The causal status effect in categorization: An overview

Woo-kyoung Ahn

Vanderbilt University

Nancy S. Kim

Yale University

To appear in D. L. Medin (Ed.) Psychology of Learning and Motivation

When we categorize objects in the world or think about concepts, some aspects of entities matter more than others. In our concept of tires, for instance, roundness is more important than a black color. Even children believe that origins are more primary than certain specific aspects of appearances in categorizing animals (Keil, 1989). This chapter discusses why some features of concepts are more central than others. We will begin with a review of possible determinants of feature centrality. Then, the main body of this chapter will focus on one important determinant, the effect of people's causal background knowledge on feature centrality.

Different Approaches to Understanding Feature Centrality

In general, three approaches have been taken to understand feature centrality in concepts: the content-based approach, the statistical approach, and the theory-based approach. We will review each of these approaches, and discuss how they complement each other.

Content-based Approach

In what we have termed the content-based approach, specific dimensions are documented as being central for certain categories or domains, but not necessarily for others. For instance, it was found that for children, the shape of a rigid object is the most important dimension used to name the object. Landau, Smith, and Jones (1988) presented two- and three-year-old children with a small, blue, wooden inverted U-shaped object, and told the children that the object was a "dax." When asked to select other objects that were also a dax, children preferred objects with the same shape over those with the same size or material. The results, however, changed for different kinds of categories. If the named objects had eyes, children generalized names along the shape and texture

dimension (Jones, Smith, & Landau, 1991). If the objects were made of non-rigid substances such as shaving cream, children generalized names along the color and texture dimensions (Soja, Carey, & Spelke, 1991).

Another example of the content-based approach comes from studies contrasting natural kinds with artifacts. These studies demonstrated that features internal to the object, such as molecular structure, were more important for categorization of an object as a natural kind than for categorization of an object as an artifact. In contrast, features external to the object, such as function, were generally shown to be more important for artifact category membership than for natural kind category membership. An intuitive example is illustrated in Smith Dick King's (1993) "Babe; The gallant pig" (later made into the movie "Babe"): A pig who wanted to be a sheepdog by performing all the correct functions of a sheepdog was scoffed at by other barnyard animals because performing the right function cannot change a natural kind's categorical identity.

Barton and Komatsu (1989) presented adult participants with either natural kinds (e.g., goat) or artifacts (e.g., tire) that had undergone various changes. The results showed that molecular changes (e.g., a goat having a changed chromosomal structure or a tire not made of rubber) mattered more for natural kinds but functional changes (e.g., a female goat not giving milk or a tire that cannot roll) mattered more for artifact membership. Based on these findings, Barton and Komatsu (1989) drew a conclusion that takes the strong content-based approach, arguing that "having a particular molecular or chromosomal structure is necessary and sufficient for (i.e., defining of) membership in a particular natural kind category and that having a particular function is necessary and,

perhaps to a lesser extent, sufficient for membership in a particular artifact category (p. 444)."

Other researchers also found a similar pattern of results. Keil's studies (1989) show that for fourth graders and adults, changing the perceptual appearance of animals was less likely to change the animals' identity than changing the origin of the animals. However, the opposite pattern occurred for artifacts. Similarly, Rips (1989) asked adult participants to read stories about various kinds of transformations. In one story, an animal called a sorp transformed from a bird-like creature into an insect-like creature as a result of either chemical hazards or maturation. Rips' (1989) participants rated chemical hazards as less likely to cause a change in the natural kind's identity than maturation. With artifacts, in contrast, the critical determinant in changing their category membership was the change in the function intended by its designer. It should be noted that unlike Barton and Komatsu (1989), Keil (1989) and Rips (1989) did not take the content-based approach, taking instead the theory-based approach, which will be discussed later.

A final example of the content-based approach is presented in Bloom (1996), who argued "the extension of artifact kind X to be those entities that have been successfully created with the intention that they belong to the same kind as current and previous X's (p.10)." For instance, an object is a chair if it was successfully created with the intention that it would be a chair. Thus, for artifact kinds, the designer's intention to create that object is a defining, and the most central, feature. One critical problem with this argument lies in its circularity: The designer's intention that the object belongs to artifact kind X presupposes X, and therefore, it cannot explain why we have X in the first place.

More generally, a critical problem with the content-based approach is that it does not offer a more fundamental account for why a certain feature is more psychologically central in a concept. That is, simply asserting that a designer's intention or functional features are defining to artifacts does not explain why that became the case in the first place. In relation to the shape bias for rigid objects discussed earlier, Jones et al. (1991) proposed that this was a product of statistical regularities in language; rigid objects with the same name tend to share the same shape. Again, however, it does not explain why our language contains such statistical regularities in the first place. That is, what psychological constraints or mechanisms compelled us to label similarly-shaped rigid objects with a common name? Thus, this approach has difficulty with making novel predictions as to what type of feature would be central for a novel kind.

The second general problem with the content-based approach is that the empirical results do not always agree with the claims. For instance, Malt and Johnson (1992) found that physical properties of artifacts (e.g., for boats: "is wedge-shaped, with a sail, anchor, and wooden sides," as selected by undergraduate subjects in the study) were judged to be more important than, or just as important as, functional features (e.g., "to carry one or more people over a body of water for purposes of work or recreation"). In a study that more directly examines the role of functional features, Ahn (1998) showed systematically conditions under which functional features can be central for natural kinds and compositional can be central for artifact. We will later revisit this study.

Statistical Approach

The second approach to the issue of feature centrality is to examine statistical patterns of features in a category, such as category validity (the probability that an object

has a certain feature given that it belongs to a certain category; e.g., the probability that an object has wings given that it is a bird) and cue validity, or diagnosticity (the probability that an object belongs to a certain category given that it has a certain feature; e.g., the probability that an object is a bird given that it has wings). For instance, Rosch and Mervis (1975) used category validity ("the number of items in a category that had been credited with that attribute," p. 578) as a measure of family resemblance, and showed that this measure was highly correlated with subjects' typicality ratings. That is, having features shared by many members of the same category makes an exemplar "good." Tversky (1977) showed how the diagnosticity of features can affect similarity judgments. Kruschke's ALCOVE (1992) learns attention strengths (i.e., feature centrality) as a function of the diagnosticity of features. Corter and Gluck (1992) proposed a measure called category utility which is the product of the base rate, cue validity and category validity of a feature. They also showed that category utility is a good predictor for which level in a hierarchical structure is selected to be the basic level of categorization. Thus, the general consensus has been that the cue and category validities of a feature determine feature centrality to some extent.

However, the centrality of features does not seem to depend exclusively on the probabilities associated with features. For instance, a square basketball and a square cantaloupe both have category validities and cue validities of zero, but people are more willing to accept a square cantaloupe than a square basketball (Medin & Shoben, 1988).

Furthermore, the importance of cue and category validities in determining feature centrality depends on the type of categorization task. Cue validity would play a more important role in a task where one determines whether an object is A or B than in a task

where one conceptualizes about an object. For instance, curvedness is a highly diagnostic feature in determining whether an object is a banana or a lemon. However, it is not as central when we think about bananas because a straight banana is acceptable (Medin & Shoben, 1988). Similarly, all purple seedless grapes are purple, but being purple is not so much conceptually central for purple seedless grapes in that it is merely a surface feature of purple seedless grapes rather than an essential feature of what really makes them purple seedless grapes, such as their origin or DNA (Ahn & Sloman, 1997; Sloman & Ahn, 1999).

Theory-based Approach

The third approach to feature centrality posits that the centrality of any given feature is determined by its importance in the principles or theories underlying the categories (Murphy & Medin, 1985). Categories are connected in many complex ways that resemble structured theories. For example, our concept of "boomerang" is connected with other concepts such as "throwing," "air," "speed," and so on, all of which are intricately connected as in a scientific theory. Features that play an important part in commonsense or naive theories seem more essential in categorization than those that do not. For example, the feature *curvedness* plays a role in a theory of physics, and it thereby becomes a central feature for categorizing objects as boomerangs. In the case of bananas, however, curvedness does not play a critical role in a naive biological theory, and it is therefore not so central. Although the previous theory-based approaches have all discussed the importance of causal background knowledge, the exact mechanism has rarely been articulated (Gelman & Kalish, 1993; Murphy, 1993). That is, previous approaches have not specified how to determine the role a feature plays in one's domain

theory. We extended the theory-based approach and proposed a causal status hypothesis (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis, in press). This hypothesis provides one way to operationalize feature centrality in terms of one's causal background knowledge. This chapter will extensively discuss the causal status hypothesis.

Before we discuss the causal status hypothesis, it should be pointed out in all fairness that the theory-based approach alone cannot completely account for all issues in feature centrality even if its underlying mechanism is specified. Instead, there are a number of different kinds of determinants of feature centrality. At the very least, a feature's centrality is determined by the domain an object belongs to, a feature's statistical value, and a feature's role in one's background knowledge. For now, we will focus only on how one's causal background knowledge can determine a feature's centrality and will later return to the issue of different kinds of feature centrality.

Causal Status Hypothesis: General Introduction

In accordance with the theory-based view, we assume that concepts consist of richly structured features, rather than a set of independent features (Murphy & Medin, 1985). We assume that people have naive theories about how these features are connected and that the majority of features of existing concepts are causally connected[1] (Carey, 1985; Gelman & Kalish, 1993; Wellman, 1990). Given such representational assumptions, the causal status hypothesis states that people regard cause features as more important and essential than effect features in their conceptual representations.

The causal status hypothesis is intuitive, and its real-life examples are abundant. We form an illness category based on the virus that causes the symptoms rather than by

[1] The last section of this chapter will provide more extensive discussion about these assumptions.

the symptoms per se. DNA structure causes many other properties of plants and animals, and hence is considered the most important feature in their classification (e.g., a plant that is found to lack tulip DNA will never be classified as a true tulip). In law, the severity of the crime often depends more on the suspects' intentions rather than their surface behaviors (e.g., killing someone by accident is a much less serious offense than intending to kill someone but inadvertently botching the plan). In judging whether somebody is nice or not, we look at their intentions rather than what they did.

At a moment's reflection, one can also generate several counterexamples to the causal status hypothesis. When doctors diagnose Alzheimer's disease, the diagnosis is not based on the cause but rather on symptoms alone. When we recognize plants and animals, we normally do not have direct access to their genetic structures, but we can confidently categorize them based only on their surface features. An orchestra selects its members based on their performance abilities rather than their training background. Then, the important question arises: to what extent does the causal status of features determine their centrality within our conceptual representations?

In the remainder of the chapter, we will focus on the causal status hypothesis. We will present our rationale for predicting the causal status effect, and empirical results supporting the hypothesis under various contexts. We will describe previous categorization studies that can be accounted for by the causal status hypothesis. We will also discuss moderating factors for the causal status effect, including how other types of determinants for feature centrality reviewed in this section might alter how strongly the causal status effect is manifested. Followed by this discussion of the causal status hypothesis in the context of categorization literature, we will present one specific

application of the causal status effect; namely, the effect of people's theories about mental disorders on their diagnosis decisions, and the possible ramifications of this effect for DSM-IV (APA, 1994) criteria. Finally, we will examine the potential consequences of focusing only on causal relations among features in studying the effect of lay theories on feature centrality.

