pick a salient dimension or memorize individual items in Phase 2 and base the test responses on similarities to memorized exemplars.

5. General discussion

The starting point of the present research was the observation that causal induction is based on relationships between *categories* of causal events. The way objects or events are categorized influences the outcome of causal induction with otherwise identical input. We have developed a learning paradigm in which people first learn to categorize novel exemplars and then, in a second learning phase, observe contingencies between these exemplars and a novel causal effect. This paradigm allowed us to investigate under what conditions people continue to use the initial categories or to abandon these categories and induce new categories based on the predictive relationships between exemplars and the effect. The results generally show that people do not uniformly activate initially trained categories, although all participants learned the categories and encoded the statistical relation between categories and causal effect. Thus, our findings go beyond a pure bottom-up perceptual learning account that would generally predict transfer from initial categories in suitable training contexts. The results rather support the view that learners activate knowledge about the possible causal relation between categories and causal effect. Whenever the category labels suggest natural kinds, learners tend to use this category information when making inductive predictions about the test exemplars. Interestingly, this behavior could be seen in all experiments although we have always emphasized in the instructions of the natural kind conditions that the categories were solely based on superficial features. Apparently the label "virus" suggested deeper causal commonalities despite this instruction.

Our research also shows that people tend to ignore initial categories if these categories seem to be arbitrary collections of exemplars. In this case, learners rather induce new

categories in the causal learning phase. Moreover, Experiment 3 shows that learners do not use category-level information when the causal effect seems unrelated to the natural kind category (e.g., kinds of viruses and liking judgments).

We also obtained some interesting results about the categories people use when they decide to ignore the initial categories. Depending on the difficulty of the category structure, these new categories may attempt to maximize predictability (Lien & Cheng, 2000) (Experiment 2), but sub-optimal categories may be induced when the optimal categories are too complex (as in Experiment 3).

Finally, we found evidence in all experiments that the choice between alternative category systems occurred in the test phase, probably based on a knowledge-driven decision. In all conditions, the relation between categories and effect was encoded not only on the exemplar but also on the category level (Phase 1 categories).

5.1. Relations to previous research

In Section 1, we have already pointed out relations to other research paradigms that have addressed similar questions. Here, we would like to summarize our findings in the context of this research.

5.1.1. Causal contingency learning

Research on causal learning has focused on the acquisition of knowledge about causal links. The question how prior categories referring to causes and effects may affect contingency learning has not been addressed. Lien and Cheng (2000) have investigated how causal contingencies may underlie the induction of maximally predictive categories in novel uncategorized domains, but they did not investigate how categories from a previous learning context interact with current contingency learning. Our research goes one step beyond what Lien and Cheng have found. We showed that people continue to use previous categories whenever they are causally relatable. Learners transfer these categories even at the cost of sub-maximal predictability. However, whenever learners decide to neglect the prior categories they tend to induce new categories. In Experiment 2, we found evidence that people tend to induce categories that maximize predictability. In more complex domains other strategies including focusing on salient features may also be observed.

We believe that there is a tradeoff between the number of categories people are using for a set of exemplars and maximizing predictability. Technically people could achieve maximal predictability by inducing a new set of categories for each predicted feature. However, this is not parsimonious so that people often seem to settle for suboptimal predictability when it buys them a smaller number of category schemes.

5.1.2. Category learning as theory modification

Our research is also linked to categorization research. We have shown that the impact of categories on causal learning cannot be reduced to perceptual bottom-up sensitizations or unitizations (see Goldstone et al., 2000). Consistent with the view that categories are theory-based (Murphy & Medin, 1985), we have found that cause and effect categories are not simply represented as referring to a collection of similar exemplars but rather have an internal theoretical structure that interacts with the causal learning process. Thus, it seems more appropriate to view the relation between categories and causal learning as a process in which an initial causal theory underlying the category representation is modified and

extended by further causal learning processes. Causal knowledge, in everyday life as well as in science, is typically not acquired at one point in time after which it remains stable but is rather the result of a long process in which it undergoes dynamic changes, such as continuous modifications or even paradigm shifts (see Carey, 1991; Horwich, 1993). The development of neuroscience is a recent example. Many neuroscientific studies use categories coming from psychological or medical paradigms (e.g., studies on memory or on psychiatric disorders), and use these categories in novel hypotheses (e.g., about the neural basis of schizophrenia). Thus, new hypotheses interact with theories that underlie the old categories. These old categories may turn out to be useful but it is also possible that scientists switch to new categories that give them a better grasp of the present causal domain.

