Running head: Unbroken Mechanism Hypothesis

Category Transfer in Sequential Causal Learning:

The Unbroken Mechanism Hypothesis

York Hagmayer[1], Björn Meder[2], Momme von Sydow[1] and Michael R. Waldmann[1]

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

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

Address for proofs:
York Hagmayer
Department of Psychology
University of Göttingen
Gosslerstr. 14
37073 Göttingen
Germany
Phone: +49-551-398293
Fax: +49-551-393656
york.hagmayer@bio.uni-goettingen.de

Abstract

The goal of the present set of studies is to explore the boundary conditions of category transfer in causal learning. Previous research has shown that people are capable of inducing categories based on causal learning input, and often transfer these categories to new causal learning tasks. However, occasionally learners abandon the learned categories and induce new ones. Whereas previously it has been argued that transfer is only observed with essentialist categories in which the hidden properties are causally relevant for the target effect in the transfer relation, we here propose an alternative explanation, the *unbroken mechanism hypothesis.* This hypothesis claims that categories are transferred from a previously learned causal relation to a new causal relation when learners assume a causal mechanism linking the two relations that is continuous and unbroken. The findings of two causal learning experiments support the unbroken mechanism hypothesis.

Category Transfer in Sequential Causal Learning:

The Unbroken Mechanism Hypothesis

1. Introduction

There is a strong tendency in cognitive psychology to compartmentalize research into different areas, such as memory, learning, categorization, or decision making. As a result, there has been little contact between these fields, which has led to notable blind spots. We will focus on one particular example of mutual blindness, namely research on causal and category induction, which traditionally have been treated as separate learning phenomena.

Typically, studies on causal learning present learners with pre-categorized potential causes (e.g., presence or absence of fertilizers) that could be potentially related to pre-classified effects (e.g., presence or absence of blooming). The categories referring to causes and effects have been treated as unproblematic givens; thus, the only remaining task was to learn about the existence and strength of the causal relations (e.g., Shanks, Holyoak, & Medin, 1996; De Houwer & Beckers, 2002).

Research on categorization has largely neglected the role of causal information for category induction. This is particularly clear for theories that solely focus on the role of similarity in the formation of categories (see Murphy, 2002). However, even within the paradigm of theory-based categorization (Murphy & Medin, 1985), the main focus has been on internal category structure, that is, the causal and functional relations that link features within categories. For example, disease categories can often be represented as common-cause models with the category features representing causes (e.g., viruses) and effects (e.g., symptoms). It can be shown that the type of causal model linking separate features within such categories influences learning, typicality judgments, and inductive inferences (Rehder, 2003a, b; Rehder & Hastie, 2001, 2004;

Waldmann, Holyoak, & Fratianne, 1995; Waldmann, 1996, 2000, 2001). However, the interrelation between learning categories of cause and effect and the induction of the causal relations linking the category members with other events has been neglected.

1.1 The Tight Coupling of Category and Causal Induction

In a seminal study, Lien and Cheng (2000) explored the relationship between category learning and causal induction. In their learning experiments, participants were presented with a set of uncategorized cause exemplars, which could be classified at different hierarchical levels of abstraction. No category labels were provided. Instead participants only observed which cause exemplars (different types of substances) generated a specific causal effect (blooming of flowers). The question was how participants would categorize the cause events in the absence of any explicit information on category structure. The results of the experiments showed that learners categorized the exemplars at the hierarchical level that was most predictive for the effect. Lien and Cheng (2000) interpreted this as evidence for their *maximal contrast hypothesis*: People tend to induce categories that maximize predictiveness.

More recently, Marsh and Ahn (2009) have reported converging results. In their studies, participants were presented with exemplars that varied on a continuous dimension (e.g., high, intermediate, low height of bacteria). Marsh and Ahn manipulated the assignment of the exemplars to a binary effect (e.g., presence or absence of a protein). For example, in one condition only exemplars with high and intermediate values, but not those with low values caused the effect. This was contrasted with a condition in which only exemplars with high values caused the effect, but not exemplars with intermediate or low values. The results showed that learners tended to categorize the cause exemplars according to the boundaries entailed by the effect. Thus, the ambiguous intermediate value was classified together with the high value when

both caused the effect; otherwise it was classified with the low value. These findings support Lien and Cheng's (2000) theory by showing that learners attempt to categorize causes according to their effects and create categories with a maximum in predictability. In sum, previous research has shown that subjects categorize cause exemplars according to their effects. Here, we will investigate whether effect exemplars are categorized according to their causes (i.e., whether people form cause-based categories).

Whereas Lien and Cheng (2000) and Marsh and Ahn (2009) showed that people categorize exemplars according to the features that maximize predictiveness, Kemp, Goodman, and Tenenbaum (2010) were interested in whether people categorize non-discriminable objects according to their causal power. In their experiments they presented subjects with perceptually indistinguishable blocks that either activated a machine or did not. In Experiment 1 Kemp and colleagues manipulated the grouping of the blocks with respect to their causal power. For example, in one condition four blocks never activated the machine and four blocks activated the machine half of the time. The machine was never active in the absence of a block. It was expected that learners would induce two classes of otherwise indistinguishable blocks which differ with respect to their causal power. To test this prediction, learners were confronted with single trials of novel test blocks. Although subjects saw the test blocks either activating or not activating the machine only once, they were capable of predicting the effects of a test block in a hypothetical setting in which the block would be placed inside the machine multiple times. For example, when the test block activated the machine, subjects inferred in the condition described above that the test block probably belongs to the category of blocks with intermediate causal power, whereas it probably belonged to the other category when it failed to activate the machine. Thus, learners used previously acquired category knowledge to make the inductive inference.

Although learners had difficulties in a condition in which both categories had probabilistic causal power (0.1 vs. 0.9), additional experiments clarified that this difficulty can be overcome if additional feature information aiding the categorization process is made available. Another interesting finding was that learners tended to abandon the previously induced categories and induce new ones when the behavior of the test object seemed to be inconsistent with the previously observed categories. For example, when a new test block activated the machine often, whereas previously observed blocks did not, learners tended to conclude that they are observing an example from a new category. This finding shows that learners may use bottom-up statistical knowledge about causal power to decide whether a novel exemplar belongs to a previously seen or a new category.

1.2 Category Transfer across Multiple Causal Relations

The studies reported in the previous section examined the interplay between category and causal learning within single cause-effect relations. While maximal predictiveness can easily be defined when only a single cause-effect relation is considered, the situation becomes more complex when the category members are involved in *multiple* causal relations. In such a situation, each causal relation may in principle entail a different maximally predictive categorical scheme. These schemes may potentially be conflicting.

In the present research we studied categories involved in multiple causal relations. In particular, we presented participants with causal chains linking three entities[1], $A \rightarrow B \rightarrow C$. Assume, for example, that an initial cause *A*, radiation, influences an intermediate entity *B*, viruses, which in turn may cause a swelling of the spleen (i.e., splenomegaly) (event *C*). Whereas *A* and *C* are both presented as pre-categorized binary events (radiation vs. no radiation, and splenomegaly vs. no splenomegaly), *B* is a set of uncategorized exemplars (viruses). Thus, the

intermediate entity *B* is part of two causal relations, each of which could be used as a basis of causally motivated categorization. Using one of the two relations to induce categories yielding maximal contrasts will lead to optimal categories for the respective single relation, but these categories may not necessarily be optimal for the other relation.

If both causal relations were learned at once, the computational problem would be to create categories for entity *B* that are globally optimal for predicting both related events, although they may not be locally optimal with respect to either. Alternatively, one could try to induce two category schemes from the causal information which would not necessarily have to overlap. In the present research we will not study learning situations in which information about multiple, interconnected causal links is simultaneously presented. We will instead focus on a case which seems more frequent in real-world learning. We rarely learn about all relations of a causal model at once but rather acquire causal knowledge in fragments, which later are integrated into more complex causal models (Fernbach & Sloman, 2009; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Waldmann, Cheng, Hagmayer, & Blaisdell, 2008; Waldmann, Hagmayer, & Blaisdell, 2006). For example, people may first learn about the causal relation *A*→*B* and use *A* to categorize *B* (i.e., induce a cause-based category). In a later learning context they may be presented with the causal relation *B*→*C*. The main question of interest in the present research is under which conditions the categories for *B* (induced in the initial learning phase; *A*→*B*) will be transferred to learning about *B*→*C*, at the possible cost of reduced predictiveness for the second causal relation. Alternatively, learners could abandon the previously acquired category scheme for *B* and induce new, effect-based categories for *B* better suited for predicting event *C*.

Waldmann, Meder, von Sydow, and Hagmayer (2010) presented a first set of studies investigating this type of category transfer within causal chains. In one experiment we used the three events mentioned above, radiation (*A*), viruses (*B*), and splenomegaly (*C*) within a two-phase causal learning paradigm. Importantly, no feedback about category labels of the events *B* was provided. Learners observed the virus exemplars (event *B*) along with information about their causes (event *A*) in the first causal learning phase, and the virus exemplars with information about their effects (event *C*) in the second causal learning phase. The general finding was that participants indeed tended to stick to the initially acquired virus categories, and used this category scheme to learn about the second causal relation. Interestingly, this was the case although a more predictive classification with respect to the final effect of the chain existed. Thus, learners did not form different category schemes for different causal relations but tended to favor category parsimony over flexibly re-categorizing the same entities with respect to the learning context.

1.3 Boundary Conditions of Category Transfer

In the present set of studies we are interested in exploring the boundary conditions of category transfer. Although Waldmann et al. (2010) generally found evidence for category transfer, this may not universally be the case. In fact, previous experiments (Waldmann & Hagmayer, 2006) suggest that there may be circumstances in which learners might be reluctant to transfer categories. In these studies, a two-phase learning paradigm was used in which a supervised category learning phase was followed by a causal learning phase involving the same or similar exemplars. Thus, unlike in the studies discussed above, explicit feedback about category labels was given in the initial category learning phase. For example, participants first learned to categorize fictitious viruses into two mutually exclusive classes (e.g., allovedic vs. hemovedic

viruses). In the subsequent causal learning phase the exemplars were presented along with information about the presence or absence of a causal effect (splenomegaly). In these studies, participants also revealed a strong tendency to continue to use old conceptual schemes rather than inducing new ones. However, category transfer depended on the relevance of the categories for the causal effect. Whenever the category labels suggested natural kinds that could be plausibly related to the causal effect, transfer was observed. But when the categories were arbitrary, or could semantically not be linked to the causal effect, learners abandoned the categories and induced a novel set of categories that mirrored the category boundaries entailed by the causal effect. This was demonstrated in Waldmann and Hagmayer's (2006) Experiment 3, where participants first learned that exemplars could be classified into two types of viruses. In the subsequent causal learning phase the same virus exemplars were introduced as blueprints for aesthetic patterns used in interior design. Participants' task was to learn which virus patterns people liked. Learners abandoned the previously acquired categories and induced a new scheme that was maximally predictive for the causal relation.