Essentialism and the Causal Status Hypothesis

One traditional and pervasive view about concepts is that things have essences that make them what they are (e.g., Locke, 1894/1975). Essentialism is particularly related to the causal status hypothesis in that both are concerned with the causal potency of essential (or central) features. For instance, Locke stated, “And thus the real internal, but generally in Substances, unknown Constitution of Things, whereon their discoverable Qualities depend, may be called their *Essence*.” (p. 417) Similarly, Putnam (1977) wrote, "If I describe something as a lemon, or as an acid, I indicate that it is likely to have certain characteristics (yellow peel, or sour taste in dilute water solution, as the case may be); but I also indicate that the presence of those characteristics, if they are present, is likely to be *accounted for by some 'essential nature'* which the thing shares with other members of the natural kind" (p. 104, emphasis added).

It is controversial as to whether or not things actually have essences in the philosophical sense. Instead, Medin and Ortony (1989) proposed the idea of psychological essentialism: people act *as if* things have essences which are responsible for the surface features of objects. If psychological essentialism holds true, the deepest cause feature in a category would be our best guess as to what an essence would be, and

it is reasonable to categorize objects based on the deepest cause feature[2] of a category (i.e., the essence in our conceptual representation). In this way, psychological essentialism provides a basis for the causal status hypothesis.

However, there are important differences between essentialism and the causal status hypothesis. (See Ahn et al., in press; Ahn, 1998 for more detail.) First, a strong version of essentialism might state that essences (whether known or unknown) serve as defining features of a category, and thus assume all-or-none categorization where surface features do not affect categorization. (See Braisby, Franks, & Hampton, 1996; Diesendruck & Gelman, 1999; Kalish, 1995; Keil, 1989; Malt, 1994 for similar descriptions of essentialism.) In contrast, the causal status hypothesis does not make any statement about whether or not there is a certain class of features that serve as defining features. Indeed, the gradient structure of feature centrality is built into the hypothesis. That is, because a feature is more central than its effect but less central than its cause, it is logical to predict that the deeper a cause is in a causal chain, the more conceptually central that feature is.

Second, essences are often described as internal, hidden, or unobservable properties such as atomic weight, genetic structure, and so on, whereas surface features are external or perceptual features such as wings and feathers in birds (e.g., Putnam, 1975). Locke (1894/1975) even stated that real essences are undiscoverable. The causal status hypothesis does not impose any restrictions on the type of cause or effect features, or the type of central or peripheral features. Third, essentialism as claimed by philosophers (e.g., Kripke, 1972; Putnam, 1975) is a statement about things in the world,

[2] A causal chain is infinite and one might wonder how we decide when to stop explanations. We briefly discuss this issue later under "Domain generality."

and it argues that categorization of natural kinds is independent of our knowledge of essential properties[3]. For instance, even if it turns out that water does not consist of H_2O, and instead consists of XYZ, what we have been referring to as water is still water. In that sense, natural kinds are like proper nouns in that even if we find out, for example, that Shakespeare did not write *Romeo and Juliet*, Shakespeare is nonetheless Shakespeare. However, Braisby et al. (1996) demonstrated that after discovering changes of essential properties (e.g., cats being robots controlled from Mars), people responded that category membership of the object should also be somewhat changed (e.g., the object formerly known as a cat is not a cat any more). Likewise, the causal status hypothesis assumes that more responsibility is given to specific knowledge because feature centrality should change as our knowledge about the causal relations among features changes. For instance, if cats' behavior turns out to be caused by Martians rather than cat DNA, then cat DNA would not be as central any more in categorizing cats.

Summary. Essentialism serves as a basis for the causal status hypothesis. If essences are the deepest cause in a category, both essentialism and the causal status hypothesis acknowledge the special status of essences in that they are the most central features in the category. However, unlike essentialism, the causal status hypothesis does not assume that there are defining features in concepts, categorization is all-or-none, deeper features are internal or hidden, or that categorization is independent of knowledge.

[3] It should be noted that related construct of psychological essentialism (e.g., Medin & Ortony, 1989) does not make this claim, because psychological, as opposed to philosophical, essentialism is about mental representations of things in the world.

In the next section, we summarize empirical support for these differences and for other predictions of the causal status hypothesis.

Recently, Stevens (2000) argued that previous studies by Gelman and her colleagues that are considered to be evidence for psychological essentialism can in fact be accounted for simply by assuming beliefs in causal laws without having to resort to claiming the existence of essences in our conceptual representation. For instance, Gelman and Markman (1986) showed children pictures of a gray squirrel, a kaibab (a kind of squirrel that looks like a rabbit), and a rabbit. When told that the gray squirrel eats bugs and that the rabbit eats grass, most 4-year-olds said that the kaibab eats bugs. Stevens argued that it is much more parsimonious to account for this kind of result by assuming that children believe that something about being a squirrel causes an animal to eat bugs, and that the fact that the something is an essence adds no explanatory power. On one hand, we hesitate to agree fully with his claim because no convincing evidence has been presented to rule out psychological essentialism. On the other hand, we agree that causal laws were most likely a critical driving factor for these results. After reviewing empirical support for the causal status hypothesis, we will also review how previous studies can be accounted for in terms of the causal status hypothesis.

Main Empirical Results Supporting the Causal Status Hypothesis

The most direct test of the causal status hypothesis is reported in Ahn et al. (in press). Participants in their Experiment 1 read about three characteristic features of a target category (e.g., animals called “roobans” tend to eat fruits, have sticky feet, and build nests on trees). Participants in the control condition received no further information about the target category. In contrast, participants in the experimental condition were told

that one feature tends to cause the second feature, which in turn tends to cause the third feature (e.g., eating fruits tends to cause roobans to have sticky feet because the fruit sugars are secreted through pores on the undersides of their feet, and that having sticky feet tends to allow roobans to build nests on trees because they can climb up trees easily with their sticky feet). Finally, all participants were presented with three exemplars, each of which had two features characteristic of the target category and one non-characteristic feature (e.g., an animal that likes to eat worms, has feet that are sticky, and builds nests in trees). Participants were then asked to rate how likely it was that this animal is a member of the target category (e.g., a rooban).

For participants in the control condition, likelihood ratings remained constant regardless of which feature the exemplar animal was missing. However, in the experimental condition, the likelihood judgments varied as a function of the missing feature's causal status. Specifically, when an exemplar was missing the target category's fundamental cause in the causal chain, the mean likelihood of being a target category member was lower than when an object was missing its intermediate cause in the causal chain, which in turn was lower than when an object was missing its terminal effect. That is, the deeper a feature was in a causal chain, the more central it was in a categorization judgment.

One alternative explanation for the above results is that participants might have assumed that category validities of features vary as a function of causal status. Whereas category validities of features in a category can be an important determinant of feature centrality as discussed earlier, high category validities do not necessarily entail high causal status. For instance, "being purple" in purple seedless grapes has a category

validity of 1 but it has low causal status because it does not cause any other features in that category. Note, however, that many cause features in basic-level categories also tend to have high category validities. For example, birds' genetic codes are causally central, and they are also present in all birds. This observation suggests that, due to this correlation in real-life categories, participants in Experiment 1 of Ahn et al. (in press) - which did not explicitly specify category validities of features - might have assumed that the cause features must have high base rates. Thus, it could be argued that these inferred high category validities of the cause features determined their conceptual centrality rather than their causal status per se.

Experiment 2 of Ahn et al. (in press) controlled for this problem. In this experiment, participants observed 12 representative samples of the target category before they made judgments on transfer items (and before those participants in the causal condition received information about the causal relations). The actual category validities of three characteristic features of a target category were held constant in these samples. After observing 12 samples, participants were asked to judge the frequency of each feature. Following these frequency judgments, participants performed the identical tasks as in the first experiment with half the participants in the causal condition and the other half in the no-causal control condition. The results showed that all participants judged frequencies fairly accurately; that is, there was no significant difference between features. Yet, the causal status effect was replicated.

Ahn et al. (in press, Experiment 3) also used a different categorization task and found the causal status effect. Imagine Jane who is depressed because she has low self-esteem. Now determine who should be categorized with Jane: Susan who is depressed

because she has been drinking, or Barbara who is defensive because she has low self-esteem. Participants in this study who received sets of items similar to this example preferred to categorize objects and persons based on a matching cause (e.g., categorizing Jane and Barbara together) rather than a matching effect (categorizing Jane and Susan together).

Thus, the basic causal status effect was documented in tasks involving both category likelihood judgments and free sorting. Furthermore, the causal status effect occurs above and beyond any confounding effect of category validities. The existence of the effect seems clear, but how far-reaching are its ramifications? In the next section, we will demonstrate how feature centrality phenomena shown in previous experiments can be explained by the causal status hypothesis.

Related Phenomena Accounted For By The Causal Status Hypothesis

Natural Kinds Versus Artifacts

As discussed earlier under the content-based approach to feature centrality, previous studies (e.g., Barton & Komatsu, 1989; Gelman, 1988; Keil, 1989; Rips, 1989) have shown that different features are central for natural kinds and artifacts: in natural kinds internal or molecular features are more conceptually central than functional features, but in artifacts functional features are more conceptually central than internal or molecular features. As discussed earlier, it is tempting to take the content-based approach based on these findings and conclude that there is something inherently special about molecular features for natural kinds and functional features for artifacts.

In contrast, Ahn (1989) argued that the mechanism underlying this phenomenon is the causal status effect. That is, in natural kinds, internal / molecular features tend to

cause functional features (e.g., cow DNA determines whether or not cows give milk) but in artifacts, functional features determine its compositional structure (e.g., chairs are used for sitting, and for that reason, they are made of a hard substance).

Experiments 1 and 2 in Ahn (1998) examined the real-life categories used in previous studies (Barton & Komatsu, 1989; Malt & Johnson, 1992). Participants were asked to draw causal relations among features within the same category. At the same time, they judged the centrality of features as measured by the degree to which a feature impacts categorization when that feature is missing. It was found that across natural and artifactual kinds, the more features any particular feature caused, the more influential the feature was in categorization. In addition, Ahn (1998) directly manipulated the causal status of features using artificial stimuli, and showed that when a compositional feature caused a functional feature, a compositional feature was more influential in categorization of both natural and artifactual kinds, whereas the opposite was true when the causal direction was reversed. Thus, the results suggest that the more fundamental determinant of feature centrality is not the specific content of features but rather the causal role that a feature plays. This way, the causal status hypothesis offers a more fundamental account than has previously been put forth for why people might treat natural kinds and artifacts differently.

Category Use Effect

We speculate that additional existing findings in the literature may be viewed as having been mediated by a causal status effect. For example, Ross (1996, 1997, 1999) found that using category knowledge to perform a non-classification task resulted in changes in feature weighting on a subsequent classification task. In a series of

experiments in Ross (1997), participants learned to categorize hypothetical patients as having one of two novel diseases, on the basis of four relevant symptoms. During this training phase, participants also learned to prescribe one of two treatments for each of the diseases, on the basis of two relevant symptoms for each disease. For instance, four symptoms of buragamo disease were a fever, a runny nose, dizziness, and abdominal pain, and buragamo patients with a fever were treated by lamohillin and buragamo patients with runny noses were treated by pexlophene. Ross found that although all four symptoms perfectly predicted the disease, symptoms that predicted the treatment to be prescribed were treated as more predictive of the disease too.

