Essentialism as a Generative Theory of Classification

Bob Rehder

Department of Psychology

New York University

Send all correspondence to:
Bob Rehder
Department of Psychology
6 Washington Place
New York, NY 10003
Phone: (212) 992-9586
Email: bob.rehder@nyu.edu
Essentialism as a Generative Theory of Classification

It is obvious that we classify the objects we encounter by their *appearance*, that is, by the particular features, aspects, or characteristics that they display. But, after a moment’s reflection it becomes clear that appearance is sometimes not all there is to it, that there are other factors not available to immediate inspection that might contribute an object’s identity. A study of Rips’ (1989) serves to illustrate. College students were told a story about a bird that had normal bird-like features (wings, ate seeds, lived in a nest in a tree, etc.) that was exposed to hazardous chemicals. As a result, the bird began to take on properties more characteristic of an insect: the wings with feathers were replaced with wings made of a transparent membrane; the nest was abandoned in favor of living on the underside of tree leaves; it a developed a brittle iridescent outer shell; and so on. When asked whether the animal was now a bird or an insect, most students judged that it was still a bird. The important point to note is that they made this decision despite the fact that the animal no longer *looked like* a bird at all—apparently, there is something to category membership more than just how an object appears. In fact, there is evidence that even children as young as three-years-old believe that the “insides” of objects are relevant to determining its class membership (Gelman & Wellman, 1991; also see Gelman, 2003; Keil, 1989).

The idea that some features might be especially important to category membership is not (to say the least) new. In philosophy, it dates at least as far back as Aristotle who distinguished between an entity’s *essential properties* (which determine *what* something is) from its *accidental properties* (which determine *how* it is, that is, what properties just happen to inhere in it). The idea that essential properties might be inaccessible to perception has an equally impressive legacy. Even as central a British Empiricist as John Locke distinguished *real essences* (what an object really is, which, according to the Locke, was unknowable in principle) from *nominal essences*, which
could be perceived and which formed the basis for everyday categorization. In psychology, even after the Roschian view that membership in natural categories is probabilistic, or graded, replaced the classical (i.e., defining features) view, the distinction between perceivable-but-only-characteristic and real-but-unobservable properties has persisted in various forms (although of course without Locke’s pessimism regarding the knowability of the latter). For example, Miller and Johnson-Laird (1976) distinguished between a categories’ core properties (which could be used during, e.g., reasoning), versus their identification procedures which inferred category membership on the basis of perceptual information.

But the separation of perceptual and core properties in this manner would seem to leave us without any elucidation of the relation between two. One is reminded of Descartes’ famous (non)solution to the mind-body problem which proved unworkable because it failed to specify how the two domains (mind and body) interacted. In the absence of any specification of how core and perceptual features interact, we are left, for example, with no principled explanation for why category membership will be decided by observable features in some (i.e., normal) conditions but by core features in others (e.g., those instantiated in Rips’ study).

This chapter presents a solution to the categorization field’s own mind-body problem, that is, how core and perceptual features interact. It does so by adopting a move that should be familiar from other fields of cognition. Namely, it will describe a generative theory of categories in which a category’s core knowledge is represented in such a way that it produces or generates the perceptual features which one might observe. But a unique characteristic of the current approach is that the relations between features will be defined in terms of generative causal relations. As will be shown, once the manner in which perceptual features are causally generated by core properties is specified, one can then “work backwards,” so to speak, to specify how perceptual information implies the presence of core properties, and hence category
membership.

To begin, consider in Figure 1 the hypothetical causal relations among features of one of the real-world categories used by Rips (1989), namely birds. The figure includes the observable features of birds, such as having wings, flying, building nests in tree, singing, and eating seeds. It also includes what might be considered a core property of birds, namely, bird DNA. Indeed, for many Western-educated adults bird DNA would seem to approximate a defining feature for birds, as an object with bird DNA is virtually certain to be a bird, and one without it is virtually certain to not be. But the dilemma is that whereas flying and eating seeds are features of an animal which are observable, bird DNA is not.