One possible interpretation of these findings is that they support *psychological essentialism* (Medin & Ortony, 1989), which is claimed to underlie the naïve representation of natural kinds from childhood on (see also Ahn et al., 2001; Gelman, 2003; Rehder, 2007; Rehder & Kim, 2006). According to this theory, people tend to ascribe stable hidden essences to natural kind categories, such as viruses, which may cause various visible features. Once learners believe that their categories refer to something real and stable in the world, they should be reluctant to change these categories even when they only generate weak probabilistic relations in future causal relations. For example, many people treat gender or race as a natural kind category, and are perfectly willing to accept weak probabilistic relations instead of looking for more predictive

categorizations of people (see Hirschfeld, 1996). Thus, transfer should be observed in our paradigm when learners believe that the same essentialist natural kind categories are causally relevant in the two presented causal relations. Viruses are certainly plausible generators of splenomegaly so that learners may have felt that the virus categories are causally relevant in the causal learning phase, and should therefore be re-used. However, in Experiment 3 of Waldmann and Hagmayer (2006), the cover stories suggested that the virus categories from the first learning phase were not relevant for the relations between the viruses and the aesthetic assessment in the causal learning phase, which may have led to the tendency to abandon the initial categories and re-categorize the stimuli.

1.4 The Unbroken Mechanism Hypothesis

The results of Waldmann and Hagmayer (2006), however, are theoretically ambiguous. The observed category transfer may only occur when people believe that the hidden ("essentialist") features of natural kind categories are linked to the effect presented in the causal learning phase. Although we believe that psychologically essential features do play an important role in eliciting category transfer, category transfer in causal induction needs not to be restricted to the use of essential features. Instead, category transfer may depend on more abstract features of people's causal theories, whether essences play a role or not.

We have developed the *unbroken mechanism hypothesis*, which posits that it is the causal relevance of the involved features and peoples' assumptions regarding the involved causal mechanisms that drive category transfer. This hypothesis is inspired by the idea that people have strong intuitions that statistical contingencies arise from the operation of (often unobservable) causal mechanisms that specify how the cause events generate (or inhibit) the effect events (see Ahn, Kalish, Medin, & Gelman, 1995; Ahn & Kalish, 2000; Griffiths & Tenenbaum, 2009;

Waldmann, 1996). These intuitions need not be very precise or correct, they may also be vague or faulty (Rozenblit & Keil, 2002). Nevertheless, they might play an important role in the way we perceive, interpret, and represent data.

The central idea behind the unbroken mechanism hypothesis is that transfer of causally induced categories to further learning episodes depends on whether learners assume a continuous causal mechanism connecting different categories of events. Whenever learners assume an unbroken causal mechanism, we expect participants to induce a coherent category scheme comprising all involved causal relations. Thus, for transfer it is not sufficient that the same entities are involved in multiple relations to form a complex causal model; rather the *same properties or features* of the entities must be causally relevant for both causal relations so that the same reference classes are picked out for the two causal relations. If *different, causally unconnected* features of the same entities are involved in multiple causal relations, learners might opt for inducing new, more predictive classification schemes. In this case, different sets of categories may be used for the two causal relations.[2]

The simplest variant of an unbroken mechanism are cases in which the same features of the category members are causally relevant. Assume the causal chain's initial event is radiation which affects the DNA of a set of uncategorized virus exemplars (second event). The viral DNA, in turn, might determine whether a given virus exemplar does or does not cause a disease (final event). In this case the same features (i.e., DNA) of the category members (i.e., viruses) are causally relevant for both relations. By contrast, if the chain's initial event affects the surface features of the viruses while the viruses' DNA is responsible for the final effect in the causal chain, different features of the entity would be relevant for the two causal relations. In this case,

different category sets may be induced for each of the relations, hence no category transfer should be observed.

There are cases in which different properties of the category members are involved in the two causal relations, but nevertheless an unbroken mechanism exists. For example, if in the first relation the DNA of the category members is involved, and the second causal relation is triggered by surface features, then an unbroken mechanism might still be in place if learners assume that the surface features of the category members are caused by their DNA. Although different causal features are involved in the two relations, the internal causal structure of the category links the different features with each other. This causal model entails that categories describing the status of the DNA are viewed as direct causes of the categories, summarizing the surface features. Therefore the DNA is viewed as an indirect cause of the target effect. Thus, this is again an example of an unbroken mechanism. Hence transfer is predicted.

Waldmann and Hagmayer's (2006) Experiment 3 provides an example of a special case in which the initial cause is linked to the surface features of the viruses via the DNA, and a configuration of the surface features is linked to the final effect (aesthetic judgment). Nevertheless, in our view this is a case of a broken mechanism. Participants had no reason to assume that the relevant features and their configurations that are caused by the DNA are the same as the causally relevant features and feature configurations influencing aesthetic judgments. Therefore different categories were derived for the two causal relations (see also General Discussion).

1.5 Preview of Experiments

In the present set of experiments, we studied the role of unbroken versus broken mechanisms in causal chains containing three entities $A \rightarrow B \rightarrow C$. The learning input was kept constant across

conditions, while learners' assumptions regarding the underlying hidden mechanisms causally connecting the three entities were manipulated. In particular, learners' assumptions about the causal relevance of different features of the intermediate entity and its internal causal structure were manipulated.

Insert Figure 1 about here

There are a number of ways in which the two causal links constituting a causal chain model can be linked to each other. The range of these possibilities is illustrated in Figure 1. One possibility, tested in Experiment 1a, is that the initial entity *A* affects hidden features of the intermediate entity *B.* These hidden features are not directly observable for the subjects but can be inferred from visible surface-features which are caused by the hidden features. For example, different types of microbes (entity *A*) might affect hidden features of protozoa's DNA (property of entity *B*) thereby systematically altering the protozoa's visible appearance (second property of entity *B*). If the hidden features are assumed to be causally responsible for the final effect (a swelling of the spleen), the causal mechanism underlying the observable correlations is unbroken. This is because the feature (i.e., the DNA) affected by the initial cause also affects the final effect (Causal Model I in the left column of Fig. 1). For this scenario, the unbroken mechanism hypothesis predicts category transfer.

We contrasted this case in Experiment 1b with an example of a broken mechanism in which the two causal relations were not linked within entity *B.* In particular, we investigated a situation in which the initial cause entity *A* affects the visible surface features of the intermediate entity but the hidden features of this intermediate entity are causally responsible for the final effect *C* (Causal Model II in the left column of Fig. 1). Here, different features of entity *B* are

causally relevant for the two relations and there is no causal link between hidden and surface features. In contrast to essentialism, the unbroken mechanism hypothesis predicts that the initially induced categories of *B* would not be used when learning about the second causal relation, because there is no causal link between the surface and the hidden features of *B*.

The goal of Experiment 2 was to further explore the boundary conditions of category transfer. The experiment consists of a set of four closely related studies, in which participants were presented with different causal models containing different types of causal mechanisms. Whereas the unbroken causal mechanism in Experiment 1 is mediated through the hidden features of entity *B*, there are further cases of unbroken mechanisms which could link an initial cause *A* with a final effect *C* (Fig. 1, right column). A second possibility for an unbroken causal mechanism is that the initial cause *A* affects the hidden features of the intermediate entity *B* which in turn influence its visible surface features (Causal Model II in the right column of Fig. 1). The surface features of the entity *B* are then causally relevant for the final effect *C*. This is a different version of an unbroken mechanism linking the components of the causal chain. A further case of an unbroken mechanism is a situation in which the initial cause *A* directly affects the surface features of entity *B*, and these features in turn cause the final effect *C* (Causal Model IV in the right column of Fig. 1). Note that no hidden (or essentialist) features are involved in this case. Nevertheless, transfer is predicted due to the presence of a continuous series of mechanisms. Such a finding would be critical for the position that transfer should only be observed when essentialist features of natural kinds are involved.

Experiment 2 investigated all three kinds of unbroken mechanisms (Causal Models I, II, and IV in Fig. 1, right column) and contrasted them with a case in which the causal mechanism is broken. In this case, the initial cause *A* affects hidden properties of entity *B*, but the final effect *C*

is only causally dependent on surface features of *B* with no causal relation linking the hidden properties with the causally relevant surface features (Causal Model III in Fig. 1, right column). Although the essential features are linked to the first cause, no category transfer is predicted by the unbroken mechanism hypothesis.

In summary, the central goal of both experiments was to test a novel hypothesis about the boundary conditions for category transfer in sequential causal learning. Whereas previous accounts, inspired by psychological essentialism, assume that transfer is primarily governed by the causal role of hidden, essentialist features of natural kinds, we propose a more general hypothesis. According to this hypothesis, people's beliefs in the connectedness of the causal mechanisms underlying a causal chain drive category transfer, rather than the involvement of essentialist features. Thus, the hypothesis predicts that no transfer will be observed if the mechanism is broken even when the natural kinds' essentialist features are involved (either as causes or as effects). Conversely, category transfer is predicted in situations involving an unbroken mechanism even when no essential features of the categorized entities are causally relevant.

2. Experiment 1

Experiment 1 provides a first test of the unbroken mechanism hypothesis. We ran two nearly identical experiments (Experiments 1a and 1b) using the two causal models shown in the left column of Figure 1. Experiment 1a focused on the upper causal chain (Model I) in which the initial event and the final effect are connected by an unbroken causal link. The causal chain is continuous in the sense that the first cause affects the intermediate entity´s hidden ("essentialist") features which, in turn, causally influence the final effect. By contrast, in Model II (Experiment 1b) the first event causally affects the visible surface features of the intermediate entity, but the

second causal link originates in the hidden ("essentialist") features of the intermediate entity. In this case the causal mechanism is broken because different, not causally related features of the intermediate entity are involved in the two causal relations. Hence, we expected to see a transfer of the cause-based category scheme of the intermediate entity to a second learning phase (concerning the relation of the intermediate entity and the final effect) given Model I but not Model II.

In order to test this hypothesis we used a two-phase learning paradigm in which subjects consecutively learned about two causal relations ($A \rightarrow B$, $B \rightarrow C$) overlapping in a middle entity. Whereas *A* and *C* were dichotomous causal events, the intermediate entity *B* was a set of uncategorized objects. In both Experiment 1a and 1b we manipulated between subjects the categories that cause *A* entailed for the uncategorized objects of *B*. The second causal learning phase was identical for all participants. The manipulation of the category structure in the initial causal learning phase allowed us to examine whether subjects transfer the category scheme entailed by the causal relation $A \rightarrow B$ to the second causal relation $B \rightarrow C$. A transfer would lead to different judgments about the causal power of objects of entity *B* with respect to the final effect *C*, despite identical learning input in the second causal learning phase (see Methods section for specific predictions). Alternatively learners may ignore the initially acquired conceptual scheme and induce a novel, optimally predictive set of categories in the second learning phase. In the latter case no difference between conditions should be obtained with respect to the causal power of objects of *B*.