5.1.3. Natural kinds

Our research also builds on previous research on natural kinds (see Ahn et al., 2001; Gelman, 2003; Hirschfeld, 1996; Medin & Atran, 2004; Rehder & Hastie, 2004; Sloman & Malt, 2003; Wellman & Gelman, 1992). Our experiments are consistent with the view that there is a strong essentialist bias for natural kind concepts (Medin & Ortony, 1989). Despite the fact that our instructions only mentioned superficial criteria for grouping viruses, the two category labels (allovedic vs. hemovedic viruses) along with information that scientists settled on these categories seemed to be a strong cue for assuming natural kinds that may be responsible for further, yet unknown effects. In contrast, when viruses were grouped in blatantly arbitrary categories, no transfer was observed.

We also found that people had general expectations about the causal effects that categories might generate. This is consistent with Heit and Rubinstein's (1994) finding that people expect that categories have specific types of effects. Apparently people use abstract framework theories that specify the types of causes and effects that go together even when they lack specific knowledge (Wellman & Gelman, 1992).

Strevens (2000) has proposed an alternative view, the minimalist view, to psychological essentialism. According to this view causal laws that connect kinds of concepts with visible features underlie natural kind representations. These laws do not require the assumption of a stable uniform underlying essence (see Ahn et al., 2001, for a response). According to the minimalist view features of natural kinds are caused, but what causes it may vary from feature to feature. Moreover, people may have no intuitions about the kind of causes.

The present experiments are consistent with the view that natural kinds are viewed as causal models in which visible features are related to hidden causes. In this respect they are consistent with minimalism as well as the variant of essentialism that ascribes a belief in causal essence placeholders to people (e.g., Ahn et al., 2001; Gelman, 2003). It seems unlikely that our learners have elaborate beliefs about the nature of the essence underlying allovedic and hemovedic viruses (see also Rozenblit & Keil, 2002). Also our results do not necessarily require the essence placeholders to refer to a single common cause (Ahn et al., 2001). It may well be that people have a rather diffuse belief in a hidden causal structure, which is consistent with Streven's view that different features may be viewed as being generated by different parts of the hidden causal model.

However, other aspects of our data seem more consistent with essentialism than minimalism, which lacks constraints for the kind of causes underlying the categories. For example, Experiment 2 has shown that an arbitrary collection of viruses is not seen as a causally homogeneous kind. Although each virus exemplar has features that are certainly caused by something, learners' reluctance to use arbitrary categories seems to signify that they did

not believe that there is a uniform common causal structure underlying all exemplars of the categories. Therefore, they did not use these categories for predicting novel effects. Moreover, Experiment 3 shows that learners had some specific beliefs about what kind of novel features the categories can generate. Although, there was no prior knowledge about the relation between our artificial viruses and splenomegaly, learners tended to have an abstract belief that these viruses are potential causes of all kinds of disease-related symptoms including splenomegaly. This is also consistent with the view that our causal learning is governed by a causal grammar that specifies the kinds of effects causes can potentially have (Tenenbaum, Griffiths, & Niyogi, in press). Our experiments clearly show that learners believe that novel natural kinds share a hidden, unknown causal structure which is a possible generator of specific kinds of novel features and is inconsistent with others (e.g., liking).