We conjecture that this result might have been obtained because the treatment-relevant symptoms were perceived to be causally central to the disorder. For instance, when a patient with a fever is always treated by lamohillin, as in Ross's (1997) study, this implies to participants that lamohillin acts specifically upon a fever. A treatment that is able to cure a disease is generally most likely to act upon the causal symptoms of the disease, not the peripheral effects. For example, one would not expect an ear infection caused by bacteria to be treated effectively with medicines that simply suppress symptoms (e.g., pain killers). Instead, to cure the ear infection, antibiotics that act directly upon the bacteria should be given. Moreover, it may be that treatment-relevant symptoms were considered by Ross's participants to be causally central in a broader sense, in that they determined which treatment plan to adopt. Thus, we suggest that Ross's category use effect may be mediated by a causal status effect. That is, within our causal status framework, the fact that Ross's participants felt that treatment-relevant symptoms were most important in diagnosing the disease can be traced to the experiment's implication

that these symptoms are the most causal. Thus, it would be extremely difficult to pit the category use effect against the causal status effect because whenever a feature is relevant for a certain use, it becomes a causal feature in that the feature determines the use of the category.

The causal status hypothesis also appears to be compatible with findings involving the use of categories in problem solving. For example, Chi, Feltovich, and Glaser's (1981) classic experiment assessed the categorical representations of physics problems in advanced physics graduate students and professors (experts) and undergraduates who had just completed a semester-long course in mechanics (novices). They found that experts consistently based their categorization of physics problems on deeper physics principles, such as conservation of momentum or the work-energy theorem, whereas novices categorized based on the surface features of the problems, such as whether there was an inclined plane or pulley in the problem. Their findings seem to be related to the causal status hypothesis in that both indicate a tendency to not categorize based on surface features when knowledge about a deeper structure is available. That is, experts in Chi et al.'s (1981) study were aware of both the surface features and the deeper structure, but preferred to categorize based on the latter. Similarly, novices in Chi et al.'s study did not categorize the problems based on even more surface features, such as the number of lines in the problem (see Keil, Smith, Simons, & Levin, 1998 for a similar argument). It should be noted, however, that in Chi et al.'s study, participants were specifically asked to sort the problems "based on similarities of solution" (p.124). Thus, it remains to be seen whether they would sort problems by deeper properties (i.e., deeper properties within their understanding) even without these instructions. Given that Ahn et

al.'s (in press) Experiment 3 described earlier showed that people spontaneously create categories based on a matching causal feature rather than a matching effect feature, we would predict that both experts and novices will spontaneously sort problems based on the deepest cause known to them.

Appreciation for Intentionality

Recently, several developmental studies demonstrated that even young children appreciate intentionality behind external representations (e.g., drawings) more than their appearance in naming the representations. Bloom and Markson (1998) asked 3- and 4-year-old children to draw, for instance, a balloon and a lollipop. As one might expect, these drawings looked almost identical. Later, when the children were asked to name these drawings, the children named the pictures on the basis of what they had intended for them to depict. Similarly, Gelman and Ebeling (1998) showed that when the same picture was produced either intentionally or accidentally (e.g., drawings had been intentionally created to be a bear versus somebody spilled paint), intentional representation led to higher rates of naming responses (e.g., “bear”) than did accidental representation. Clearly, the intention behind drawings is a causal factor of the drawings' appearance. Thus, appreciation for the drawer's intention when naming drawings can be construed as an example of the causal status effect.

Developmental Studies

It has been shown that even children discriminate features when they categorize objects or use concepts. For instance, Keil (1989) demonstrated that fourth graders base their categorization on origins of animals (e.g., being born from another raccoon), rather than on perceptual appearance (e.g., black with a white stripe on the back). Gelman

(1988) found that second graders are more likely to generalize internal parts of an animal (e.g., a spleen inside a rabbit), rather than functional features, to other instances of the same kind. Once again, these results seem to be an example of the causal status effect: origins or internal parts of animals can be thought of as determining the surface features, and therefore, they are more essential in children's categorization.

Unfortunately, however, previous studies did not provide direct evidence that children believe that essential features cause other features. This leaves open the question of whether children's preference for deeper features is derived from accumulated past knowledge, or whether it instead manifests a belief brought to the task of acquiring knowledge, such as the causal status effect. For instance, Keil's (1989) transformation experiments further showed the content-specific effect in that fourth graders can differ from adults when different stimulus materials were used. Fully half the fourth graders in these experiments thought that changing a tiger into a lion by injection given early in life would make it a lion, whereas most adults did not. Keil (1992) argued that this difference is due to the fact that understanding is knowledge-dependent. That is, the argument is that the difference occurred because the fourth graders did not have the right sort of knowledge rather than because they lack a bias toward weighing causal features.

One way of testing this claim is to use artificial categories and directly manipulate causal relations among features to see whether causal status can determine feature centrality in children's categorization. Using this framework, Ahn, Gelman, Amsterlaw, Hohenstein, and Kalish (under review) found the causal status effect in 7- to 9-year-old children. In this study, adults and children learned descriptions of novel animals, in which one feature caused two other features. When asked to determine which transfer item was

more likely to be an example of the animal they had learned, both adults and children preferred an animal with a cause feature and an effect feature than an animal with two effect features. Thus, children at this age group do indeed show the causal status bias. It remains to be seen whether even younger children have the causal status bias.

Summary

The causal status effect seems to be prevalent in various aspects of categorization and our use of concepts. It explains why different features are central for natural kinds and artifacts, in that internal/molecular features tend to be causally central in natural kinds, whereas functional features tend to be causally central in artifacts. It may be the underlying mechanism for phenomena involving the use of categories in reasoning, including Ross's (1997) category use effect and Chi et al.'s (1981) classic experiment on experts' and novices' sorting of physics problems. It may mediate our appreciation for intentionality as demonstrated by Bloom and Markson (1998). Finally, it is able to explain one way in which children weight features in categorization. Given that the causal status effect appears to be present in a variety of phenomena, a way to implement this general effect computationally would provide the means to facilitate further research. In the next section, we review a model of the causal status effect that fulfills this purpose.

Computational Modeling

The causal status hypothesis provides one well-defined way of constraining feature weights, within the framework of the theory-based approach to categorization. An example of a computational implementation of the causal status hypothesis is presented in Sloman, Love, and Ahn (1998). This model is based on general "dependency"

relationships rather than being restricted only to causal relationships. (See the last section of this chapter for further discussion about the difference between these two.) The idea of the model is that the more other features depend on (e.g., are caused by, are determined by, are followed by, etc.) a feature i, the more central this feature i becomes to the concept. More specifically, the model states that conceptual centrality of a feature ($C_{i,t}$) is a function of dependency as follows:

$$C_{i,t} = \sum A_{ij} C_{j,t-1} \qquad (1)$$

where A_{ij} is the strength of a dependency link from feature j to feature i. This formula states that the centrality of feature i is determined at each time step by summing across the immutability of every other feature multiplied by that feature's degree of dependence upon feature i. For instance, suppose feature X causes feature Y, which causes feature Z, and the causal strengths of both relations are 3, and the initial value for feature centrality was an arbitrary value of, say, 1. After two iterations, the centralities of features X, Y, and Z become 16, 7, and 1, respectively. These qualitative differences in feature centrality predicted by the model are consistent with the results found in Experiments 1 and 2 of Ahn et al. (in press). This model is just one of many possible ways of computationally modeling the causal status effect. In fact, there are results that this particular model cannot account for, as we shall see later in the chapter.

Moderating factors

Although the causal status effect is robust, it is important to specify its boundary conditions. This section discusses various factors that can moderate the causal status effect. Another goal of this section is to account for apparent counterexamples to the causal status effect in terms of these moderating factors.

Influence Of Other Determinants Of Feature Centrality

At the beginning of this chapter, we reviewed different approaches to feature centrality. Similarly, Sloman et al. (1998) examined conceptual centrality (mutability, or the degree to which a feature in a concept can be transformed while maintaining the concept's coherence), category centrality (perceived relative frequency of a feature within a category), diagnosticity (evidence provided by a feature for one category relative to a set of categories), and prominence (how salient the feature seems to people when thinking about the category). In order to show that these are empirically dissociable, Sloman et al. (1998) asked participants questions that were designed to measure each of these centralities. For instance, conceptual centrality was measured using a number of questions, including a measure of surprise (e.g., How surprised would you be to encounter an apple that did not grow on trees?), ease-of-imagining (e.g., How easily can you imagine a real apple that does not grow on trees?) goodness-of-example (e.g., How good an example of an apple would you consider an apple that does not ever grow on trees?) and similarity-to-an-ideal (e.g., How similar is an apple that doesn't grow on trees to an ideal apple?). In a factor analysis, they found that questions selected to measure the same centrality loaded on the same factor, and questions measuring different centralities loaded on different factors. (One exception to this result was that the questions measuring the conceptual centrality and those measuring categorical centrality loaded on the same factor. Sloman and Ahn [1999] later address this issue, as is discussed in a later section of this chapter.) Thus, in general, these results show that different kinds of feature centrality are empirically dissociable.

One intuitive example for understanding this phenomenon is illustrated in a Calvin and Hobbes cartoon in Figure 1. Calvin has just lost his beloved stuffed tiger Hobbes, so he is working on a "lost" ad. The ad should be designed to facilitate a perceptual categorization task, but instead of describing perceptually central features, Calvin listed Hobbes' features that are conceptually central to him, making this cartoon amusing.

Insert Figure 1 about here.

Among these various types of feature centrality, the causal status effect is related to conceptual centrality; the measures of surprise and goodness-of-example are similar to the ones used in Ahn et al. (in press). Furthermore, Sloman et al. found that conceptual centrality was the only type of centrality that correlated reliably with a feature's status in a concept's dependency structure (i.e., centrality predicted by equation [1]) across various levels of categories (see also Ahn & Sloman, 1997; Sloman & Ahn, 1999). As explained earlier, equation (1) is one way of computationally modeling the causal status effect. Thus, these results suggest that the causal status effect is likely to be most pertinent to tasks involving conceptual centrality, and may be less influential in tasks such as perceptual categorization (which would be determined by perceptual prominence), or discrimination tasks (which would be determined by diagnosticity of features).

Levels of Abstraction and Name versus Conceptual Centrality

The fact that there are different types of determinants for feature centrality that act on different types of categorization tasks can account for a set of examples that

apparently contradict the causal status hypothesis. For instance, although "being purple" is not causally central in our concept of purple seedless grapes, it is a necessary feature for the category of purple seedless grapes. Similarly, in our concept of pine trees, "having needles" is not as causally central as "having a trunk", but we would not call an object a pine tree if it did not have needles. These apparent counterexamples can be explained if we understand the distinction between name centrality and conceptual centrality.

Sloman and Ahn (1999) argued that the name centrality of a feature (i.e., the feature's power to determine the appropriateness of a category label for an object) is a function of feature frequency in a category whereas conceptual centrality is a function of features' relational structure (e.g., causal status). In most cases, the two types of centrality are highly correlated (e.g., "having wings" for birds). But oftentimes, we discriminate between categories that do share conceptually central properties, and assign different names to them. Such cases tend to occur at a high level of specificity. For instance, grapes have conceptually central properties (e.g., is juicy, grows on vines), and we subdivide them into purple versus green seedless grapes. By virtue of specifying a distinct category, the name must refer to one or more features that make the distinction, but are not necessarily conceptually central (e.g., "being purple").