In sum, according to the unbroken link hypothesis, we predicted to observe category transfer in Experiment 1a, but not in Experiment 1b. Crucially, this should be the case, even

though the categorization of the intermediate entity *B* is based on exactly the same visible features of *B* in both studies.

2.1 Experiment 1a

2.1.1 Method

2.1.1.1 Participants and Design

Forty students from the University of Göttingen, Germany, participated for course credit. Participants were randomly assigned to one of two causal categorization conditions (color vs. size). We manipulated the category structure entailed by the first causal relation (*A*→*B*) between conditions. The second causal learning (*B*→*C*) phase was identical for all participants.

2.1.1.2 Materials

Participants received instructions about a causal chain relating an initial binary cause event (*A*; genetic mutation vs. no mutation) to complex biological molecules (*B*; uncategorized intermediate entity), which in turn were related to a final binary effect (*C*; cell death vs. no cell death).

The intermediate, uncategorized entity involved the "molecules" depicted in Figure 2. Depending on the experimental condition, the visible feature affected by the initial cause event (genetic mutation) was either the molecules' color (mostly yellow vs. mostly orange; in Figure 2 depicted as grey and white) or the molecules' size (predominantly large atoms vs. predominantly small atoms). Henceforth we will refer to these four basic exemplar types as item type 11 (white and large molecules), type 10 (white and small molecules), type 01 (grey and large molecules), and type 00 (grey and small molecules) (cf. Figure 2). In addition, there were three features that were not relevant for the two causal relations: presence or absence of a strong bond, presence or absence of a circular sub-structure, and presence or absence of a v-shaped sub-structure. For

example, the molecule shown in the upper left corner of Figure 2 consists of predominantly large and white atoms, has a strong bond, no circular substructure, and no v-structure. By contrast, the molecule at the right bottom corner consists of predominantly small and grey atoms, has no strong bond, a circular substructure on the left-hand side and a v-structure on the right-hand side of the molecule. Thirty-two different molecules were constructed by factorial combination of these five binary features. Additional five variants of each of these 32 exemplars were created by switching the positions of the atoms within the molecules (resulting in 160 items in total). This item space allowed us to use perceptually different items for the learning and test phases.

Insert Figure 2 about here

Figure 2 also outlines the statistical structure between the chain's initial cause event and the molecules' features. In the *color condition*, the initial dichotomous cause event (mutated vs. normal gene) was related to the color of the exemplars of the intermediate entity (number of grey vs. white molecules). This relationship was deterministic (the probabilities shown in Figure 2 refer to the second causal learning phase). Molecules generated by mutated genes had more white than grey atoms (exemplars 11 and 10), whereas molecules generated by normal genes had more grey than white atoms (exemplars 01 and 00). In the s*ize condition*, the type of gene was deterministically related to whether the molecule consisted of predominantly small or large atoms. Molecules generated through the mutated gene consisted of more large than small atoms (exemplars 11 and 01), whereas the opposite was true of molecules caused by a normal gene (exemplars 10 and 00). All other features were statistically unrelated to the initial cause event. The assignment of features (size and color) to the cause events (normal vs. mutated gene) was counterbalanced across participants.

Figure 2 also provides information about the statistical structure of the second causal relation (molecules → cell death). These learning data were identical regardless of condition. Molecules of type 11 (white + large) always caused cell death and molecules of type 00 (grey + small) never caused cell death. The remaining molecule types, exemplars 10 (white + small) and 01 (grey + large), had a probability of 50% to cause cell death. The assignment of probabilities to feature combinations was counterbalanced across participants.

2.1.1.3 Procedure

The experiment consisted of two consecutive causal learning phases, each followed by a test phase. In the first causal learning phase (Phase 1), participants' task was to learn about the relation between the initial dichotomous cause (normal vs. mutated gene) and the uncategorized exemplars of the intermediate entity (the molecules depicted in Figure 2). The second causal learning phase (Phase 2) concerned the causal impact of these molecules on the destruction of living cells (binary final event). Whereas in Phase 1 the category structure entailed by the initial cause *A* was manipulated across subjects (color vs. size), Phase 2 learning was identical in all conditions.

In the initial instructions prior to Phase 1 participants were told that biologists had investigated the causal relation between genes (*A*) and the formation of certain biological molecules (*B*). It was pointed out that every result was possible: the genes might deterministically or probabilistically cause the generation of specific molecules, or have no influence at all on their formation. Then participants were presented with 64 index cards (two randomized blocks of the 32 basic exemplars) in a trial-by-trial learning procedure. Information about the state of the cause event (normal vs. mutated gene) was given first, followed by a picture of the resulting molecule (intermediate event) on the back of the index card. Across

conditions the category structure (color vs. size) was manipulated. This initial learning phase was followed by a test phase, in which participants were presented with four new molecules of each type (11, 10, 01, and 00), resulting in a total of 16 test trials. For each molecule participants were requested to estimate how likely this exemplar had been generated by the mutated gene. The rating scale ranged from 0 ("molecule was definitely not generated by the mutated gene") to 100 ("molecule was definitely generated by the mutated gene"). No feedback was provided.

Prior to the second learning phase (Phase 2), which was identical for all participants, the instructions stated that a team of other researchers had investigated the causal *effects* of the molecules and wondered whether they caused cell death. Again it was pointed out that every result was possible (i.e., the molecules might deterministically or probabilistically cause cell death, or have no influence at all). The learning and the test procedures were similar to the first learning phase. Participants were first presented with a molecule on the front side of an index card and then were informed on the back side whether it had caused cell death. Participants were presented with 64 different molecules (two randomized sets of the 32 basic exemplars). In the subsequent second test phase participants were shown 16 previously unobserved molecules (four of each type) and asked to estimate the probability that the shown molecule would cause cell death. The rating scale ranged from 0 ("never causes cell death") to 100 ("always causes cell death"). No feedback was provided.

2.1.1.4 Transfer Tests

The critical test exemplars for assessing whether participants transferred the category scheme entailed by the first causal relation ($A \rightarrow B$) to the second causal learning phase ($B \rightarrow C$) were exemplars of type 10 and 01, which in the second phase had a probability of 50% to cause cell death (see Figure 2). Participants in the color condition had previously learned that molecules of

type 10 (white + small molecules) were generated by the same cause event as molecules of type 11 (white + large molecules), which in the subsequent learning phase always caused cell death. Thus, if participants induced the category "white molecules caused by a mutated gene" (exemplars 10 and 11) based on the causal input of the initial learning episode, they should infer that members of this category would cause cell death with probability of 75%. Conversely, molecules of type 01 (grey + large molecules) were generated by the same cause as molecules of type 00 (grey + small molecules), which never caused cell death. Thus, if participants induced the category "grey molecules caused by a normal gene", they should infer that cell death will be caused by molecules of this category with an average likelihood of 25%. In sum, if learners transferred the color-based category scheme of the initial learning phase, they should give high probability estimates for molecule exemplars of type 10 and low estimates for exemplar type 01, although the data in the second learning phase indicated that these two molecule types generated the effect with a probability of 50%.

The opposite pattern of judgments should arise if people had previously learned to classify the molecules according to their size. In this condition, learners had observed that all small molecules were generated by the same cause event (i.e., items 10 and 00 belonged together) and that all large molecules had the same cause (i.e., exemplars 01 and 11 belonged together). In the second phase large molecules had a probability of 75% and small ones had a probability of 25% to cause cell death. If learners transferred these size-based categories, they should give low estimates for exemplars of type 10 (white + small) and high estimates for molecules of type 01 (grey + large).

A different pattern of causal judgments should arise if participants did not transfer the initially acquired categories to the subsequent learning episode. If they preferred to induce a

novel category scheme based on the contingencies observed in the second causal learning, they should rate the probability of both types of critical test molecules (types 01 and 10, respectively) to cause cell death at a level of 50%.

Whereas transfer of the categories acquired in the first causal learning phase implies diverging judgments for test items of type 10 and 01, no difference is predicted for test items 11 and 00. Regardless of the feature constituting the category boundary in the initial causal learning episode, molecules of type 11 always belonged to the category strongly related to cell death (*P*(cell death) = .75), whereas exemplars of type 00 always were members of the category weakly associated with the effect in the second learning phase (*P*(cell death) = .25) (see Figure 2). Therefore, participants should tend to give high ratings for large and white molecules (item type 11) and low ratings for small and grey molecules (item type 00), regardless of condition.

2.1.2 Results and Discussion

The mean probability ratings for the two conditions (color vs. size) are shown in the upper half of Table 1. To analyze participants' judgments, we conducted a number of planned pair-wise comparisons. We first analyzed the ratings of the test exemplars obtained after the first learning phase (genes → molecules). Within both categorization conditions there were large differences between estimates for molecules generated by the mutated and normal types of genes (color condition: $t(19) = 14.9$, $p < .01$; size condition: $t(19) = 17.3$, $p < .01$).[3] There were no reliable differences between conditions, indicating that in both conditions participants successfully induced cause-based categories in the first learning phase.

Insert Table 1 about here

Table 1 also displays the results for the second causal relation (molecules→cell death). The data show that learners' judgments of the causal efficacy of the molecules causing cell death with a probability of 50% (exemplars of type 10 and 01) differed strongly between conditions (upper half of Table 1; item 01: 44.1 vs. 60.6, and item 10: 58.9 vs. 33.4). Crucially, the mean ratings of these exemplars were reversed in the two conditions, resulting in a significant interaction contrast, $t(38) = 2.97$, $p < .01$. This result indicates that learners' causal judgments concerning the second causal relation were systematically influenced by the previously acquired causal category scheme.

While transfer of categories entails that learners' ratings for the intermediate items should differ between conditions, no such effect is expected for exemplars that always or never caused cell death (items 11 and 00). In line with this prediction, these molecules received very similar ratings in both conditions. A significant difference between molecules that always or never caused cell death was obtained within both conditions (color condition: $t(38) = 4.96$, $p > .01$, size condition: $t(38) = 5.37$, $p < .05$), but, as expected, there was no significant interaction between conditions ($t(38) = .04$, $p = .97$).

Taken together, the observed pattern of causal judgments indicates that participants made use of the categories induced in the first causal learning phase when learning about the second causal relation instead of inducing novel categories based on the contingency information available in the second learning episode. This finding is predicted by the unbroken mechanism hypothesis.

2.2 Experiment 1b

In Experiment 1b we used a nearly identical scenario as in Experiment 1a. The only difference was that in Experiment 1b we used atmospheric pressure as the initial cause of the chain instead

of genetic mutation. Pressure is more likely to be viewed as a cause that merely affects superficial features of the chain's intermediate entity (molecules), whereas cell death (the chain's final event) is more likely to be considered an effect of the molecules' essential, hidden features. Based on these assumptions there is no continuous causal mechanism linking the initial and final event of the causal chain (cf. Figure 1, left-hand side). Hence, the unbroken mechanism hypothesis predicts that there will be no transfer of categories.