5.1.4. Category-based induction

Category-based induction is a paradigm in which questions related to our research have been addressed (see Coley, Medin, Proffitt, Lynch, & Atran, 1999; Gelman, 2003; Heit, 2000; Murphy, 2002, for overviews). Although our theoretical predictions were strongly influenced by this research (see Introduction), it is important to point out the differences between the paradigms. In a typical category-based induction task participants may be informed that one or more specific categories of exemplars have a novel feature (e.g., "All robins have a spleen"). Participants are then asked whether a different category probably has this feature as well (e.g., "All animals have a spleen" or "All ostriches have a spleen"). There are a number of key differences to our paradigm (see also Fig. 1): (1) The categories in these tasks are not learned but given. In contrast, we manipulated the categories by training novel categories using a trial-by-trial learning procedure. (2) Moreover, in our causal learning phase we teach the contingencies between exemplars and an effect in a trial-by-trial learning procedure so that the responses in the test phase are in part driven by the learned contingencies. In contrast, in category-based induction research no contingency learning takes place on the exemplar level; the relation between the category and the features is simply asserted. (3) The most important difference is that in our paradigm learners have a choice to either activate category-level information or exemplar-level information (see Fig. 1) during learning and testing, whereas in category-based induction participants are forced to use the given categories. Thus, our paradigm enabled us to investigate the question under what conditions people activate prior categories and under what conditions they rather choose to neglect these categories.

Nevertheless, our predictions and findings are consistent with a number of discoveries in this research area. Consistent with the findings of Murphy and Ross (1994) we have found that people like to use category knowledge when they are asked to predict relations between features (see also Malt, Ross, & Murphy, 1995; Ross & Murphy, 1996). However, we also demonstrated that prior knowledge influences whether people go through the upper category-level route (Fig. 1) or neglect the previously learned categories.

Our findings are also consistent with the findings of Gelman and her collaborators (Gelman & Coley, 1990; Gelman & Markman, 1986; see also Gelman, 2003) that people prefer to use natural kind categories over similarity-based orderings when making inductive inferences, and that they only generalize features that are consistent with general assumptions about effects of natural kinds (see also Heit & Rubinstein, 1994). Finally, some of our data (Experiments 1 and 2) is consistent with findings that show that inductions are

stronger when the test exemplars are more typical for the category (see Osherson, Smith, Wilkie, López, & Shafir, 1990; Rehder & Hastie, 2004).

5.2. Directions for future research

A number of different questions were left open in the present research, which should be addressed in future studies.

5.2.1. Natural kinds vs. artifacts

Our research has focused on natural kinds. It would be interesting to contrast natural kinds with artifacts. This might provide further evidence on the differences between kinds of concepts, and may also be relevant for the debate between essentialism and minimalism. There is agreement in the literature that artifacts such as cars, refrigerators, or chairs do not have a hidden, unknown causal essence in the sense of natural kinds (see Ahn et al., 2001; Bloom, 2000; Gelman, 2003; Medin & Atran, 2004; Medin et al., 2000; Rehder & Hastie, 2004; Sloman & Malt, 2003), although it is argued by some researchers that the function intended by the designer may play a corresponding role. On all these accounts it is less likely that artifacts exhibit novel, yet unknown causal features that may be discovered in the course of research.

The crucial factor underlying this prediction is whether the causal structure underlying the categories is *known or unknown*. Natural kinds are not constructed but discovered. Therefore, they typically have a hidden, unknown causal structure that is open to future discovery of novel properties (Putnam, 1975). In contrast, artifacts are constructed by human designers so that it is less likely that their internal causal structure will surprisingly show novel, unexpected effects. Few studies have addressed this question. Gelman (1988), and Gelman and O'Reilly (1988) report studies that show that natural kind categories support more inductive inferences than artifact categories. However, other researchers found no difference (Rehder & Hastie, 2004). It may be interesting to use our task to test the hypothesis that natural kinds afford a wider range of novel inductive inferences than artifacts.

Although we expect differences in the willingness to accept novel causal effects for natural kinds compared to artifacts, we can imagine a number of cases in which artifact categories might support causal learning of novel effects. Of course, we would predict transfer if the causal effect belongs to the functions of the artifacts or is a plausible known side effect. For example, the noise of a refrigerator brand may be plausibly related to different brands (with more expensive ones being less noisy). Or with art objects we might have the intuition that viewers find classical art more pleasing than modern art. However, this would not be an effect that is newly discovered, it rather is an effect that we would already expect based on our prior knowledge. Transfer may also be observed when the causal effect is related to the materials of the artifact (e.g., chemical substances of the cooling system of a refrigerator as a cause of disease). However, in this case the task is not primarily about an artifact (refrigerator) but about a natural kind (cooling fluid).