The counterexamples we provided at the beginning of this section can be explained by this framework. For instance, "having needles," which has low causal centrality, has high name centrality (i.e., we would not name an object a pine tree unless it had needles), but it still has low conceptual centrality (i.e., it is not difficult to imagine an object that is in all ways like a pine tree except that it did not have needles). Sloman and Ahn (1999) showed that participants also shared this intuition. The question format

used to measure name centrality was, "Suppose an object is in all ways like an X except that it does not have feature Y. How appropriate would it be to call this object an X? " and the question format used to measure conceptual centrality was, "How easy is it to imagine an object that has ALL the features of an X except that it does not have feature Y?" Sloman and Ahn (1999) found that features that distinguish categories at a specific level have high name centrality but low conceptual centrality. In addition, they found that few features of the same concept depended on these specific-level features, consistent with the causal status hypothesis. Hence, the above "counterexamples" of the causal status effect become moot when we distinguish name centrality from conceptual centrality.

The Case of a Common-Effect Structure

In the previous two sections, we focused on the fact that different determinants of feature centrality are dissociable. However, real-life categorization often involves multiple categorization processes. For instance, names of concepts might influence the way we think about concepts (e.g., Roberson, Davidoff, & Davies, 1999). Or when we think about object concepts, we might also visualize the object. Under such circumstances, different determinants of feature centrality can simultaneously act on categorization. For instance, a feature with high causal status but low category validity might end up having mediocre feature centrality. When other determinants of feature centrality heavily favor the effect feature over its cause, it is reasonable to expect that these multiple influences can override the causal status effect. Alternatively, the causal status effect can be enhanced when other determinants of feature centrality favor the cause feature over its effect.

This discussion about the influence of other determinants of feature centrality on the causal status effect is useful in analyzing a case that has been frequently posed as a counterexample to the causal status hypothesis. The case we will examine in this section is a common-effect structure, in which multiple causes have a single effect in common (see Figure 2 for an abstract structure). We will first discuss real-life examples involving a common-effect structure, followed by Rehder and Hastie's experimental results (1997, in preparation).

Insert Figure 2 about here.

Real-life example. Some real-life concepts have a common-effect structure[4]. One example is a syndrome: a syndrome is a cluster of symptoms with some predictive value in terms of treatment and prognosis without a single known etiology, and thus it tends to have multiple causes for a common set of symptoms. Another example is pneumonia: it is a serious infection or inflammation of the lungs which can have over 30 different causes, including bacteria, viruses, chemicals, or even inhaled objects like peanuts or small toys.

In these examples, a cause seems less crucial than its effect in determining category membership. For instance, in determining whether or not a person has pneumonia, the cause of the lung inflammation is less important than the lung

[4] The fact that we have concepts containing common-effect structures might be considered as a counterexample to the causal status hypothesis. That is, why would we have any category that was not created based on a single cause? In the "Domain Generality" section, we will discuss the issue of why people sometimes create concepts with a common-effect structure.

inflammation itself[5]. As discussed above, however, feature centrality may be determined by other factors, such as category validity. In the case of pneumonia, the effect feature has a category validity of one (i.e., all patients with pneumonia have lung inflammation), whereas each of the cause features has a much lower category validity because there are many possible causes for pneumonia. For the feature centrality of each of these causes, any amount of the causal status effect that occurs is cancelled out due to its low category validity. Furthermore, in the case of the common-effect structure, the effect feature has a greater advantage over the cause features in terms of the number of relations in which it participates. Gentner and her colleagues (e.g., Gentner, 1989) have suggested that relational features are more important than isolated features in analogical reasoning. If the number of relations in which a feature participates also determines feature centrality, then an effect feature in a common-effect structure can outweigh one of the cause features because the effect feature participates in more relations than any one of these cause features.

Because these multiple counteracting forces act upon the cause features in a common-effect structure (low category validity, fewer relations to participate in), the overall centrality of one of these causes in a common-effect structure can end up becoming lower than the overall centrality of its effect feature. Exactly how all these potential determinants for feature centrality might be combined awaits more future research. With these examples we attempted to argue in this section that a failure to

[5] In diagnosing pneumonia, causes for lung inflammation are also crucial information because they determine different treatment plans. However, in simply determining whether or not someone has pneumonia, causes would not be as critical as lung inflammation per se because there are many potential causes whereas lung inflammation is a necessary feature for pneumonia.

observe the causal status effect in common-effect structures as in the above situations is not evidence for a counterexample to actual occurrence of the causal status effect because in these situations, there are a disproportionate number of counteracting forces.

Consider again the case of pneumonia: the causal status hypothesis is supported when these counteracting forces are removed. Indeed, the American Lung Association (1998) states explicitly that "pneumonia is not a single disease," precisely because there are many causes. Instead, they break the disease down into various sub-categories including Bacterial Pneumonia, Viral Pneumonia, Mycoplasma Pneumonia, and so on, based on the type of cause, just as the causal status hypothesis would predict. Furthermore, note that each sub-type no longer has a common-effect structure, and within each sub-type, the category validity of the cause feature and the category validity of the effect feature become equal. For instance, the probability of *having lung infection* given that a person has bacterial pneumonia is the same as the probability of *being infected with pneumonia bacteria* given that a person has bacterial pneumonia (i.e., 1). In addition, both of these features participate in the same number of relations. Now that all the other counteracting forces are removed, the cause feature (e.g., being infected with pneumonia bacteria) may be measured to be more central than the effect feature (e.g., having a lung infection). For instance, a patient infected with pneumonia bacteria, but who has not yet developed lung inflammation (missing effect) is at least more likely to be considered as specifically having bacterial pneumonia than a patient who has lung inflammation from inhaling a peanut (missing cause).

Rehder and Hastie's study (1997). Rehder and Hastie (1997) posed what seemed to challenge our causal status hypothesis, reporting that a cause feature in a

common-effect structure was less central than its effect. Moreover, our argument in the previous section might not appear to hold in this case because they equated the category validities of features in their training exemplars. Thus, this study merits a detailed discussion in the current chapter. Readers who are not interested in the details of this issue can safely skip to the next section.

Participants in Rehder and Hastie learned novel categories' features on four dimensions (A1, A2, A3, and A4 in Table 1), each with two values (0 and 1 in Table 1). Prototype values were 1's. In the Common-Effect condition, 1's on A1, A2, and A3 were described to cause 1 on A4 as in Figure 2. Participants learned a novel category by observing training exemplars where 1's occurred equally frequently across the four dimensions. In the test phase, participants rated the category membership likelihood of 16 exemplars which were all possible combinations of binary values on the 4 dimensions. (Table 1 lists 6 of the 16 categorization test exemplars in pairs of missing-cause and missing-effect items with 3, 2, or 1 prototype values.) Their regression analysis, performed on average participants' ratings on category membership, showed that in the Common-Effect condition the weight associated with A4 (i.e., effect) was higher than that associated with each of A1, A2, and A3 (i.e., cause), presumably contradicting the causal status hypothesis.

Insert Table 1 about here.

It should be noted that in Rehder and Hastie (1997) there is no mention of whether the participants were informed that the exemplars they saw during training were

representative samples of each category. These would have been critical instructions to add, because the observed category validities of features during training conflicted with the category validities that one would expect from the common-effect structure. More specifically, we argue that due to the common-effect structure, participants expected that the effect feature should have a higher category validity than each of the causes, even if they happened to observe the same category validities across all features. That is, if they believe that A1 can cause A4, A2 can cause A4, and A3 can cause A4, it is reasonable to expect that A4 would be more frequent than A1, A2, and A3, and that they might have accidentally seen a subset of category members with a different distribution. Without specific instructions that the training exemplars were representative of the entire category, participants could have discounted evidence in favor of their expectations derived from causal background knowledge. In the following experiment we tested whether participants indeed have different expectations about category validities of cause and effect features in a common-effect structure.

For the sake of simplicity, we presented only the three pairs of missing-cause and missing-effect exemplars in Table 1, because they represent all 16 test exemplars used by Rehder and Hastie, assuming equal salience of dimensions without the causal background knowledge (which was ensured through various measures they took). For instance, although we did not use 1011 and 1101, they are fundamentally redundant with 0111 because all three are missing-cause items with three prototype values.

As in Rehder and Hastie (1997), we first taught participants the common-effect causal background knowledge and then measured the likelihood of category membership for each of the 6 test exemplars. In addition, we included an inference task which

measured the likelihood of a non-prototype value (Value 0) for the missing dimension, given values on other dimensions as specified in Table 1 for the inference task. In this way, we measured the participants' expected category validity of value 0 on each of the dimensions, given the causal background knowledge of the common-effect structure.

The prediction was that measures from the categorization task would correlate with the measures of the inference task. For instance, Rehder and Hastie found that 1110 (missing-effect) yielded lower ratings from the categorization task than 0111 (missing-cause), and we argue that this occurred because given the common-effect structure, A4=0 in 111? is expected to be less likely than A1=0 in ?111. People would judge A4=0 in 111? to be very unlikely because all the causes have a value of 1. In contrast, A1=0 in ?111 is somewhat likely because there is no reason to expect 1 on A1. Similar predictions are made for 1100 and 0011.

However, 1000 and 0001 would be expected to lead to the opposite result. That is, A4=0 in 100? would be somewhat unlikely given that A1=1 and is expected to cause A4=1, but given that A2 and A3=0, it is not unreasonable to predict A4=0. In contrast, A1=0 in ?001 would be very unlikely because neither A2 nor A3=1, so A1 must have a value of 1 to account for the fact that A4=1. That is, the category validity expected for a value of 0 on the missing dimension is lower in the missing-cause item (?001) than in the missing-effect item (100?). Indeed, Rehder and Hastie (Rehder, personal communication, September 22, 1997) found that the categorization likelihood judgments for 0001 (missing-cause; M = 20.7) was lower than that for 1000 (missing-effect; M = 29.5), which runs counter to their results from the overall regression analyses.

The results of our experiment supported our predictions, as shown in Table 1 under "Mean" ratings. First, the categorization task results replicated the results from Rehder and Hastie's study. That is, missing-effect items led to lower categorization judgments than missing-cause items only from the 1110 vs. 0111 and 1100 vs. 0011 pairs. For the 1000 vs. 0001 pair, the missing-cause item led to lower categorization judgments than the missing-effect item. More importantly, the directions of these results are also mirrored by the corresponding inference problem results. The correlation between the average likelihood judgments for the six categorization task problems and the six inference task problems was significant (r_s=.83; p<.04).

This experiment thereby provides support for the hypothesis that in a common-effect structure, missing effect features presumably mattered more than missing cause features in Rehder and Hastie (1997) because having a non-characteristic value for the effect dimension in their stimulus materials is less likely than having a non-characteristic value for cause dimensions. Although the training exemplars equated for the category validities of cause and effect features in their materials, the causal relations among features were structured in such a way that effect features are more likely to be expected to occur in most exemplars (as shown in the inference task of this study). Based on this imbalance between the effect and the cause features, we submit that their results do not necessarily contradict the causal status hypothesis.