2.2.1 Method

2.2.1.1 Participants and Design

Forty-four students from the University of Göttingen, Germany, participated for course credit. They were randomly assigned to one of two conditions. Again we varied which feature (color vs. size) of the object involved in the intermediate event was causally affected by the initial cause.

2.2.1.2 Materials and Procedure

The stimuli and experimental procedure were almost identical to the ones in Experiment 1a, including all counterbalancing. The only difference was that we used a different initial cause in the first learning phase. Prior to the first learning phase participants were informed that some researchers hypothesized that atmospheric pressure might have a causal influence on the molecules. Therefore they had conducted an experiment in which the formation of molecules was studied under high and low pressure. As in Experiment 1a, it was pointed out that any outcome was possible (i.e., a deterministic or probabilistic causal relation, or no relation at all).

Except for the different cover story, the procedure was identical to Experiment 1a. Participants received 64 learning trials in which they learned about the causal relation between atmospheric pressure and molecules. In the *color condition* high pressure always caused white molecules (exemplar types 11 and 10) and low pressure generated grey molecules (exemplars 01

and 00). In the *size condition* participants observed that high pressure always caused large molecules (exemplars 11 and 01) whereas low pressure generated small molecules (exemplars 10 and 00). In the subsequent test phase participants were asked to rate the probability that a particular molecule was generated by high or low pressure. The same learning and test items as in Experiment 1a were used. The second learning and test phase were completely identical to the ones in Experiment 1a. Participants were again presented with molecules and had to estimate their likelihood of causing cell death.

2.2.2 Results and Discussion

The means displayed in the lower half of Table 1 show that participants induced categories in the first learning phase (pressure → molecules). Reliable differences were obtained between the induced cause-based categories in both conditions (color condition: $t(21) = 12.97$, $p < .01$, size condition: $t(21) = 9.87$, $p < .01$). No differences between conditions were obtained, indicating that all participants induced categories according to the initial cause event, namely atmospheric pressure.

Next we analyzed participants' judgments of causal efficacy obtained after the second learning phase (molecules → cell death). The results show that learners' causal judgments of the critical items (molecule types 10 and 01) did not differ between conditions (Table 1, lower half, right hand side). Contrary to Experiment 1a the interaction contrast between conditions was not significant ($t(42) = .03$, $p = .97$). In fact, participants estimated the probability of cell death to be around 50%, regardless of condition. These findings indicate that the initially induced category scheme was not transferred to the subsequent learning phase. The judgments obtained after the first causal learning phase clearly reveal that learners had no problems with inducing the cause-

based category scheme. Thus, the lack of category transfer was not due to a learning deficit in the initial causal learning phase.

The findings for molecules of type 11 and 00 replicate the results of Experiment 1a. There was a difference between these items within each condition (color condition: $t(21) = 4.19$, $p < .01$; size condition: $t(21) = 6.73$, $p < .01$), but as before there was no significant difference between conditions (interaction contrast: $t(42) = .47$, $p = .64$). This result further supports the claim that the failure of obtaining transfer effects cannot be traced back to a general lack of learning. Rather, in the second learning phase participants preferred to induce a new, tripartite category scheme which mirrored the observed statistical contingencies.

2.3 Discussion Experiment 1

Taken together, the observed category transfer in Experiment 1a and the lack of transfer in Experiment 1b are consistent with the unbroken mechanism hypothesis. Category transfer was only observed when there was an unbroken causal mechanism relating the three events in the causal chain. Although an essence played a causal role in Experiment 1b, we obtained no transfer effect. However, one may defend essentialism arguing that only categories being based on an essence in the first place may be transferred. Therefore we ran another set of studies in which learners were presented with causal scenarios for which the two accounts make diverging predictions.

3. Experiment 2

The previous studies showed that category transfer depended on the existence of an unbroken causal mechanism linking the events in the causal chain. An alternative explanation, however, might be that category transfer additionally depends on the involvement of hidden (essential) features of the intermediate event. The goal of the second study was to rule out this explanation

by directly manipulating learners' assumptions about the causal mechanisms in the chain. We ran four studies with varying instructions but otherwise identical procedures and materials. Therefore, we will present them together.

Figure 1 (right column) shows the different variants of causal chains investigated in Experiments 2a-d. *Causal Model I* represents an unbroken mechanism in which the chain's entities are connected via hidden essential features of the intermediate entity. Like in Experiment 1a, we expected transfer in this condition. *Causal Model II* embodies a different kind of causal chain with an unbroken mechanism. The initial cause of the chain affects the intermediate entity's hidden features, which generate the observable features of the respective exemplars. The surface features, in turn, are causally responsible for the final effect. Since there is a continuous mechanism relating the causally relevant features to each other, the unbroken mechanism hypothesis predicts category transfer in this scenario too. By contrast, in *Causal Model III* the two causal relations of the chain are disconnected. While the initial cause affects the intermediate entity's essential features, the second causal relation originates in the surface features. Since the surface features already exist prior to the causal modification of the hidden properties and are not causally linked to these properties, the mechanism that links the initial and the final event of the causal chain is broken. Therefore, no category transfer should be obtained according to the unbroken mechanism hypothesis. Note that this is a different type of broken mechanism than the one examined in Experiment 1b, in which the first causal arrow pointed towards the surface features and the second link originated in the (essential) hidden features (cf. Figure 1, left column). The predictions of psychological essentialism are unclear here. It may still predict transfer because essentialist categories were established in the first learning phase. However, in this causal model these categories are described as causally irrelevant for causal relation in the

second learning phase. Therefore an alternative view might be that psychological essentialism does not predict transfer (e.g., Waldmann & Hagmayer, 2006). Finally, *Causal Model IV* presents a further variant of an unbroken mechanism in which no essence is causally involved at all. The initial cause affects the surface features of the intermediate entity, which in turn are directly causally responsible for the final effect. Note that in this case no essential property of the mediating entity is involved. Hence, no categories based on essential properties can be established. Therefore no transfer of categories should be observed according to essentialism. If the connectedness of the assumed causal mechanisms is crucial, however, category transfer should result.

In summary, the unbroken mechanism hypothesis predicts category transfer when participants assume Causal Models I, II, and IV, since in these three causal models the initial cause and final effect of the causal chain are linked by an unbroken causal mechanism. By contrast, since there is no unbroken mechanism in Causal Model III, learners should tend to abandon the initial categories and induce a novel category scheme when learning about the second causal relation. The alternative theoretical account, psychological essentialism, predicts transfer of categories for Causal Models I, II, and possibly Causal Model III, since in all these scenarios the initial event causally affects the essential features of the intermediate entity. By contrast, no transfer is predicted for Causal Model IV, since here only the surface properties of the intermediate entity are involved in the causal relations.

3.1. Method

3.1.1 Participants and Design

Twenty students from the University of Göttingen participated in each of the four studies (Exp. 2a-d) for course credit (i.e., $N = 80$). Each of the experiments provided learners with different

instructions about the causal mechanism underlying a causal chain $A \rightarrow B \rightarrow C$. Within each experiment, participants were randomly assigned to one of two counterbalancing conditions (A vs. B). These conditions were defined by the category boundary entailed by the causal relation between the chain's initial and intermediate entity. The second causal learning phase was identical in all conditions and experiments.

3.1.2 Materials

Participants in all four experiments were presented with a causal chain connecting the presence of alpha and beta microbes (initial binary cause) to the properties of protozoa (set of objects constituting the intermediate entity), some of which caused an inflammation of the spleen (binary final effect).

The materials used in Experiments 2a-d are depicted in Figure 3. The protozoa (uncategorized intermediate event) were constructed from factorially combining four binary features: Number of corners (pentagonal vs. octagonal), shape (rounded vs. elongated), number of surface molecules (two vs. four), and shape of surface molecules (round vs. squared). Prototype 1 (item 1111) was octagonal and rounded, and had four squared surface molecules. Prototype 2 (item 0000) was pentagonal and elongated, and had two round surface molecules (Figure 3).

Insert Figure 3 about here

In contrast to Experiments 1a and 1b (in which the initial cause of the chain was causally related to a single feature of the uncategorized intermediate event) we here used a family resemblance structure for the first causal relation (microbes → protozoa). Thus, the chain's initial cause (alpha vs. beta microbes) was probabilistically related to all features of the

intermediate entity (protozoa). We used this more complex category structure to broaden the evidence for the unbroken mechanism hypothesis.

Within each experiment we counterbalanced the category boundary (condition A vs. B) entailed by the first causal relation (Figure 3). In Condition A, alpha microbes caused Prototype 1 (item 1111) and the protozoa that had at least three features in common with this prototype (i.e., items 1110, 1101, 1011, and 0111). The remaining protozoa were caused by beta microbes. In Condition B, the category boundary was shifted. Here, beta microbes caused Prototype 2 (item 0000) and the protozoa that had three features in common with this item (i.e., items 0001, 0010, 0100, and 1000). The remaining exemplars were caused by alpha microbes (cf. Figure 3). Thus, depending on the condition, the intermediate exemplars sharing two features with each of the two prototypes either belonged to the same category as Prototype 1 (1111) or Prototype 2 (0000).

The statistical structure entailed by the second causal relation (i.e., protozoa→inflammation) is shown in Figure 4. Prototype 1 (item 1111) and protozoa having three features in common with this prototype never caused an inflammation, while Prototype 2 (item 0000) and protozoa having three features in common with this prototype always caused inflammation (the assignment of exemplars and causal effects were counterbalanced across participants). Protozoa that were equally similar to both prototypes (i.e., items sharing two of four features with each of the prototypes) caused the effect with a probability of 50%. Thus, the observed contingencies of the second causal relation actually entail a tripartite category structure (Figure 4).

Insert Figure 4 about here

3.1.3 Procedure

The materials and learning data were identical in all four experiments. Participants' assumptions regarding the causal mechanisms that related the events in the causal chain were manipulated. In *Experiment 2a* (Figure 5, Causal Model I) participants were told that alpha and beta microbes infiltrate the protozoa's nucleus and change their genetic make-up by integrating their microbial genes into the protozoa's genome. The modified genome, in turn, affects the body structure of the protozoa. Later on the protozoa's genes enter the spleen cells and cause an inflammation. Thus, this is an example of an unbroken mechanism. *Experiment 2b* (Figure 5, Causal Model II) presented participants with a slightly different mechanism, which also represents an unbroken mechanism. Again the microbial genes causally influence the protozoa's genetic make-up, which affects the protozoa's surface features. In contrast to Experiment 2a, the inflammation of the spleen is now caused by these surface features. Participants were told that the body structure of the protozoa may attach itself to receptors of spleen cells, thereby blocking these receptors and causing an inflammation. In *Experiment 2c* (Figure 5, Causal Model III), the instructed causal mechanism was broken. Like in Experiment 2b, the microbes causally influence the genetic make-up of the protozoa, and the protozoa's surface features are causally responsible for an inflammation of the spleen. In contrast to Experiment 2b participants were told that the changes in the protozoa's genome caused by the microbes do not influence the surface structure of the protozoa. Thus, the causal mechanism between essentialist and surface features was broken. In *Experiment 2d* (Figure 5, Causal Model IV) the microbes directly affected the surface features of the protozoa without affecting their genome. Participants were told that the microbes attach to the outside of the protozoa and modify their body structure through biochemical processes. It was explicitly pointed out that the protozoa and their genetic make-up were not affected by the

changes of the cells' surface. Like in Experiments 2b and 2c participants were also instructed that the surface features cause the inflammation. This causal model represents an example of an unbroken mechanism in which essential features of the intermediate entity are not involved.