5.2.2. The role of the learning input

We contrasted our dynamic theory modification hypothesis with the perceptual learning hypothesis. Although the results support our theory, this does not mean that bottom-up factors do not play a role. We only intend to argue that the effects we investigated cannot

be reduced to mere perceptual learning. In the present research we have focused on top-down factors influencing whether people stick to old categories or induce new ones. It would be interesting to extend this research and investigate possible bottom-up factors that also may have an influence on people's learning strategies.

In all experiments we chose to train the categories (Phase 1) with corrective feedback, whereas causal learning was observational. In our view, this is a natural first choice. We were interested in possible transfer of category knowledge; therefore it was a prerequisite to make sure that all participants had learned the categories. Moreover, categories are typically mutually exclusive with each exemplar belonging to only one of the categories. In contrast, causal effects are in most cases probabilistic; we even accept small probabilities as valid reflections of causal relations (e.g., pollution and lung cancer). A corrective feedback-based training regime would have created the wrong impression that the task was to learn a deterministic relation. Of course, with sufficient training and corrective feedback almost all participants may switch to alternative categories. But our goal was to investigate the early stages of learning in which we could observe whether learners choose to stick to the old categorical schemes or induce new ones. Nevertheless, it may be interesting to explore other training regimes.

Another interesting research question might focus on the statistical structure of the learning input. In the present experiments, the prior categories were fairly strongly related to the effect. The covariation was not perfect but also far from zero. Thus, the prior categories were pragmatically useful. However, there may be situations in which people become aware of the fact that a continued use of categories, even when their labels suggest relevance, does not allow them to make good predictions. This may be the case when the covariation between categories and effect turned out to be close to zero. Thus, although the categories may sound promising, they may not generate useful causal knowledge that furthers learners' ability to predict and explain. We hypothesize that this might also be a situation in which people tend to abandon prior categories. Our general hypothesis is that there is a trade-off between sticking to familiar conceptual schemes and predictability. As long as the familiar categories yield satisfactory predictions, people might rather opt against learning alternative categories for the same domain. But when predictability is below some threshold, then people might be inclined to start from scratch. This process seems to be analogous to what we occasionally see in the dynamics of scientific research.

Moreover, one could further explore the relative difficulty of the categories. In Experiments 1 and 2 we used linearly separable categories in Phase 1 and the category boundary that separated effects from non-effects in Phase 2 was also linearly separable. In contrast, in Experiment 3 the initial categories were based on a one-dimensional rule whereas the categories distinguishing exemplars that generate effects from the rest required multidimensional boundaries. Our design did not allow us to study the induction strategies of individual participants in Phase 2; we were mainly interested in whether they used prior categories or neglected them. It would be interesting to study the inductions in this phase further. Moreover, we would like to know whether people are aware of such difficulty gradients between Phases 1 and 2, and whether their tendency to switch to new categories is affected by the perceived relative difficulty of the categories.

A related research question might address the processes of inducing new categories in Phase 2 when learners have decided to abandon the initial Phase 1 categories. In Experiment 2 we found evidence for the induction of fuzzy categories that reflected the predictability gradient of the set of exemplars. In other cases people may fall back to sub-optimal

rules. Interestingly, participants in these conditions nevertheless encoded the relationship between the initial categories and the effect, even though this knowledge was not used in the test phase. We would be interested in finding out how knowledge about the old categories (upper route in Fig. 1) and the new categories (lower route in Fig. 1) is acquired in parallel, and how the two routes interact.

5.3. Conclusion

The hypothesized impact of pre-existing categories on causal learning constitutes a new type of transfer effect. Unlike in research on analogical transfer (see Holyoak & Thagard, 1995), no specific relational knowledge is transferred. The transfer effect is rather based on people's decisions to apply the categories, which themselves are based on knowledge about the causal structure underlying the concepts and the observable contingencies. Thus, categories and causal inferences are related in a dynamic interplay of theory change and new beginnings. The present results show that the outcome of this dynamic process of theory development may be crucially dependent on how it started.