Plausibility of causal beliefs

Thus far, we have not considered how the causal status effect might be moderated by how plausible people find the causal relations in the first place. We believe that plausibility does moderate the causal status effect, such that low plausibility reduces the

effect. In general, the causal beliefs that people have about concepts come in various strengths. People might strongly believe, for instance, that bottles have a bottom so that they can contain liquid, or that robins are robins because they were born from another robin. However, that Echinacea contains a substance that bolsters the human immune system, or that yellow furniture increases children's emotional intelligence might be implausible to some (or most) people. It seems reasonable to expect that the amount of the causal status effect would be moderated by the degree to which the causal relations among features are plausible. That is, the more plausible the causal relations between two features are, the stronger the causal status effect should be from these two features.

From among many possible ways of manipulating causal plausibility, Ahn et al. (in press, Experiment 6) chose to examine the effect of compatibility between old and new causal background knowledge. Most people would believe that viral infections or genetic defects cause symptoms in disorders, rather than the other way around. Because of this old causal background knowledge, the statement that "Virus XB12 causes a low insulin level" is more plausible than the statement that "a low insulin level causes infection of Virus XB12."[6] The stimulus materials that conformed to laypeople's existing knowledge were developed for the Canonical condition (similar to the first example in the above), and the ones that conflicted with it were developed for the Reverse condition (similar to the second example in the above). Indeed, the results showed that the causal status effect was much more pronounced in the Canonical condition than in the Reverse condition.

[6] There are exceptions to this (e.g., AIDS), but in general cases, this difference in plausibility between the two types of items seems to hold up, as was confirmed in Ahn et al.'s (in press) pilot study.

This effect of the plausibility of causal relations on the causal status effect suggest one interesting possible difference between experts' and novices' categorization in their domain of expertise. Novices who have just completed initial training in a certain domain might not be confident in their background knowledge. They might not yet have internalized newly learned theories, and they have not had the chance to develop their own theories. Experts, on the other hand, might have a better understanding of the underlying mechanisms of the concept, which has been shown to be an important factor in increasing the plausibility of causal relations (e.g., Ahn, Kalish, Medin, & Gelman, 1995). In addition, experts, after years of practicing in their domain, might have developed their own idiosyncratic theories about, for instance, how symptoms are causally related or how different components of a machine interact. If so, the causal status effect might be more pronounced in experts than in novices. We are currently testing this issue in the domain of psychological disorders. (See the section entitled "Application" for more detail.)

Domain generality

Another potential moderating factor for the causal status effect is domain. Schwartz (1978) argues that nominal kinds, such as "white things," are different from natural kinds, such as "gold," in that nominal kinds do not have essences. According to him, nominal kinds are conventionally established, and therefore, if the criterial properties change, it no longer belongs to the category. For instance, if an object is stained with mud, it no longer belongs to the category of "white things." In support of the idea of domain differences in essentialism, Diesendruck and Gelman (1999) found more essentialistic, all-or-none categorization of natural kinds than of artifacts. Thus, if

psychological essentialism is responsible for the causal status effect, it might be predicted that a stronger causal status effect would occur for natural kinds than for nominal kinds.

The results obtained to date, however, are somewhat at variance with that prediction because no significant differences for the causal status effect have been found across various domains, including diseases and symptoms, artifacts, natural kinds, and social situations (Ahn, 1998; Ahn et al., in press). However, these studies did not systematically test the issue of domain generality by using randomly selected categories and features.

In particular, domain differences might occur when people create new categories for the following reason. Causal chains can be infinite, and an explanation must stop at some point. For natural kinds, this point might be their essences, or the deepest cause known to people. For nominal kinds, it is up to humans to select a stopping point based on their goal in creating these categories.

Thus, for categories that people treat as natural kinds (i.e., kinds that naturally occur in the world), we would expect people to create new categories as deeper underlying causes are revealed. For instance, to laypeople pneumonia might be a single disease, because laypeople do not know that there are multiple possible causes for it (thus, from a layperson's perspective, pneumonia does not have a common-effect structure.) Experts, on the other hand, do not treat pneumonia as a single disease, and partition pneumonia into more specific categories based on its different causes.

In contrast, in creating nominal kinds, one can choose any criteria that meet one's functional goal for the category, ignoring underlying causes. This is because nominal kinds are by definition conventionally fixed. For instance, the selection criteria for new

orchestra members usually focus on the quality of playing (an effect) rather than the details of how the person was trained (a cause). The orchestra category is a nominal kind, and the creator of the category can therefore intentionally set up such criteria, in order to meet their goal of getting the best performers. Similarly, one might create an ad-hoc category based on a goal, such as, "things to take out of a house in case of fire", without further querying about why one might have such a goal.

Fast vs. slow judgments

All of the experiments showing support for the causal status hypothesis involved slow judgments, where participants responded at their own pace. It is still an open question as to whether the causal status effect would be manifested in a time-pressured task, and as to whether it can account for some of the time course differences found in the categorization literature, such as reaction time differences in responding to typical versus atypical instances.

There are reports of influence of background knowledge even under a very short response deadline (e.g., Lin & Murphy, 1997; Palmeri & Blalock, 2000). Lin and Murphy's (1997) studies are of particular relevance to the causal status hypothesis. Participants first learned the following background knowledge about novel categories. For participants in Group A, "tuks" were said to be animal-catching devices (Category A, henceforth), whereas for participants in Group B, they were said to be pesticide-spraying devices (Category B, henceforth). Although both groups saw identical pictures of the learning exemplars, features that were crucial to the function of Category A were different from features that were crucial to the function of Category B. After learning exemplars, participants were presented with, among various items, an item retaining the

crucial part of Category A while omitting that of category B (consistent A item), and an item retaining the crucial part of Category B while omitting that of Category A (consistent B item). They found that, in general, Group A responded positively to Consistent A items and negatively to Consistent B items whereas Group B showed the opposite pattern. Most interestingly, they found the same pattern of results even when the stimuli were presented only for 50 ms followed by a mask. Note that, consistent with the causal status hypothesis, the features that had a greater impact in this study were the ones that had causal relevance to the function of each category. One reason to be cautious about generalizing Lin and Murphy's study to the causal status hypothesis is that their task measured perceptual recognition of stimuli, rather than directly measuring conceptual centrality per se. Future research can further examine this issue.

Summary

In this section, we have reviewed some moderating factors for the causal status effect. In general, if a sufficient number of other centrality-determining factors are pitted against the causal status effect, it may be difficult to observe in experimental measures. These other determinants of centrality include category validity and cue validity (diagnosticity), in addition to conceptual centrality, which we have argued embodies the causal status hypothesis most closely. We proposed that in many cases, apparent counterexamples to the causal status effect are actually instantiations of the case in which other factors are overwhelmingly pitted against it. In a study of the common-effect causal structure, we demonstrated some direct evidence for this proposal. In addition, low plausibility of the causal relations appears to diminish the causal status effect. When creating categories, domain differences appear, such that natural kind categories may be

based on essences, whereas artifact categories are based on conventionally fixed, functional goals. Finally, we speculated that the causal status effect may not be affected much by time pressure, given related findings such as Lin and Murphy's (1997). Thus, this section has addressed concerns about counterexamples to the causal status effect and laid out some of its boundaries. In the next section, we take a step in the opposite direction, explaining what predictions the causal status hypothesis may make for features in a category that exist outside the main causal structure.

Causal vs. Isolated features

The causal status hypothesis concerns the difference between features within the same causal structure. What about the difference between the centrality of features that participate in a causal structure and the centrality of features that are not involved in any causal relationships with other features? For instance, consider four symptoms of a mental disorder, A, B, C, and D, where the first three symptoms form a causal chain (A→ B→C) and the last symptom, D, is isolated. Would D affect diagnosis as much as the most fundamental cause (A), the intermediate cause (B), or the most terminal effect (C)? Alternatively, could it be that D affects diagnosis even less than C? Previous studies testing the effect of causal status of features on categorization have not examined this issue, although it is important to do so because real-life concepts sometimes have isolated features. For instance, most people would find it difficult to explain why tires are black and how this feature is causally related to other features of a tire.

Some clues to answering this question come from the induction literature. According to Gentner's (1989) structure-mapping theory, relational features (statements taking two or more arguments; for instance, x is smaller than y) are more important than

attributes (statements taking only one argument; for instance, x is yellow) in analogical inference. For instance, in the analogy "an atom is like the solar system," attributes such as "yellow," "hot," and "massive" for the sun are not useful in making the analogy and are therefore discarded. However, relational features such as "more massive than," and "revolves around" can be used to draw the analogy that electrons revolve around the nucleus in an atom as planets revolve around the sun in a solar system. Lassaline (1996) provides evidence for Gentner's theory in category-based induction. In a similar vein, Billman and her colleagues have shown that it is easier to learn a rule that links two features when the link is part of a system of correlations than when it occurs in isolation (e.g., Billman, 1996; Billman & Knutson, 1996). Given these findings, we hypothesized that features related to other features would be given greater weight in categorization than isolated features.

In our experiment (Kim & Ahn, 1999), we developed 6 artificial mental disorders, each comprised of 6 symptoms taken from various DSM-IV (APA, 1994) disorders[7]. For example, the 6 symptoms of "Methinismus" were (1) intense fear of gaining weight, (2) recurrent unjustified suspicions of the spouse's infidelity, (3) making fake suicide attempts, (4) perfectionism that interferes with task completion, (5) unreasonably scorning authority, and (6) exaggerated startle response. No two symptoms within an artificial disorder were taken from the same DSM-IV (APA, 1994) disorder.

Each participant received 2 disorders in the First-Cause condition (the first three symptoms were causally connected), 2 in the last-cause condition (the last three

[7] We were particularly interested in the domain of psychological disorders because these studies were carried out as a part of a larger project to understand expert / novice differences in the domain of psychological disorders. See the section "Application" for preliminary data on this.

symptoms were causally connected), and 2 in the No-Cause condition (none of the symptoms was causally connected). When the symptoms were causally related, they also received a plausible explanation (e.g., "An intense fear of gaining weight causes these patients to have recurrent unjustified suspicions of their spouses' infidelity, because they fear having become fat and unattractive to their spouses."). When the symptoms were not causally related, the symptoms were given a common grouping factor in order to make the length and the saliency of symptoms equivalent to the causal conditions[8] (e.g., "At a recent conference on mental illnesses, a speaker talking about intense fears of gaining weight was praised because she remained collected when the microphone broke. At a recent conference on mental illnesses, a speaker talking about recurrent unjustified suspicions of the spouses' infidelity was praised because she had charisma. At a recent conference on mental illnesses, a speaker talking about making fake suicide attempts was praised because she handled the audience's questions well").

After receiving descriptions of six symptoms of each disorder, participants answered questions in the format of "if a patient is in all ways like a typical person with [disorder X] EXCEPT that he or she does NOT have [symptom Y], does the patient have [disorder X?]" on a scale of 0 (definitely no) to 100 (definitely yes)[9]. For clarity of presentation, we inverted these scores in Figure 3 so that the higher the number is, the more conceptually central the symptom is to the disorder.

---

[8] The same pattern of results were obtained when no common factor was given in the control condition (Kim & Ahn, in preparation).

[9] The background knowledge was available to participants during this task.

Insert Figure 3 about here.

Figure 3 summarizes the results. In the No-Cause condition, the six symptoms did not differ. However, when the same symptoms were causally related, the causal depth determined the membership likelihood (i.e., In the First-cause condition, A > B > C; In the Last-cause condition, D > E > F), replicating Experiment 1 in Ahn et al. (in press). More importantly, isolated symptoms were significantly less central than even the most terminal effect in a causal chain (i.e., in the First-cause condition, F < C; in the Last-Cause condition, C < F).