Insert Figure 5 about here

Experiments 2a-d comprised two consecutive causal learning phases. Participants were requested to learn the causal relationship between the two events in each phase (microbes → protozoa, and protozoa → inflammation, respectively). In the first phase participants learned the causal relation between microbes (alpha vs. beta) and uncategorized protozoa (Figure 3). Like in Experiments 1a,b it was then investigated whether this cause-based categorization of the protozoa would be transferred to the second causal learning phase (protozoa → inflammation). The only difference between the four experiments concerned the provided information regarding the causal mechanism underlying the chain. Again it was explicitly pointed out that any strength of the relations including no causal influence was possible. Prior to each learning phase the causal mechanism underlying the relation was verbally described through instructions and graphically illustrated as shown in Figure 5.

Before learning about the first causal relation (microbes → protozoa), participants were presented with a diagram similar to Figure 3 (without the numbers), showing which protozoa were affected by the two types of microbes. Participants were asked to memorize the relations. Studying this figure prior to the trial-by-trial learning procedure simplifies and speeds up learning. We introduced this step to compensate for the fact that a family resemblance structure is more complicated than the one-dimensional category structure used in Experiments 1a, b. The

diagrams varied across the two conditions A and B (see Figure 3). The assignment of microbe type (alpha vs. beta) to protozoa was counterbalanced across participants.

Next participants were shown index cards in a trial-by-trial learning fashion. Each card provided information about the presence or absence of the chain's initial cause (alpha or beta microbes) on the front side and a corresponding protozoon on the back side. To ensure that participants saw the same number of trials in both causal learning phases two of the 16 exemplars were not shown during learning (exemplars 1010 and 0101; see below for the rationale behind excluding these items in the second learning phase). Each participant received five randomized blocks of 14 items each (i.e., 70 trials in total). Upon completion of the first causal learning phase participants were shown 10 of the 16 protozoa in randomized order, including all six exemplars which had two features in common with each of the two prototypes. For each exemplar participants were asked to judge whether this protozoon was causally affected by alpha or beta microbes. Estimates were given on a scale ranging from 0 ("the protozoon was definitely affected by alpha microbes") to 100 ("the protozoon was definitely affected by beta microbes"). No feedback was provided.

Prior to the second causal learning phase (protozoa → inflammation) participants were informed about this relation and the respective underlying causal mechanism. Again a graphical representation of the mechanism was provided (see Figure 5, right hand side). Then the learning data were presented in a trial-by-trial fashion on index cards. On each trial, participants were first shown a protozoon and then, on the back side, whether an inflammation had occurred or not. Two of the 16 items (exemplars 1010 and 0101) were not shown during the second learning phase to ensure that all individual features were equally predictive for inflammation. The removal of these two exemplars also allowed us to investigate whether participants relied on

previous categories regardless of whether they encountered a specific exemplar during learning. In total, five randomized blocks of the 14 exemplars were presented.

Subsequently, participants were asked to rate the probability of different protozoa causing inflammation. The scale ranged from 0 ("the protozoon never causes inflammation") to 100 ("the protozoon always causes inflammation"). Ten protozoa were presented in randomized order: the two prototypes, one exemplar sharing three features with Prototype 1, one exemplar sharing three features with Prototype 2, and all six items sharing two features with each prototype (i.e., items being equally similar to both prototypes). Four of these six microbes had been shown in the second causal learning phase, and two had not been shown. No feedback was given.

3.1.4 Transfer Tests

To test for category transfer we focused on the items that were equally similar to both prototypes which actually caused inflammation with a probability of 50% (henceforth denoted as *critical items*). The family resemblance structure allowed us to use a sensitive and powerful within-subjects test of transfer. We compared the ratings for the critical items to the ratings for the two prototypes of the categories entailed by the first causal relation. If the categories entailed by the first causal relation were transferred to the second learning phase, then the ratings for the critical items should be more similar to the ratings of the prototype of the cause-based category they belonged to (*within-category prototype*) than to the prototype of the other category (*between-category prototype*). Thus, we expect that the critical items are assimilated to the category to which they were assigned to based on the causal learning input in the first learning episode. Assume, for example, that participants had learned in the first phase that exemplars which are equally similar to both prototypes are generated by the same type of microbe as Prototype 1. Now consider exemplar 1001 in Figure 5: Based on the previously learned categorization this

test exemplar and its within-category prototype (Prototype 1) should receive similar ratings regarding their causal capacity to generate the final causal effect. By contrast, the causal judgment regarding this test exemplar should clearly differ from those for Prototype 2, the prototype of the other category. To measure this assimilation effect we computed the differences between the ratings of the critical items and the ratings of the within and between-category prototypes. If participants transferred the categories implied by the first causal relation, these differences should be smaller for the within-category prototype than for the between-category prototype. By contrast, if participants induced a new category structure in the second causal learning phase and ignored the previous categories, ratings for the critical items should be roughly at equal distance from the ones obtained for the two prototypes.

Insert Table 2 about here

3.2 Results and Discussion

We first analyzed participants' estimates regarding the categories entailed by the first causal relation (Table 2, left hand side). A significant difference between protozoa generated by alpha and beta microbes, respectively, was obtained in all studies (Exp. 2a: $t(19) = 7.05$, Exp. 2b: $t(19) = 11.85$, Exp. 2c: $t(19) = 5.71$, Exp. 2d: $t(19) = 15.42$, all $p < .01$). Thus, all learners correctly encoded the relation between the first and the intermediate event of the causal chain by inducing cause-based categories.

Did participants transfer the category scheme entailed by the first causal relation? Table 2 (right hand side) shows the estimates obtained after the second learning episode. The distance of the critical test items from the two prototypes was computed by recoding and pooling the data over the two counterbalancing conditions A and B. An inspection of the data reveals that in

Experiments 2a, 2b, and 2d the distance of the critical test items to the within-category prototype was smaller than the distance to the between-category prototype (see the right column of Figure 1 for the corresponding Causal Models I, II, and IV). Contrary to that no significant difference resulted in Experiment 2c.

Since the four studies were identical except for the manipulation regarding learners' assumptions about the causal mechanisms, we first conducted a cross-experimental comparison. An analysis of variance with the mean differences of the six critical test items from the two prototypes as a within-subjects factor and causal model (I – IV, cf. Figure 1) as a between-subjects factor yielded a main effect of prototype [$F(3, 76) = 2.99$, $p < .05$, $MSE = 676.2$] and a main effect of causal model [$F(1, 76) = 18.5$, $p < .001$, $MSE = 603.2$]. To specifically test the prediction that there would be no difference in Experiment 2c, but differences in Experiment 2a, 2b, and 2d, we computed a weighted interaction contrast with causal model as a between-subjects factor and the distance to the two prototypes as a within-subjects factor. This interaction was significant [$F(1, 76) = 4.78$, $p < .05$, $MSE = 603.2$]. These results support the unbroken mechanism hypothesis.

We next analyzed the four studies separately. The crucial analyses concern learners' causal judgments of the critical exemplars. We separately analyzed participants' judgments for the four critical test items presented in the second learning phase and the two exemplars that were not shown. The data shown in Table 2 indicate that the categories entailed by the first causal relation strongly affected learners judgments regarding the second causal relation in Experiments 2a, 2b and 2d, but not in Experiment 2c. In the former experiments, ratings for the critical test exemplars deviated significantly less from the within-category prototype than from the prototype of the contrasting category (Table 2). Particularly the judgments obtained in studies

2c and 2d provide strong support for the unbroken mechanism hypothesis. Psychological essentialism predicts no transfer in Experiment 2d since no essential features are involved, and possibly transfer in Experiment 2c as the first causal relation affects essential features. The unbroken mechanism hypothesis makes the opposite prediction, which is in line with participants' judgments.

The results also indicate small differences between the causal judgments for items which were actually presented in Phase 2 versus the ones being omitted. Generally, category transfer seemed to be stronger for exemplars not previously shown. Memory for the observed exemplars which had a probability of .5 to generate the effect in Phase 2 may have interfered with the category assimilation effect, thus slightly weakening the category transfer effect (see also Waldmann & Hagmayer, 2006; Waldmann et al., 2010).

Taken together, Experiment 2 provides further evidence for the unbroken mechanism hypothesis. Transfer of categories depended on the assumption of an unbroken causal mechanism, regardless of whether essential features of the intermediate entity were involved or not.

4. General Discussion

Our main goal in the present set of studies was to investigate how causal induction and categorization interact in sequential causal learning. We presented participants with tasks in which they separately learned about two relations of a causal chain which overlapped in an intermediate entity. Whereas the initial and final events were pre-categorized binary causal events or entities, the intermediate entity in the chain consisted of a set of uncategorized objects taking part in both causal relations. Thus, along with learning about the two causal relations, participants also were confronted with the task of inducing categories for the objects constituting

the intermediate entity. Our main goal was to investigate under which conditions people transfer categories from one causal relation to the other, and when they tend to abandon the initially induced categories.

Our studies provide a number of novel findings. First, learners used the binary initial *cause* to categorize the intermediate entity (see also Waldmann et al., 2010). This result extends previous research (Lien & Cheng, 2000; Kemp et al., 2010; Marsh & Ahn, 2009) which has shown that exemplars may be categorized according to their *effects*. The most important novel contribution of this paper is to flesh out the conditions under which people transfer categories between two causal relations that overlap in one entity. We proposed the *unbroken mechanism hypothesis*, which claims that category transfer within causal chains depends on people's assumptions about the causal mechanisms underlying the observed contingencies between categories of events or entities. Whenever learners assume an unbroken mechanism linking the three entity types *A*, *B*, and *C* of a causal chain, transfer is predicted. An unbroken mechanism in a chain exists when the same features of the middle entity *B* are causally relevant for the two causal relations $A \rightarrow B$ and $B \rightarrow C$. Should different features of *B* be relevant in the two causal relations, then the mechanism is only unbroken if these features within category *B* are causally linked (e.g., the relation between DNA and surface features) so that the effect category of the first relation coincides with the cause category of the second relation. Otherwise the mechanism is broken. In two sets of experiments we found support for our hypothesis.