References

Ahn, W., Kalish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., et al. (2001). Why essences are essential in the psychology of concepts. *Cognition, 82*, 59–69.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61*, 1178–1199.

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General, 130*, 77–96.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Science, 21*, 547–611.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 257–291). Hillsdale, NJ: Erlbaum.

Clark, A., & Thornton, C. (1997). Trading spaces: Computation, representation and the limit of uninformed learning. *Behavioral & Brain Sciences, 20*, 57–90.

Coley, J. D., Medin, D. L., Proffitt, J. B., Lynch, E., & Atran, S. (1999). Inductive reasoning in folk-biological thought. In D. L. Medin & S. Atran (Eds.), *Folkbiology*. Cambridge, MA: MIT Press.

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.

Dupré, J. (1993). *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 241–256.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234–257.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20*, 65–95.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*, 796–804.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183–208.

Gelman, S. A., & O'Reilly, A. W. (1988). Children's inductive inferences within superordinate categories: The role of language and category structure. *Child Development, 59*, 876–887.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227–247.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*, 178–200.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130*, 116–139.

Goldstone, R. L., Steyvers, M., Spencer-Smith, J., & Kersten, A. (2000). Interactions between perceptual and conceptual learning. In E. Diettrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 191–228). Mahwah, NJ: Lawrence Erlbaum Associates.

Goodman, N. (1978). *Ways of worldmaking*. Indianapolis, IN: Harvester Press.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts and theories*. Cambridge, MA: MIT Press.

Hacking, I. (1993). Working in a new world: The taxonomic solution. In P. Horwich (Ed.), *World changes: Thomas Kuhn and the nature of science* (pp. 275–310). Cambridge, MA: MIT Press.

Hacking, I. (2000). *The social construction of what*. Cambridge, MA: Harvard University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review, 7*, 569–592.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 411–422.

Hirschfeld, L. A. (1996). *Race in the making: Cognition, culture, and the child's construction of human kinds*. Cambridge, MA: MIT Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps*. Cambridge, MA: MIT Press.

Horwich, P. (Ed.). (1993). *World changes: Thomas Kuhn and the nature of science*. Cambridge, MA: MIT Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kuhn, T. S. (1962). *Structure of scientific revolutions*. Chicago, IL: Chicago University Press.

Kunda, Z. (1999). *Social cognition: Making sense of people*. Cambridge, MA: MIT Press.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 754–770.

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

Malt, B. C., Ross, B. H., & Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 646–661.

Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review, 111*, 960–983.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32*, 49–96.

Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology, 51*, 121–147.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge, MA: Cambridge University Press.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology, 27*, 148–193.

Nozick, R. (2001). *Invariances: The structure of the objective world*. Cambridge, MA: Harvard University Press.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 811–828.

Putnam, H. (1975). The meaning of "meaning". In H. Putnam (Ed.), *Mind, language, and reality. Philosophical papers* (Vol. 2, pp. 215–271). London: Cambridge University Press.

Putnam, H. (1987). *The many faces of realism*. LaSalle, IL: Open Court.

Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141–1159.

Rehder, B. (2003b). Categorization as causal reasoning. *Cognitive Science, 27*, 709–748.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130*, 323–360.

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

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*, 736–753.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross classification and category organization in a complex real-world domain. *Cognitive Psychology, 38*, 495–553.

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

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21*, 1–54.

Shanks, D. R., Holyoak, K. J., & Medin, D. L. (1996). *The psychology of learning and motivation, Vol. 34: Causal learning*. San Diego: Academic Press.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition, 52*, 1–21.

Sloman, S. A., & Malt, B. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Language and Cognitive Processes, 18*, 563–582.

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

Tenenbaum, J. B., Griffiths, T. L., Niyogi, S., in press. Intuitive theories as grammars for causal inference. In: A. Gopnik & L. Schulz (Eds.), *Causal learning*: Psychology, philosophy, and computation. Oxford University Press.

Thagard, P. (1999). *How scientists explain disease*. Princeton, NJ: Princeton University Press.

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: Academic Press.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 53–76.

Waldmann, M. R. (2001). Predictive versus diagnostic learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin and Review, 8*, 600–608.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124*, 181–206.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*, 337–375.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 776–795.