These results are comparable to findings in Wisniewski's study (1995). He taught participants pairs of artifact categories and asked them to classify novel exemplars, manipulating whether they were told the artifacts' functions or not during training. He found that when people knew an artifact's function, they were more likely to classify new exemplars based on the exemplar features' relevance to that function than when they did not know the artifact's function. That is, causally isolated features (i.e., features that did not have any functional relevance) were less central to the concept than causal features. From this perspective, Wisniewski's results might have been due to the causal status of features rather than something inherent in the relevance of function in categorizing artifacts.

More recently, Gopnik and Sobel (1999) showed that 2-, 3-, and 4-year-old children categorize objects on the basis of causal power rather than similarity in appearance. Children saw that an object labeled as a "blicket" would set off a machine. When asked to select another blicket, they were more likely to choose an object that also

set off the machine than an object similar in appearance. Because there was no clear causal connection between appearance and being able to set off the machine, these results demonstrate that children gave more weight to causal features than to isolated features (i.e., perceptual appearance) in categorization, consistent with our findings.

The finding that features participating in a relation are judged to be more central than isolated features have some implications for the model of Sloman et al. (1998). This model currently predicts that isolated features and terminal effect features will have the same weight, but this was not the case in the current study.

Application

We now describe a study that is applied to real-life concepts. In this study, we were interested in the domain of psychological disorders, especially because of its potential implications for the Diagnostic and Statistical Manual of Mental Disorders (4th ed.; APA, 1994).

One of the earliest known classification systems of mental disorders, espoused by Kraepelin (1913), required necessary and sufficient features for the diagnosis of a disorder. This is an approach analogous to the bygone classical view in categorization research. Over the past century, this classical-view taxonomy of mental disorders has given way to the DSM system. The widely accepted DSM-IV (APA, 1994) was intended to reflect the later prototypical view of categorization, which allows for some flexibility (Barlow & Durand, 1999). For example, the prototypical patient with Schizophrenia has five symptoms, but a presenting patient need only have two of those five symptoms for a diagnosis of the disorder. In addition, as in a standard prototype model of categorization, the DSM-IV represents disorders as lists of features that are not causally connected to

each other. For example, clinicians using DSM-IV (APA, 1994) criteria would diagnose someone with obsessive-compulsive personality disorder if the patient has any combination of 4 symptoms out of a list of 8. Indeed, Medin (1989) has suggested that "the DSM-III-R guidebook (APA, 1987) provides only a skeletal outline that is brought to life by theories and causal scenarios underlying and intertwined with the symptoms that comprise the diagnostic criteria (p. 1479)."

A more serious challenge to the DSM system was put forth by Follette and Houts (1996). Their position is that the DSM-IV (APA, 1994), which claims to be atheoretical, is inadequate because it fails to provide "an organizing theory that describes the fundamental principles underlying the taxonomy" (Follette & Houts, 1996; p. 1120). The DSM's purpose in not specifying underlying theories is to avoid battles between different theoretical schools as to which theories should be included or focused on in the manual. However, these authors argue that the advantage gained by such a solution is outweighed by disadvantages concerning the negative effect a taxonomy that is silent with respect to theory has on clinical research (Follette & Houts, 1996). Specifically, Follette and Houts (1996) contend that one major practical failure of the DSM system is that its lack of unifying theories makes research difficult and slow. They propose that the first step towards remedying this failing of the DSM is to strengthen those research programs based on multiple theories (Follette & Houts, 1996).

As the first step toward studying the nature of the theory-based categorization in the domain of psychological disorders, we tested laypeople's understanding of psychological disorders (Kim & Ahn, 1999). Unlike the assumption underlying the DSM system, laypersons seem to have theories about how the symptoms of a disorder are

causally connected. For instance, consider three of the DSM-IV (APA, 1994) diagnostic criteria for anorexia nervosa: "refuses to maintain minimal body weight," "intense fear of gaining weight or becoming fat," and "disturbance in the way in which one's body weight or shape is experienced." It is easy to imagine a layperson thinking that people who have these anorexia nervosa symptoms "refuse to maintain minimal body weight" *because* they have an "intense fear of gaining weight or becoming fat," and also that they have an "intense fear of gaining weight or becoming fat" *because* they have a "disturbance in the way in which one's body weight or shape is experienced." Furthermore, we hypothesized that laypeople's causal theories would influence the way they perceive the importance of symptoms in their conceptualization of psychological disorders, as predicted by the causal status hypothesis.

In this experiment, we selected four mental disorders taken directly from the DSM-IV (APA, 1994) as stimuli. They included two Axis I clinical disorders (Anorexia Nervosa and Major Depressive Disorder), and two Axis II personality disorders (Narcissistic Personality Disorder and Obsessive-Compulsive Personality Disorder). The task was divided into 2 parts, a causal centrality task in which participants drew causal relations among symptoms within each disorder and assigned causal strengths to each relation, and a conceptual centrality task in which participants rated the importance of symptoms in each category. We hypothesized that each symptom is conceptually central to the extent that other symptoms are dependent on it (that is, to the extent that it is causally central). We will now describe the methods in more detail.

A list of "criterial" symptoms and "characteristic" symptoms was compiled for each of the four selected disorders. The criterial symptoms were taken from the list of

diagnostic criteria for each disorder described in DSM-IV. We also selected additional characteristic symptoms that were not considered to be diagnostic criteria but were described as characteristics in the DSM-IV (APA, 1994).

For each criterial symptom, a question measuring conceptual centrality was developed, in the format of "if a patient is in all ways like a typical person with X EXCEPT that he or she does NOT have the symptom of Y, does the patient have X?" where X is one of the four mental disorders and Y is a symptom. The participant's answer was collected on a rating scale of 0 (definitely no) to 100 (definitely yes). A total of 30 such questions, presented in random order and blocked by disorder, served as the conceptual centrality task. For clarity of presentation, we inverted these ratings in the analysis reported below so that the higher the number is, the more conceptually central the symptom is to the disorder.

For the causal centrality task, participants were presented with both the criterial and characteristic symptoms[10], and were asked to draw arrows indicating causal relations among them. They then assigned numbers to each arrow indicating the strength of that causal relationship, on a scale of 1-5 (where 1 means "X very weakly causes Y" and 5 means "X very strongly causes Y"). The order of the conceptual centrality task and the causal centrality task was counterbalanced across the participants.

The results showed that in general, the more causally central a symptom was, the more conceptually central it was. Figure 4 shows an example for the disorder anorexia nervosa. This figure presents averaged causal strengths among symptoms. (For simplicity

[10] We included the characteristic symptoms so that participants' causal structures could be measured as comprehensively and as accurately as possible (e.g., if a characteristic symptom X were not included, a criterial symptom Z could erroneously be measured as causally peripheral when, in fact, participants believed it to cause characteristic symptom X).

of presentation, causal strengths lower than 1.0 are omitted from the figure.) Arrows point toward effect symptoms, and the symptoms circled with thicker lines are criterial symptoms. This figure also reports the mean conceptual centrality ratings, shown by the numbers within the circles. One interesting result to notice is that even among criterial symptoms, causal centrality seems highly variable. For instance, on average, participants believed that in anorexia nervosa, "fear of being fat even when underweight" causes many symptoms, including fear of eating in public, bingeing and purging, excessive dieting, and refusal to gain weight. However, "absence of the period (in women) for 3+ menstrual cycles," another criterial symptom for Anorexia nervosa, was not judged to cause any other symptoms in that disorder. More importantly, note that a conceptually central symptom (e.g., "fear of being fat even when underweight" in anorexia nervosa) is also causally central, and a conceptually peripheral symptom (e.g., "absence of the period (in women) for 3+ menstrual cycles") is also causally peripheral.

We quantified this result by implementing Sloman et al.'s (1998) model, described earlier. To implement the model, we derived the predictions of Equation (1) based on the ratings obtained in the causal centrality task and compared them with the participants' conceptual centrality ratings obtained from the experiment. A pair-wise dependency matrix for each participant and disorder was first determined from their responses in the causal centrality task. That is, the strengths participants assigned to the causal arrows constituted the cells of the matrix. For each disorder, the matrices were averaged over all participants to yield a single matrix. Model-predicted conceptual centrality ratings were set to the initial arbitrary value of 0.5, following the procedure of Sloman et al. (1998). The matrix multiplication was performed repetitively until the

Spearman rank correlation of the model-predicted conceptual centrality ratings and the conceptual centrality ratings given directly by the participants converged to its terminal stable value. Rank correlations between the predicted and actual conceptual centrality ratings for each feature showed that the two factors were indeed positively correlated r_s = .73.

Thus, this experiment presents a number of important findings concerning laypersons' conceptual representations of mental disorders. First, laypersons seem to have richly structured causal theories about mental disorders, as shown in Figure 4. Participants were specifically told that they could leave the sheet blank if there were no causal relations among symptoms, but they spontaneously draw rich causal connections among symptoms. Second, this experiment presents the first demonstration of how naïve theories on mental disorders correlate with laypersons' diagnoses. In accord with our hypotheses, symptoms that undergraduates rated as being causally central were also conceptually central.

Possible implications for DSM-IV

We have shown that in laypeople at least, diagnosis is influenced heavily by causal theories. Therefore, another question that necessarily arises from an investigation of causal theories and diagnosis is whether the same issue applies also to clinicians. Elstein (1988), in his review of research on the cognitive processes involved in clinical inference and diagnosis, points out that clinicians do not, in general, make diagnoses by matching all the symptoms of the client to the lists of criteria in the DSM. Instead, clinicians are much more likely to devote most of their attention to several symptoms thought to be "prototypical" of a specific disorder (Elstein, 1988). Cantor, Smith, French,

and Mezzich (1980) have also argued that clinicians' reasoning adheres most closely to this prototype view of categorization. Our speculation is that the extra weight given to these "prototypical" symptoms is due in part to their place in the causal structure of clinicians' theories. Such a link between causal theories and diagnosis in clinicians seems likely. Indeed, Einhorn (1988) suggested that "cues-to-causality," or cues that indicate probable causal relations, may be used by clinicians as constraints to hypothesis construction (p. 53). That is, the clinician may use these cues to narrow down possible causal scenarios for the yet-undiagnosed disorder, which at the same time narrows down the set of possible diagnoses.

We currently have a project under way to investigate the relation between causal theories and diagnosis for clinicians. If it turns out that clinicians also show a strong causal status effect, it may shed some light on a well-known problem in DSM-IV (APA, 1994) diagnosis—in particular, the reliability problem in diagnosing personality disorders. The DSM-IV (APA, 1994) specifies the more important symptoms for depression and other clinical disorders. For instance, it specifies that more weight should be given to "depressed mood" and "lack of pleasure" than to the other 7 symptoms of a major depressive episode. For these disorders, there is high reliability of diagnosis between clinicians. However, the DSM-IV (APA, 1994) does not make such specifications for the personality disorders, and it is for these disorders that reliability of diagnosis is notoriously low. We suspect, given the findings in laypersons reported here, that clinicians may simply be administering their own weights to symptoms in diagnosis according to their personal theories of the personality disorders. We do not necessarily expect that this will be the whole answer to the problem of reliability in diagnosing

personality disorders, but we do believe that it is a potentially important factor deserving further study. Future studies of the effects of clinicians' causal theories on their diagnoses may have serious ramifications for this aspect of the DSM system.