4.1 Relations to Previous Research

4.1.1 Causally Motivated Category Induction

As pointed out in the Introduction, our research is an extension of previous studies which demonstrated that causal power may motivate categorization. Lien and Cheng (2000) and Marsh

and Ahn (2009) have shown that people tend to induce feature-based categories that maximize causal predictiveness. Kemp et al. (2010) have extended this research by showing that learners use causal power even when no other category features are available, or combine feature and causal power information when both types of information are available. While categorization was studied in the context of single causal relations in these experiments, our research went one step further by addressing the question what categories learners would induce when the category exemplars are presented in the context of multiple causal relations (i.e., causal chains; common-cause models) (see also Waldmann et al., 2010). Whereas optimal predictiveness of categories can be easily defined with respect to a single causal relation, multiple causal relations may lead to optimization problems: categories that are optimal with respect to one relation may be suboptimal with respect to the other. Thus, learners may see themselves in a situation in which they may choose to generate categories that simultaneously optimize different relations, or they may opt for maximal predictiveness with respect to one relation at the cost of the other.

We have investigated causal categorization the context of a sequential learning task in which the different relations are presented in consecutive learning phases. Assuming that people tend to induce optimal categories with respect to the initially learned causal relation, our research question was whether they would stick to these categories in a subsequent causal learning episode or abandon them and induce a new conceptual scheme.

Although restricted to single causal relations, Kemp et al. (2010) were also interested in the conditions in which learners maintain or abandon previously learned categories. They have shown that learners tend to assign exemplars to new categories when the apparent causal power of the new test exemplars does not seem to cohere with the previously induced categories. Thus, Kemp and colleagues have provided evidence for the role of bottom-up statistical information in

the decision about whether categories are transferred to new exemplars or not. Our research extends these findings by showing that top-down assumptions about underlying unobserved mechanisms may play an important role in category transfer between learning episodes. Whenever learners assume an unbroken mechanism linking two causal relations within the target category, a tendency to maintain the same category scheme across both relations was observed.

What are participants doing when the mechanism is broken? According to our hypothesis, learners should tend to view the initially induced categories as causally irrelevant for the second relation and invoke a new conceptual scheme that is statistically better suited to predict the final effect. In general, transfer of categories has the advantage of avoiding proliferations of category schemes, but the potential disadvantage is that we may be stuck with classification schemes that are not optimally predictive and do not make theoretical sense. Apparently people's beliefs about causal mechanisms guide the decision between continuing to use an existing classification scheme and inducing a novel set of categories.

4.1.2 Theory-Based Categories and Psychological Essentialism

Research on theory-based categorization also addresses the relation between categories and causality, but has a different focus from the research on causally motivated category induction. Whereas the latter field is interested in how people induce categories based on the causal relations, in which these categories are embedded, research on theory-based categorization primarily focuses on the internal causal structure of categories. According to this view many categories contain features that are causally and functionally interrelated. Thus, the categories can be represented as intuitive causal theories. It has been shown that the causal models underlying individual categories predict learning (Waldmann, Holyoak, & Fratianne, 2005), typicality judgments (Rehder & Kim, 2006), and inductive inferences about features (Rehder,

2009). Our research builds a bridge between these two separate areas. We have shown that the maintenance and transfer of causally induced categories does not only depend on the statistical relations the categories are embedded in, but also on top-down assumptions about the hidden, unobserved causal structure underlying the induced categories. Only when the assumptions about the underlying causal model are consistent with an unbroken mechanism, category transfer was observed.

Research on theory-based categorization has also been interested in modeling different types of categories, such as natural kinds or artifacts. Medin and Ortony (1989) have argued that we have a tendency to assume a hidden essence underlying the visible, variable features for natural kinds (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, 2007; Rehder & Kim, 2006). In our experiments we have provided learners with information about causal models that contain hidden, essential (e.g., DNA) and superficial, appearance-related features.

One novel focus of our research was to study the role of essential features in category transfer. In our previous research (Waldmann & Hagmayer, 2006) we have proposed that the involvement of hidden, essential features in multiple causal relations is a necessary precondition of transfer. This hypothesis was motivated by the idea that essences may be viewed as stable ontological givens that provide stability to the categories regardless of the causal powers generated by these categories. In contrast to this hypothesis, Experiment 2c showed that the involvement of hidden or essential features is not sufficient for transfer of categories entailed by a causal relation. Transfer was only observed when the essentialist features were part of an

unbroken mechanism linking the events in the causal chain; in Experiment 2c essentialist features were only causally relevant in one of the two causal relations.

The lack of transfer in this experiment could be interpreted as being consistent with psychological essentialism if the assumption was added that transfer is only observed when the essentialist features are causally relevant for both causal relations (e.g., Waldmann & Hagmayer, 2006). However, this is also not the full story. We observed category transfer when the initial cause only affected surface features of the category members of the intermediate entity, and when these surface features were causally relevant for the final effect (Experiment 2d). This finding indicates that the involvement of essentialist features is not a necessary condition for category transfer, but that people's causal beliefs about an unbroken mechanism drive category transfer.

The above mentioned modified variant of psychological essentialism, which claims that transfer only occurs when essentialist features are causally relevant in both causal relations, certainly describes an important, maybe the most important case of transfer. The involvement of an unobserved cause place holder (i.e., an essence) in several causal relations allows learners to flexibly link this cause to various features, which makes category transfer plausible in many instances. Thus, we believe that transfer is generally likely to be observed with natural kind concepts, unless there are more specific instructions breaking up the underlying mechanism. However, the present studies have shown that the driving factor is not the involvement of essentialist features but rather the assumption of an unbroken mechanism. It is certainly hard to create cover stories that present an unbroken mechanism just using surface features of natural kind categories, but it is, as we have shown, possible.

4.1.3 Categories and Causal Mechanisms

Our unbroken mechanism hypothesis assumes that people have intuitions about underlying causal mechanisms. This raises the question how mechanism information is represented. In the literature, covariation-based theories and process theories of causality have often been pitted against each other (see Waldmann & Hagmayer, in press, for an overview). Whereas covariation-based theories assume that causal mechanisms exist between types of events, process theories typically focus on sequences of singular token events and assume that causal mechanisms exist on the token level. In a Michotte task, for example, process theorists assume that we directly perceive a singular event of causation in which a specific Object *A* exerts a force to which Object *B* resists to some extent (White, 2009, see also Wolff, 2007).

One interesting novel question concerns the relationship between categories and the representation of mechanisms on the type and token level. Covariation on the type level (i.e., between categories of the events or objects involved in the causal relation) can be viewed as an indicator of causal mechanisms on the token level (Pearl, 2000). On the other hand, observations of individual causal relations on the token level may allow us to induce categories and causal mechanisms on the type level (cf. Kemp et al., 2010). Thus the two levels of representation seem to be interdependent.

Our research shows that people often need the type or category level to decide whether a mechanism is broken. Experiment 3 of Waldmann and Hagmayer (2006) exemplifies a critical case. In this experiment, participants first learned to categorize viruses on the basis of surface features. A plausible representation of natural kinds, such as viruses, is that hidden properties cause the surface features so that these features can be used as diagnostic indicators of category membership (Rehder, 2007). In a subsequent learning phase we told participants that interior

designers used images of the viruses as blueprints for aesthetic patterns. These patterns (i.e., viruses) were used in a fictitious study in which research subjects were asked to judge whether they liked a specific (virus) pattern or not. In this experiment, no category transfer was observed because there was no reason to assume that virus types necessarily line up with categories of patterns people like or dislike. This result is in line with our unbroken mechanism hypothesis, which refers to the type level, because learners probably assumed that the categories entailed by the first causal relation are irrelevant with respect to the second relation. Thus, participants' assumptions about the relation of the two categories broke the mechanism and caused them to infer new categories in the second learning phase.

When observing individual tokens, however, learners may believe that there is a continuous chain leading from DNA (*a*) to surface features (*b*), which in turn affect the aesthetic judgment (*c*) in each case. On the token level it is impossible to see whether the causally relevant properties of instance *b* of entity *B* shift between the $a \rightarrow b$ and $b \rightarrow c$ relation. Learners can only observe the configuration of features of the individual tokens but cannot distinguish within each token between causally relevant and irrelevant features. To make this distinction, it is necessary to study the mechanism on the type or category level. Further research will have to explore such dissociations between the type and token level in more detail.

4.1.4 Modeling the Unbroken Mechanism Hypothesis

Recently hierarchical Bayesian models have been proposed as a computational model of category induction through causal learning (Kemp, Goodman, & Tenenbaum, 2007, 2010; Tenenbaum & Niyogi, 2003). These models use the observation of causal relations among individual exemplars of causes and effects and their respective features as evidence (event level). Hypotheses are formulated on two different levels of abstraction. On the level of causal models,

hypotheses about the causal relations between individual object types are represented. On a higher level (the level of causal schemata or causal laws, i.e., categories in our terminology), hypotheses about classes of objects, their causal relations and their features are represented. As a constraint it is assumed that objects belonging to the same class have similar causal powers and share similar features. Thereby the hypotheses about causal schemata constrain the hypotheses on causal models. Based on the available evidence, hypotheses are simultaneously revised on both levels using hierarchical Bayesian updating (see Kemp et al., 2010, for more details).

These models have been extremely powerful in explaining current research findings about category induction during causal learning. For example, the model by Kemp and colleagues (2007) is able to explain the findings of Lien and Cheng (2000) that people prefer to induce categories of causes that are maximally predictive for the observed effects. This model is also able to capture our finding that people transfer previously acquired categories to new causal learning episodes (Waldmann & Hagmayer, 2006) by assuming that hypotheses about these categories receive a higher a priori probability for future learning episodes.

However, the current version of the model does not use top-down knowledge that is needed for implementing our unbroken mechanism hypothesis. Categories in the Kemp et al. model are formed on the basis of bottom-up information about statistical relations, such as causal power, but they are not constrained by top-down knowledge about mechanisms. Nevertheless, it should be possible to extend hierarchical Bayesian models to capture our hypothesis. One way to achieve this goal would be to add a layer above the causal schema level incorporating mechanism knowledge which additionally constrains category formation (see Griffiths & Tenenbaum, 2009, for examples of how theory knowledge could be modeled within hierarchical Bayes nets).

4.2 Directions for Future Research

There are a number of interesting questions for future research. We have focused on causal chains because they provide the most plausible cases of unbroken und broken mechanisms. However, the notion of an unbroken mechanism can also be applied to other causal models. For example, in a common-cause model category transfer may be observed when different effects are generated by the same features of the category members representing the common cause (e.g., DNA). Essentialist common cause placeholders are a prime example. As the same essential features are assumed to generate the observable surface features and external effects of natural kinds, these categories are generally transferred to new learning episodes. However, if the effects are assumed to be generated by different causal mechanisms, people may induce alternative category schemes. Our previous study on the common-cause model (Waldmann et al., 2010, Exp.2) presented participants with two causal links originating in the same features (DNA). As expected, category transfer resulted. It would be interesting to conduct a study in which different aspects of the same common cause generated its different effects. According to the unbroken mechanism hypothesis, no transfer should be expected in this case.