Fortunately, features are not all that we know about birds—we also know how their features are causally related. For example, we know that birds are able to fly because they have wings and have body weight that is small enough to be supported by those wings. We know that birds build nests in trees because they can fly (and also they are light enough to not break tree branches when they sit in them!). Finally, we (Western-educated adults) all believe that basic morphological and behavior properties like having weight, body size, singing, and eating seeds is (somehow, directly or indirectly) caused by the fact that birds have the right kind of genetic makeup and evolutionary history that lead birds to have bird features.

Hopefully, the claim regarding how core and perceptual properties interact in categorization is now clear. People of course usually classify objects on the basis of observable features (what else could they use?), but they use those observable features to infer the presence of more core properties, which are then taken as defining of (or at least more diagnostic of) category membership. In other words, the mental act of categorization can be viewed as a case of causal reasoning in which properties like weight, body size, singing, and eating seeds provide inferential support for properties like bird DNA. This account provides an explanation for why people typically use
observable features in classification, but also why they can override perceptual information in particular circumstances. When one is told a story about how a particular bird’s features are transformed through external intervention into those of an insect’s, one recognizes that the underlying core properties remain unchanged, and thus so does the animal’s category membership.

Accounting for data such as Rips’ is itself no small feat. But in fact a generative view of classification can gain a large increment in explanatory power, and in its ability to make unique predictions, by making one additional assumption, namely, that the causal relations linking core and observable properties can sometimes be probabilistic rather than deterministic in nature. To illustrate the importance of treating causality as a probabilistic relation, consider a second, somewhat simpler example of a category and its causal network in Figure 2. In Figure 2, D represents some disease, S₁ represents a symptom which is directly caused by that disease, S₂ is a symptom which is caused by S₁, and S₃ is a symptom which is caused by S₂. The disease and its three symptoms are assumed to be related by the three causal mechanisms depicted as diamonds and labeled M₁, M₂, and M₃ in Figure 2. On the one hand, if these causal mechanisms operate deterministically, then whenever the disease D is present, so too are the symptoms S₁, S₂, and S₃. In this case, S₁, S₂, and S₃ all provide equally good evidence for the presence of D (all else being equal). But if the causal mechanisms operate probabilistically instead, then D will produce S₁ with some probability less than one. Similarly, S₁ will produce S₂ with some probability less than one, but since S₁ doesn’t always accompany the disease, intuition suggests that S₂ will be produced even less often than S₁. The same argument applies to S₃, so the upshot is that \( 1 > P(S₁ | D) > P(S₂ | D) > P(S₃ | D) \). As a consequence, when causal relations are probabilistic, the generative view predicts that S₁ will serve as better evidence for the presence of D than S₂, which in turn will serve as better evidence for D than S₃ (all else being equal¹).

This example is important enough that the intuition behind it deserves to be
cashed out with some precision. First note that the representation of causal relations depicted in Figures 1 and 2 are examples of Bayesian networks. Bayesian networks consist of nodes that represent variables and directed edges that can be interpreted as representing direct causal relationships between variables (for details see Glymour, 1998; Jordan, 1999; and Pearl, 1988; 2000). By themselves Bayesian networks also convey no information regarding the details of the causal relationships that link variables (i.e., features) in a network. However, the causal relations in such networks can be assumed to take on a specific functional form, and the parameters of those functions can be specified. In Figure 2 there are a total of six parameters: \( m_1, m_2, m_3, b_1, b_2, \) and \( b_3 \). Parameters \( m_1, m_2, \) and \( m_3 \) represent the probability that the causal mechanisms \( M_1, M_2, \) and \( M_3 \), respectively, will successfully operate (that is, will bring about effects \( S_1, S_2, \) and \( S_3 \), respectively) when the cause feature is present. Parameters \( b_1, b_2, \) and \( b_3 \) represent the probability that symptom \( S_1, S_2, \) and \( S_3 \), respectively, is produced by some unspecified background cause.