Other types of dependency relations

As a final issue, this section discusses the research strategy we took, namely, focusing only on the causal relations among features in concepts. Certainly, concepts consist of relations other than causal relations. We have focused on the causal relations because they are considered to be the most critical part of theory representation (Carey, 1985; Gelman & Kalish, 1993; Wellman, 1990). How much explanatory power do we lose because of this? We will discuss this issue by analyzing both bi-directional and directional relations among features.

Bi-directional Relations

Previous research has found that correlations are often noticed and used in categorization tasks (Medin, Altom, Edelson, & Freko, 1982; but see Murphy & Wisniewski, 1989). However, Malt and Smith (1984) presented the finding that for many real-world categories, properties that are correlated in participants' feature listings are not always noticed as such. Furthermore, they reported that these unnoticed correlations do not affect categorization. In part 1 of their Experiment 1, they compiled from participants a list of properties that are generally true of various exemplars of bird, furniture, tree, fruit, and flower. These were then used as the input to the correlation analyses. Across all the categories, an average of 33.5% of the property pairs were statistically reliably correlated. The property correlations obtained in this experiment were statistical co-occurrences, and were not necessarily reflected in people's actual representations of those

concepts. For instance, “eats insects” and “sings” were found to be correlated in the bird concept. Indeed, Malt and Smith found that breaking such correlations did not affect typicality judgments. Thus, not all correlations mattered in categorization.

In Malt and Smith’s Experiment 2, participants determined whether each of the correlated property pairs from Experiment 1 was perceived to have a relationship. Of the original pairs, 33% were explicitly rated as correlated. When they manipulated these explicit correlations, Malt and Smith (1984) found that the correlations now determined participants’ typicality ratings. Specifically, the exemplars with the correlations intact were judged to be more typical than the exemplars with the correlations broken.

The question that Kevin Lee and the first author of this chapter raised was why only a subset of the correlations originally found in Malt and Smith's (1984) Experiment 1 were perceived to be correlated, and affected typicality judgments. We hypothesized that the property pairs that participants explicitly rated as belonging together were noticed because of a perceived dependency relation between them. For example, people may have rated “eats fish” and “is near ocean” to go together for birds because eating fish depends on being near the ocean. On the other hand, we posited that “eats insects” and “sings” were not perceived as being correlated (even though they were, in fact, positively correlated) because people generally do not have an explanation for how those two features are directly related.

In our experiment, we presented each of the property pairs found to be correlated in Malt & Smith’s first experiment to 18 undergraduate students at Yale University. Pairs that were rated as correlated in part 1 of their Experiment 2 were labeled as ‘explicit’ correlations, while the remainder were labeled ‘implicit.’ For each property

pair, two statements were generated, one to reflect each direction of dependency (A→B and B→A henceforth). For example, using the properties "eats fish" and "lives near the ocean," one test item was, "For a bird, whether or not it eats fish depends on whether or not it lives near the ocean," and the other test item was, "For a bird, whether or not it lives near the ocean depends on whether or not it eats fish." Along with each item, a scale marked from 1 for "strongly disagree" to 9 for "strongly agree" was displayed.

The idea is that the more asymmetric the dependency relation is between the members of a correlated pair, the greater the absolute value of the difference between ratings on A→B and B→A would be. Indeed, the absolute value of the mean difference score between A→B and B→A was greater in the explicit pairs ($M = 2.0$) than in the implicit pairs ($M = 1.4$), $p < .01$. We also examined the maximum score from each property pair. The mean maximum score for the explicit pairs (5.20) was significantly greater than that for the implicit pairs (2.96), $p < .05$. These results indicate that explicit pairs had more asymmetric dependency relations associated with them than implicit pairs. These results supported our hypothesis that perceived theoretical relationships, such as dependency, might allow correlations between some property pairs to be noticed, subsequently affecting categorization.

Let us return to the issue of the explanatory power of the causal status hypothesis. In this section, we attempted to show that correlations that are actually noticed and influence categorization are, in fact, likely to be directional relations. This proposal is consistent with Murphy and Medin's suggestion (1985) that "people are not only sensitive to feature correlations, but they can deduce reasons for those correlations, based on their knowledge of the way the world works. Perhaps, then, the connection between

those features is not a simple link, but a whole causal explanation for how the two are related (p. 300)." (See also Murphy & Wisniewski, 1989.) Thus, by focusing only on directional relations, the causal status hypothesis does not seem to be too limited in accounting for feature centrality. We will now examine cases with directional relations.

Directional Relations

We are not aware of any empirical studies that have systematically examined all possible types of directional relations among features in people's conceptual representations. Thus, we will discuss two types of directional relations (other than a causal relation) that were examined in Sloman et al. (1998). In this study, participants were explicitly told that symptom A, for example, does not cause symptom B but that symptom B follows symptom A (temporal dependency), or that the presence of symptom B depends on the presence of symptom A (contingency; e.g., the presence of a moustache is contingent upon the presence of a mouth). The results showed that temporally preceding features or features on which other features are dependent were judged to be more conceptually central. What are the implications of these results for the causal status effect? There are at least three possibilities.

First, these results might have occurred because participants imposed complex causal interpretations on the temporal dependency and contingency relations (e.g., symptom A might not directly cause symptom B, as was specified, but it might indirectly cause symptom B). In this case, the effect of dependency structure can be viewed as a special case of the causal status effect.

Second, it might be that these "non-causal" relations are in fact, a type of causal relations in a loose sense. For instance, in a restaurant script (e.g., Schank & Abelson,

1977), there is a sequence of scripted events, such as "customer goes into restaurant," "customer goes to the table and sits down," "customer picks up menu," and so on. These events are related in a temporal sequence, and it is difficult to say that one event caused another event in a strict sense. For instance, a customer's going into the restaurant did not make that customer sit down at the table. However, in a loose sense, these temporally related actions are causally linked in that changes in one may lead to changes in another. That is, a customer's going into the restaurant *allowed* the customer to sit down at the table. Sloman et al. also examined a "contingency" relation, such that the presence of a moustache is contingent upon the presence of a mouth. Again, it is difficult to say that a mouth caused a moustache to be present. But in a loose sense, there is a causal relation in that a mouth allowed a moustache to be present. If so, the causal status effect equals the effect of dependency relations, and Sloman et al.'s findings thereby suggest that there are no critical differences between at least some different types of causal relations.

Third, it could be that the causal status effect is a special case of a more general phenomenon occurring in any kind of asymmetrical dependency structure. This last possibility does not threaten the present claim that cause features are more central than their effect features, because the key ideas converge. Furthermore, even if the causal status effect is a special case of a more general phenomenon, it nonetheless appears to be a major portion of that general phenomenon, as indicated by studies showing that causal relations alone can account for a large amount of variance in feature centrality for natural categories (e.g., Ahn, 1998; Kim & Ahn, 1999). Indeed, causal relations are prevalent and serve as essential components of relations that features have in our conceptual representations (e.g., Carey, 1985; Wellman, 1990).

Conclusion

In this chapter, we have provided one mechanism by which feature centrality is determined. Our causal status hypothesis, derived from the idea of psychological essentialism, proposes that people regard cause features as more important and essential than effect features in that cause features affect category membership decisions more than effect features do. We reported several empirical studies that robustly support these main predictions of the causal status hypothesis, and reviewed a number of related phenomena that it can account for. We presented one computational model that implements the main gist of the causal status hypothesis, and reviewed direct empirical evidence for the model's validity. Boundary conditions of the causal status effect were stated and evidence for them reviewed. Empirical evidence modifying the predictions of the causal status hypothesis for feature weighting of causal versus isolated features was also discussed. A study was presented that successfully applied the causal status hypothesis to a class of real-world concepts, that of mental disorders. Finally, we provided a rationale for our decision to study only causal relations with respect to feature centrality, among the many types of possible dependency relations.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. Cognition, 69, 135-178.

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

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (in press). Causal status as a determinant of feature centrality. Cognitive Psychology.

Ahn, W., & Sloman, S. (1997). Distinguishing name centrality from conceptual centrality Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Associates, NJ: Mahwah.1-7.

American Lung Association (1998). Pneumonia [On-line]. Available: http://www.lungusa.org/diseases/lungpneumoni.html.

American Psychological Association. (1980). Diagnostic and statistical manual of mental disorders (3rd ed.). Washington, DC: Author.

American Psychological Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

Barlow, D. H., & Durand, V. M. (1999). Abnormal psychology (2nd ed.). Pacific Grove, CA: Brooks/Cole.

Barton, M. E., & Komatsu, L. K. (1989). Defining features of natural kinds and artifacts. Journal of Psycholinguistic Research, 18, 433-447.

Billman, D. (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. Language & Cognitive Processes, 4,

127-155.

Billman , D., & Knutson, J. (1996) Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. Journal of Experimental Psychology: Learning, Memory, & Cognition, 22, 458-475.

Bloom, P. (1996). Intention, history, and artifact concepts. Cognition, 60, 1-29.

Bloom, P., & Markson, L. (1998). Intention and analogy in children's naming of pictorial representations. Psychological Science, 9, 200-204.

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. Cognition, 59, 247-274

Cantor, N., Smith, E. E., French, R., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. Journal of Abnormal Psychology, 89, 181-193.

Carey, S. (1985). Conceptual change in childhood. Cambridge, MA: Plenum.

Chi, M. R., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Corter, J. E, & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. Psychological Bulletin. 111, 291-303.

Diesendruck & Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. Psychonomic Bulletin & Review, 6, 338-346.

Einhorn, H. (1988). Diagnosis and causality in clinical and statistical prediction. In D. C. Turk & P. Salovey (Eds.), Reasoning, Inference, and Judgment in Clinical Psychology (pp. 51-70). New York: Free Press.

Elstein, A. S. (1988). Cognitive processes in clinical inference and decision making. In D. C. Turk & P. Salovey (Eds.), Reasoning, Inference, and Judgment in Clinical Psychology (pp. 17-50). New York: Free Press.

Follette, W. C., & Houts, A. C. (1996). Models of scientific progress and the role of theory in taxonomy development: A case study of the DSM. Journal of Consulting and Clinical Psychology, 64, 1120-1132.

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

Gelman, S. A., & Ebeling, K. S. (1998). Shape and representational status in children's early naming. Cognition, 66, B35-B47.

Gelman, S. A., & Kalish, C. W. (1993). Categories and causality. In R. Pasnak & M. L. Howe (Eds.), Emerging themes in cognitive development (Vol. 2). New York, NY: Springer Verlag.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: the role of categories and appearances. Child Development, 58, 1532-1541.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. Cognition, 38, 213-244.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning . Cambridge: Cambridge University Press.

Gopnik, A., & Sobel, D. (1999). Detecting blickets: How young children use novel causal information in categorization and induction. Paper presented at the Society for Research in Child Development, April, Albuquerque.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. Child Development. 62, 499-516.

Kalish, C. W. (1995). Essentialism and graded membership in animal and artifact categories. Memory & Cognition, 23, 335-353.

Keil, F. C. (1991). The origins of an autonomous biology. In M. R. Gunnar, & M. Maratsos (Eds.), Modularity and constraints in language and cognition: The Minnesota symposia on child psychology, Vol. 25. (pp. 103-137). Hillsdale, NJ: Erlbaum.