Whereas causal chains and common-cause models can contain unbroken causal mechanisms linking separate causal relations, common-effect models in which multiple causes generate a joint effect represent a more complicated case. In common-effect models (e.g., $A \rightarrow C \leftarrow B$) two different mechanisms converge, hence there cannot be a continuous mechanism linking the three events. Thus, at first sight no transfer should be observed in common-effect models. However, a more general hypothesis derived from the unbroken mechanism hypothesis could claim that the *C* category in the common-effect example above will be transferred if both causes, *A* and *B*, affect the same *C* properties, and therefore generate the same effect categories.

If two independent causes *A* and *B* causally influence effect *C*, there is no reason to assume that *A* influences the same features of *C* as *B*. Thus, *A* will probably generate a different category of *C* events from *B*. For example, radiation may influence the DNA-sequence of viruses, but heat may destroy the viruses' proteins. Due to the competitive nature of independent causes, there is no a priori reason to expect that the categories of *C* generated by *A* are the same as the categories generated by *B*. Hence, in a sequential learning task presenting such a common-effect model, no transfer should be expected, analogous to our unbroken mechanism hypothesis. However, it may be possible to find examples which suggest that the same properties of *C* are affected by both *A* and *B*, for example, when the causes appear similar. For instance, it seems more likely that two food items (i.e., similar causes) cause a similar disease than a food item and radiation (i.e., dissimilar causes). The similarity of causes may not only be affected by superficial similarity but also by knowledge about underlying mechanisms. Both unprotected sex and contaminated needles cause AIDS through a common mechanism (e.g., a virus causing immune deficiency). Therefore, these two superficially dissimilar causes become similar, which makes it plausible that they entail the same category of disease.

In summary, studying the interplay between categorization and causal learning in non-stationary learning situations will provide a fruitful topic for future research and should help to overcome the segregation of different lines of research in the cognitive sciences. Both research on categorization and causality will profit from such studies.

References

Ahn W.-K., & Kalish,C. W. (2000). The role of mechanism belief in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 199-225). Boston, MA: MIT Press.

Ahn, W.-K., Kalish, C. W., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., Coley, D. J., & Shafto, P. (2001). Why essences are essential in the psychology of concepts. *Cognition, 82*, 59-69.

Ahn, W.-K., Kalish, C. W., Medin, D. L., & Gelman, S.A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299-352.

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, 55B,* 289-310.

Fernbach, P.M., & Sloman, S.A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35,* 678-693.

Gelman, S. (2003). *The essential child. The origins of essentialism in everyday thought*. Cambridge, MA: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661-716.

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

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 389-394). Austin, TX: Cognitive Science Society.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science, 34,* 1185–1243.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. Cues to causal structure. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154-172). Oxford: Oxford University Press.

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

Marsh, J., & Ahn, W.-K. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35,* 334-352.

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., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-195). Cambridge: 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.

Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.

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. (2007). Essentialism as a generative theory of classification. In A. Gopnik & L. Schulz, (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 190-207). Oxford, UK: Oxford University Press.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science*, *33*, 301-343.

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.

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 659-683.

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

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

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Cognitive Science Society* (pp. 1152-1157). Austin, TX: Cognitive Science Society.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning* (pp. 47-88). San Diego: 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 causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin & Review, 8*, 600-608.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology, 53*, 27-58.

Waldmann, M. R., & Hagmayer, Y. (in press). Causal reasoning. In D. Reisberg (Ed.), *Oxford Handbook of Cognitive Psychology*.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: a minimal rational model. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian Cognitive Science* (pp. 453-484). Oxford: University Press.

Waldmann, M. R., Hagmayer, Y, & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, *15,* 307-311.

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.

Waldmann, M. R., Meder, B., von Sydow, M., & Hagmayer, Y. (2010). The tight coupling between category and causal learning. *Cognitive Processing*, *11*, 143-158.

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

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

Author Note

Address correspondence to Y. Hagmayer or M. R. Waldmann, Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany. Electronic mail may be sent to york.hagmayer@bio.uni-goettingen.de or michael.waldmann@bio.uni-goettingen.de. We thank Nick Chater, Dave Lagnado, Michael Lee, and the anonymous reviewers for their helpful comments on an earlier draft. We also thank Anne Meier-Credner, Dennis Golm, and Almut Hagner for running the experiments.

Footnotes

[1] An entity involved in a causal relation may be a type of event or a type of object having the causal power to affect another object or event (cf. Kemp et al., 2007).

[2] Note that the hypothesis also applies to configurations of features. If different configurations are assumed to be causally relevant, no transfer is entailed by the unbroken mechanism hypothesis even when the same elemental features participate.

[3] We decided to analyze the data by using *t*-tests rather than ANOVAs as the raw data were not normally distributed. *t*-tests are generally considered to be more robust under these conditions.

Table 1

Mean probability ratings (±SE) in Experiment 1a (N = 40) and Experiment 1b (N=44).

		Test Phase 1 (First causal relation *A*→*B*)		Test Phase 2 (Second causal relation *B*→*C*)			
	Category Feature	Molecules formed by mutated genes (Exp. 1a) / high pressure (Exp. 1b)	Molecules formed by normal genes (Exp. 1a) / low pressure(Exp. 1b)	Items 00	Items 01	Items 10	Items 11
Experiment 1a	Color	91.1 (2.5)	10.3 (3.0)	32.1 (5.1)	44.1 (6.5)	58.9 (5.8)	73.3 (4.3)
	Size	91.5 (3.2)	6.3 (2.0)	32.0 (5.1)	60.6 (4.9)	33.4 (5.6)	73.4 (3.8)
Experiment 1b	Color	89.3 (2.9)	12.3 (3.3)	32.0 (5.5)	48.7 (6.2)	46.9 (6.4)	72.7 (5.4)
	Size	90.0 (3.4)	14.4 (4.6)	27.7 (4.2)	50.9 (5.3)	48.6 (5.6)	74.2 (5.1)

Table 2

Mean causal judgments (±SE) in Experiment 2.

	Test Phase 1 (First causal relation)		Test Phase 2 (Second causal relation)					
			Unobserved Test Items			Observed Test Items		
	Protozoa causally affected by alpha microbes	Protozoa causally affected by beta microbes	Difference to within-category prototype	Difference to between-category prototype	*t*-test	Difference to within-category prototype	Difference to between-category prototype	*t*-test
Experiment 2a	78.8 (4.3)	20.4 (5.4)	20.0 (8.1)	48.0 (8.1)	$t(19)= 2.37$ $p < .05$	27.5 (7.1)	40.5 (6.3)	$t(19)= 1.76$ $p < .05$
Experiment 2b	85.8 (3.8)	16.4 (4.4)	28.0 (6.6)	67.0 (7.2)	$t(19)= 2.92$ $p < .01$	36.0 (3.8)	59.0 (3.7)	$t(19)= 3.40$ $p < .01$
Experiment 2c	75.4 (4.7)	28.5 (5.7)	31.8 (9.5)	35.8 (9.4)	$t(19)= .26$ $p = .40$	33.3 (7.5)	34.3 (7.0)	$t(19)= .10$ $p = .45$
Experiment 2d	84.9 (3.2)	16.0 (3.5)	29.3 (6.2)	59.8 (5.4)	$t(19)= 2.87$ $p < .01$	38.3 (5.2)	50.8 (2.8)	$t(19)= 1.80$ $p < .05$

Note. All *p*-values refer to one-tailed *t*-tests.

Figure 1

Causal chains used in Experiments 1 and 2.

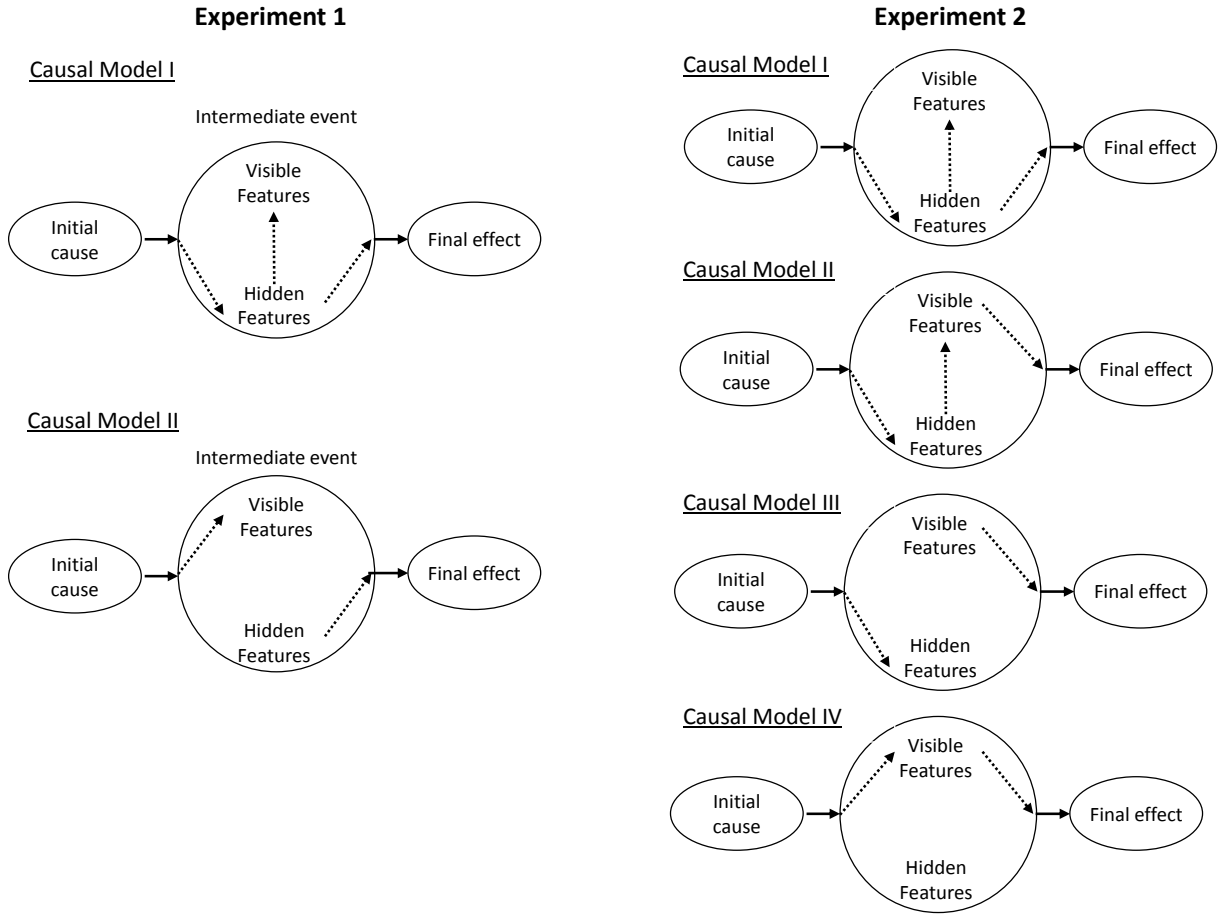

Figure 2

Examples of stimulus material, category boundaries, and probability of item types to cause the final effect (cell death) in Experiments 1a and 1b. See text for explanations.