Table 1 presents the likelihoods that the causal network of Figure 2 will generate symptoms \( S_1, S_2, \) and \( S_3 \). For example, the probability that \( S_1 \) is present when \( D \) is, \( P(S_1 | D) \), is the probability that it is caused by the causal mechanism \( M_1, m_1, \) or brought about by the background cause, \( b_1 \). Assuming independence, this probability is \( m_1 + b_1 - m_1 b_1 \). The probability that \( S_2 \) is present, \( P(S_2 | D) \), is the probability that it is caused by the causal mechanism \( M_2, m_2 P(S_1 | D) \), or brought about by the background cause \( b_2 \). This probability is \( m_2 P(S_1 | D) + b_2 - m_2 b_2 P(S_1 | D) \). And so on for \( P(S_3 | D) \). Table 2 also presents the probability that each feature will be generated by disease \( D \) when \( m_1 = m_2 = m_3 = 0.80 \) and \( b_1 = b_2 = b_3 = 0.20 \). These probabilities confirm the intuition that when causal relations are construed as probabilistic, symptoms that are causally “farther away” from their disease are generated with lower reliability.

In summary then, the solution offered to the categorization field’s mind/body problem is to assume that core and observable features are interconnected with
probabilistic causal relations, and as a result observable features serve as evidence for core properties as a function of the likelihood they are causally generated by those core features. Using a term first introduced by Waldmann, Holyoak, & Frattiane (1995), I refer to this theory as *causal-model theory*, although my application of this theory concerns classification rather than category learning (Rehder, 2003a; 2003b). It also differs in the functional form assumed to hold of causal relations. Whereas the representation of causal relations presented here is isomorphic to Cheng’s (1997) power-PC theory of causal induction, Waldmann et al. considered causal relations between continuous variables such that the level of one variable (the effect) changed as a linear function of another (the cause), as in structural equation models. This difference is not fundamental however, as the generative view can be applied to causal networks that include both continuous variables viewed as being as linearly related and binary (or ordinal) variables connected with the type of discrete causal mechanisms described above. Another extension of the generative view would be to causal mechanisms involving more than two variables. For example, I may not be exactly sure how birds fly, but I believe that the causal mechanism that produces flight somehow involves the lift produced by the wings overcoming the bird’s body weight; thus wings and body size (and the wing’s flapping motion, and the body’s aerodynamic shape, and so on) are all variables causally relevant to flight. The generative approach can be applied to cases involving three or more causally-related variables as easily as two.

In this chapter three experiments are presented that provide empirical support for the generative view of classification just described. To assess the kind of causal reasoning that occurs when a category has a defining or “essential” property, in Experiments 1-3 adult subjects are taught novel categories in which one of the category’s features is designated as occurring in 100% of members of the category, and in members of no other category. Note that the property will not be DNA or a disease. The purpose of these experiments is not to determine what the defining or essential
properties of categories really are, but rather to stipulate such a feature and then show how classification is influenced when it is causally linked to observable properties. Experiments 1-3 differ in terms of the exact network topology by which the defining feature is causally related to those observable properties.

These experiments will show classification can be viewed as causal reasoning in which observable features are taken as evidence for features which are unobservable but which are defining of category membership. However, although a generative theory of classification subsumes causal reasoning as a special case, its application is in fact not limited to categories with a defining feature. As I will argue in the Discussion, the extent to which a category is essentialized is a matter of degree, and an advantage of the generative view is that it can also accommodate categories which are essentialized only “partially”, or not at all. To this end, subjects in Experiment 4 are taught the same categories as in Experiments 1-3, with causal relations among observed features but without an unobservable defining feature. Their performance will be shown to correspond to the generative view: Objects will be judged to be good category members to the extent their features will likely to have been generated by category’s causal laws.

Experiment 1: Classification As Diagnosis

The purpose of the first experiment is to conduct a test of the example already presented in Figure 2. College undergraduates were instructed on novel categories in which one feature was described as defining but unobservable; the other features were observable and were related to the defining feature in a causal chain. The generative view predicts that the evidence that a feature provides for category membership should vary as function of its position in the causal chain, with the feature that is directly caused by the defining feature providing the strongest evidence and the most indirectly caused feature providing the weakest.