Keil, F. C. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: The MIT Press.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: implications for hybrid models of the structure of knowledge. Cognition, 65, 103-135.

Kim, N. S., & Ahn, W. (1999, November). The influence of naïve causal theories on lay diagnoses of mental illnesses. Poster session presented at the 40th annual meeting of the Psychonomic Society, Los Angeles, CA.

King, S. D. (1993). Babe; The gallant pig. Crown Publishers Inc.

Kraepelin, E. (1913). Psychiatry: A textbook. Leipzig: Barth.

Kripke, S. (1971). Naming and necessity. In D. Davidson & Harman (Eds.), Semantics of natural language . Dordrecht: D. Reidel.

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

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive Development, 3, 299-321.

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

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. Journal of Experimental Psychology: Human Perception and Performance, 23, 1153-1169.

Locke, J. (1894/1975). An essay concerning human understanding, Oxford University Press.

Malt, B. C. (1994). Water is not H_2O. Cognitive Psychology, 27, 41-70.

Malt, B. C., & Johnson, E. C. (1992). Do artifact concepts have cores? Journal of memory and language, 31, 195-217.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. Journal of Verbal Learning & Verbal Behavior, 23, 250-269.

Medin, D. L. (1989). Concepts and conceptual structure. American Psychologist, 12, 1469-1481.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. Journal of Experimental Psychology: Learning, Memory, & Cognition, 8, 37-50.

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.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. Cognitive Psychology, 20, 158-190.

Murphy, G. L. (1993). A rational theory of concepts. In G. V. Nakamura, R.

Taraban, & D. L. Medin (Eds.), The Psychology of Learning and Motivation (pp. 327-359), San Diego: Academic Press.

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

Murphy, G. L., & Wisniewski, E.J. (1989). Feature correlations in conceptual representations. In G. Tiberghien (Ed.). *Advances in cognitive science: Vol. 2. Theory and applications* (pp. 23-45). Chichester, England: Ellis Horwood.

Palmeri, T. J., & Blalock, C. (2000). The role of background knowledge in speeded perceptual categorization. Manuscript under review.

Putnam, H. (1977). Is semantics possible? In S. P. Schwartz (Ed.), Naming, necessity, and natural kinds . Ithaca, NY: Cornell University Press.

Rehder, B., & Hastie, R. (1997). The roles of causes and effects in categorization. Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Associates, NJ: Mahwah, 650-655.

Rips, L. J. (1989). Similarity, typicality, and categorization. In A. O. Stella Vosniadou (Ed.), Similarity and analogical reasoning. (pp. 21-59). Cambridge University Press: New York.

Roberson, D., Davidoff, J., & Davies, I. (1999). Are color categories universal? New evidence from a traditional culture. Paper presented at the 40th annual meeting of the Psychonomic Society, Los Angeles, CA.

Rosch, E. (1978). principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.), Cognition and categorization (pp. 27-47), Hillsdale, NJ: Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal

structure of categories. Cognitive Psychology, 7, 573-605.

Ross, B. H. (1999) Postclassification category use: The effects of learning to use categories after learning to classify. Journal of Experimental Psychology: Learning, Memory, & Cognition. 25, 743-757.

Ross, B. H. (1997). The use of categories affects classification. Journal of Memory & Language. 37, 240-267.

Ross, B. H. (1996). Category representations and the effects of interacting with instances. Journal of Experimental Psychology: Learning, Memory, & Cognition. 22, 1249-1265.

Schank, R. C., & Abelson, R. (1977). Scripts, plans, goals, and understanding. Hillsdale, NJ: Erlbaum.

Schwartz, S. P. (1979). Natural kind terms. Cognition, 7, 301-315.

Sloman, S. A., & Ahn, W. (1999). Feature centrality: Naming versus Imaging. Memory & Cognition. 27, 526-537

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. Cognitive Science, 22, 189-228.

Soja, N. N., Carey, S., Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. Cognition, 38, 179-211.

Stevens. M. (2000). The essentialist aspect of naïve theories. Cognition, 74, 149-175.

Tversky, A. (1977). Features of similarity. Psychological Review, 84(4), 327-352.

Wellman, H. M. (1990). The child's theory of mind. Cambridge, MA: MIT Press.

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 449-468.

Acknowledgments

We would like to thank Charles Kalish, Douglas Medin, and Thomas Palmeri for their extremely helpful comments on the earlier draft of this article. This project was supported by a National Institute of Mental Health Grant (RO1 MH57737) awarded to Woo-kyoung Ahn, and a National Science Foundation Graduate Research Fellowship awarded to Nancy S. Kim.

Correspondence concerning this article should be sent to Woo-kyoung Ahn, Department of Psychology, Vanderbilt University, 301 Wilson Hall, Nashville, TN 37240, woo-kyoung.ahn@vanderbilt.edu.

	Missing-Effect Items					Missing-Cause Items				
Task	A1	A2	A3	A4	Mean	A1	A2	A3	A4	Mean
Categorization	1	1	1	0	2.87	0	1	1	1	6.65
	1	1	0	0	3.41	0	0	1	1	5.20
	1	0	0	0	3.71	0	0	0	1	2.09
Inference	1	1	1	?	2.85	?	1	1	1	4.35
	1	1	0	?	3.17	?	0	1	1	3.95
	1	0	0	?	4.56	?	0	0	1	2.72

Table 1. Items used in Rehder and Hastie (1997) and in our rebuttal experiment.

Note: Items for the Categorization task constitute a subset of items used in Rehder and Hastie's study (1997) and a complete set of items used in our experiment. The Inference task was used only in our experiment. The mean ratings are results from our experiment. In the common-effect condition, A4 was a common effect for A1, A2, and A3.

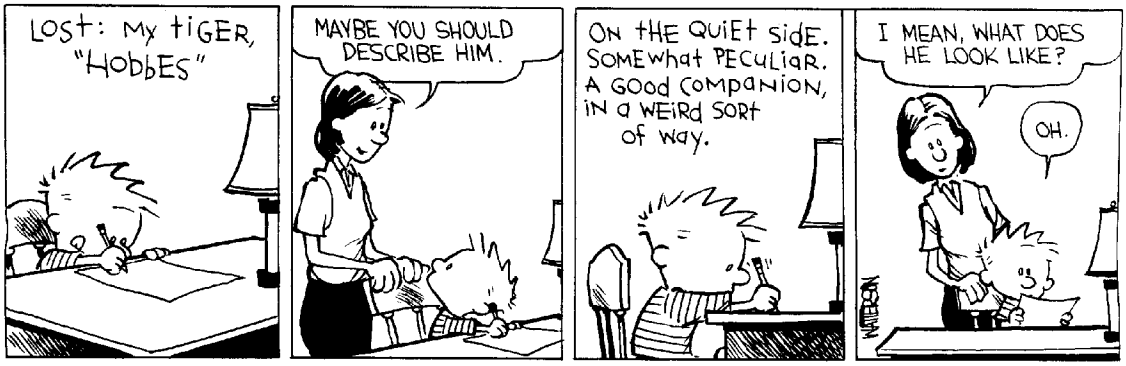

Figure 1. Calvin and Hobbes

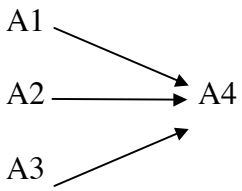

Figure 2. Common-Effect causal structure.

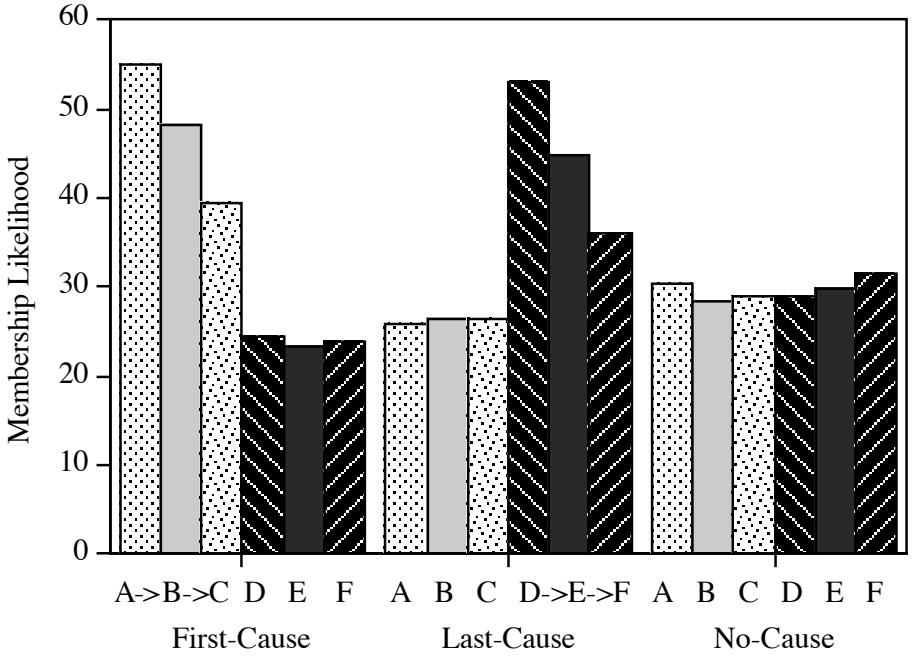

Figure 3. Results from Kim and Ahn (in preparation, Experiment 2)

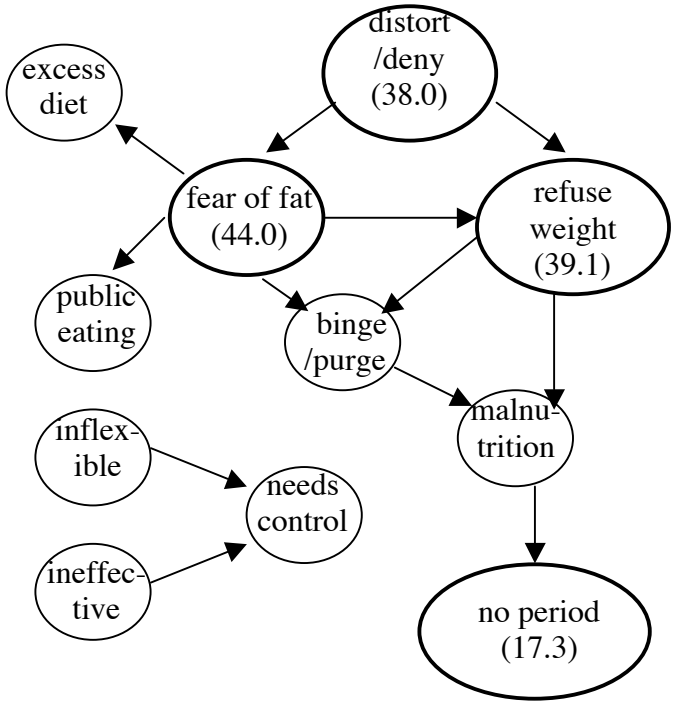

Figure 4. Laypeople's mean causal dependency structure for the disorder anorexia nervosa.

Note: “distort/deny” means “Disturbed experience of body shape or denial of the problem,” “fear of fat” means “Fear of being fat even when underweight,” “refuse weight” means “Refusal to maintain body weight at or above minimal levels,” and “no period” means “Absence of the period (in women) for 3 + menstrual cycles.”