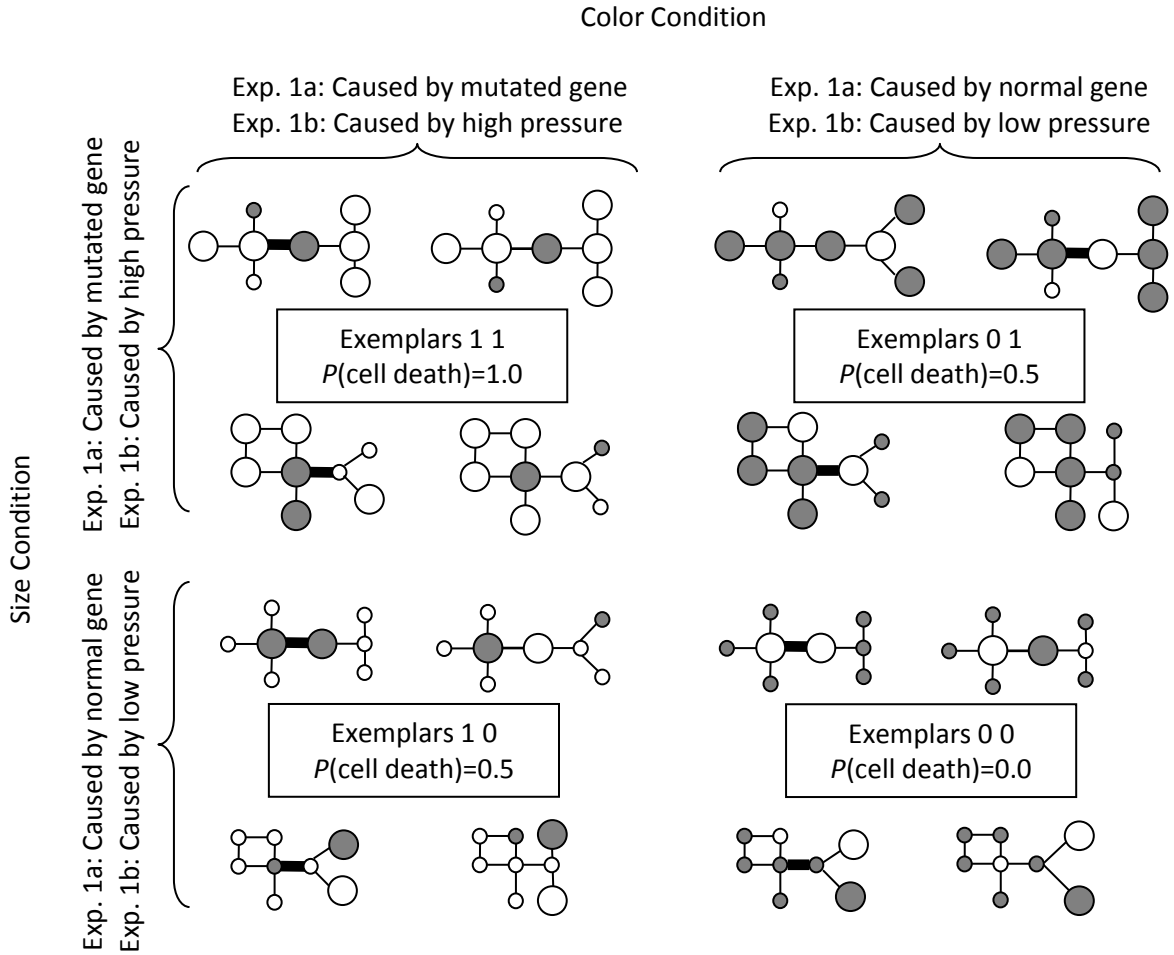

Figure 3

Stimulus materials and category boundaries entailed by the first causal relation (microbes→protozoa) in Experiments 2a-d. See text for details.

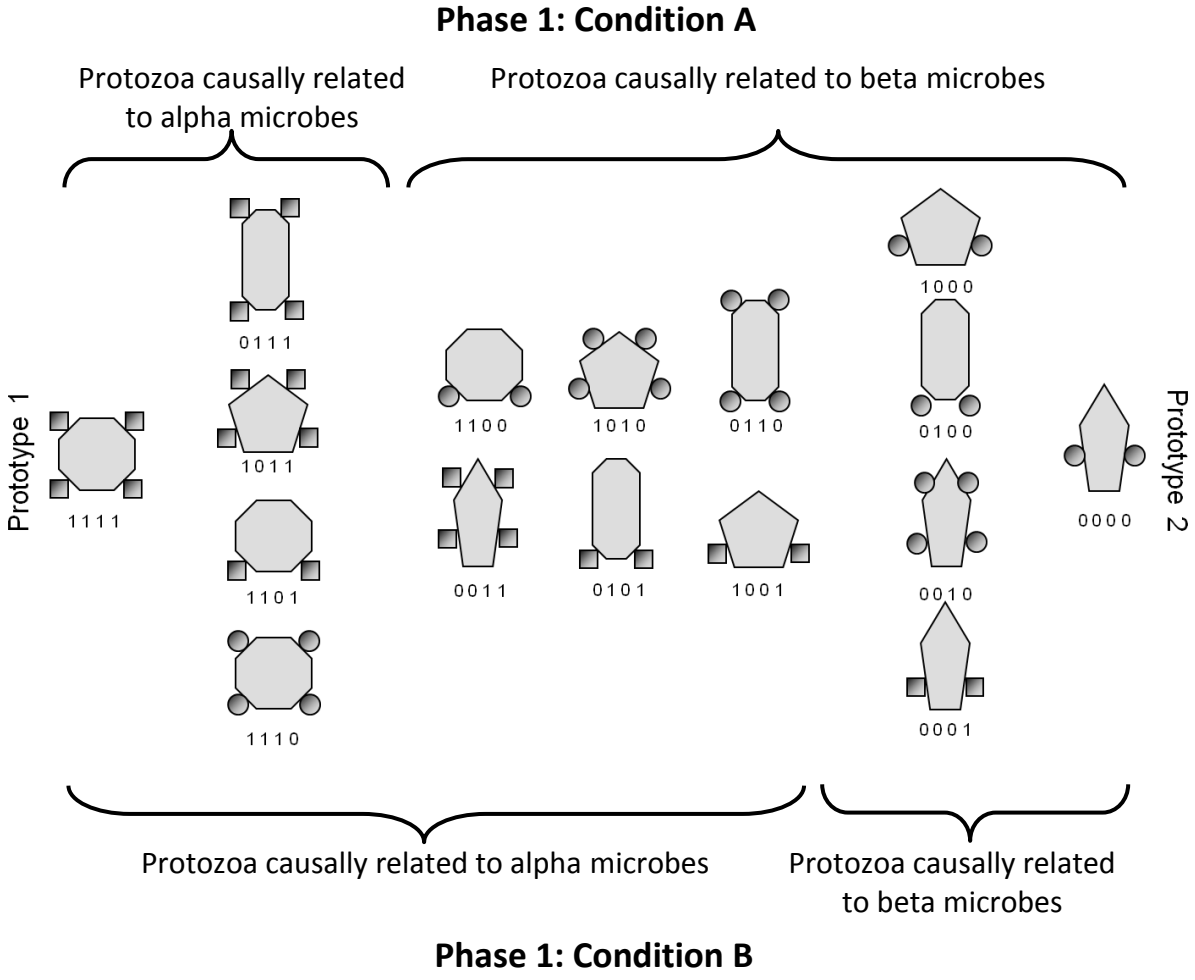

Figure 4

Statistical structure of the second causal relation (protozoae→inflammation) in Experiments 2a-d. Exemplars 1010 and 0101 were omitted in the second causal learning and only used as test exemplars. See text for details.

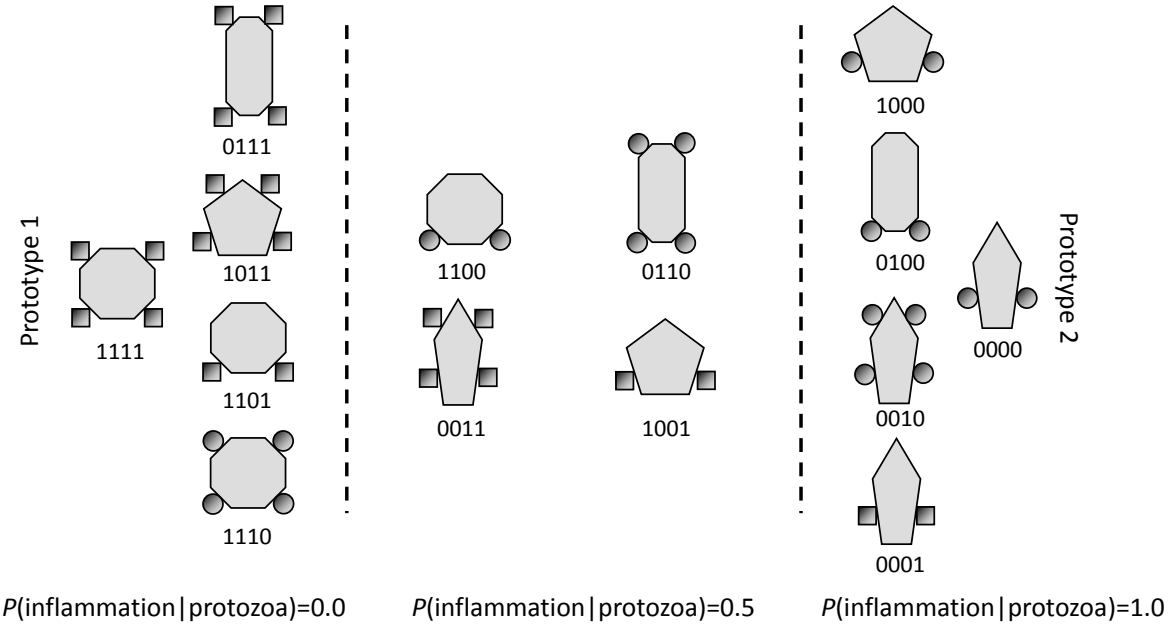

Figure 5

Graphical representations used to manipulate learners' assumptions about underlying causal mechanisms in Experiments 2a-d.