Table 2 presents an example of features and causal relationships for one of the
novel experimental categories, Lake Victoria Shrimp. Lake Victoria Shrimp were described to participants as possessing four binary features and three causal relationships among those features. One feature (which I will continue to refer to in the abstract as D because it is “defining”) was described as occurring 100% of the time in category members, and never in members of other categories. Features $F_1$, $F_2$, and $F_3$ were described as occurring in “most” category members. The causal links were arranged in a chain pattern: $D \rightarrow F_1 \rightarrow F_2 \rightarrow F_3$ (as in Figure 2). Each causal relationship consisted of one sentence indicating the cause and effect features (e.g., “A high quantity of ACh neurotransmitter causes a long-lasting flight response.”), and then one or two sentences briefly describing the causal mechanism (e.g., “The duration of the electrical signal to the muscles is longer because of the excess amount of neurotransmitter.”). The knowledge associated with categories such as Lake Victoria Shrimp was intended to be a simplified analog of real-world category knowledge, such as that bird DNA causes wings, which causes flying, which causes nests in trees, or a disease which causes a chain of symptoms. Participants in a control condition were provided with the identical category information except for the causal relations between features.

Participants learned one of six novel categories: two biological kinds (Kehoe Ants, Lake Victoria Shrimp), two nonliving natural kinds (Myastars, Meteoric Sodium Carbonate), and two artifacts (Romanian Rogos, Neptune Personal Computers). Participants first studied several computer screens of information about their assigned category, and were required to pass a multiple-choice test of this knowledge. They then performed a categorization task in which they rated the category membership of one of the three observable features ($F_1$, $F_2$, or $F_3$). For example, participants who learned about Lake Victoria Shrimp (Table 2), would be presented with a shrimp that was described as possessing a “Long-lasting flight response” and asked “Is this a Lake Victoria Shrimp?” Participants entered their rating by using the left and right arrow keys to move a bar along a response scale to a position which reflected their confidence that the exemplar
was a category member. The two ends of the scale were labeled "Definitely not an X" and "Definitely an X," where X was the name of the category. Responses were recorded as a number in the range 0-100. Thirty-six university undergraduates participated in this experiment, and were assigned in equal numbers to the causal and control conditions.

The category membership ratings for features $F_1$, $F_2$, and $F_3$ are presented in Figure 3 for both the causal and control conditions. Two things in Figure 3 should be noted. First, average ratings were significantly higher in the causal condition (56.7) as compared to the control condition (46.5), indicating that features provide stronger evidence for category membership when they are causally-linked to the defining feature. This finding was predicted by the generative view, because the presence of causal relations allows one to infer the presence of the defining feature one basis of observable features.

The second important result is that, in the causal condition, feature $F_1$ received a significantly higher rating (63.3) than feature $F_2$ (54.6) which received a higher rating than $F_3$ (52.3). (The difference between $F_2$ and $F_3$ did not reach significance.) This result was also as predicted, because the generative view assumes that features that are generated with greater reliability by the underlying defining feature (like $F_1$) should serve as greater evidence for the presence of that defining features than features which are generated less reliably ($F_3$). Apparently, when observed features are causally related to defining ones, categorizers can invoke a process of causal inference in which they work backwards from observables to defining properties, in the same way that one can backwards from a disease’s symptoms to the disease itself.

Experiment 2: Boundary Intensification

Experiment 1 instructed participants on one category and asked them to rate the likelihood of whether or not an object was a member of that category. But it will be
more often the case that there will usually be more than one category that an object may belong it. A generative view of classification can easily be applied to such situations by predicting that an object is a member of the category that is was most likely to be generated by. For example, in the context of an experiment in which an object (O) can belong to one of two categories (A or B), the probability that O is an A, \( P(O|A) \), can be expected to be a function of the relative probability that it was generated by A and B, that is, according to Bayes’ Law,

\[
P (A|O) = \frac{[P(O|A) / [P(O|A) + P(O|B)]] * P(A)}{P(O)}
\]  

In Experiment 2 participants are instructed on two categories and are asked to make a binary judgment regarding which category an object belongs to. As in Experiment 1, in a causal condition a category’s observable features will be causally related to a defining feature. However, in Experiment 2 those features are each directly caused by the defining feature, as shown in Figure 4. In Figure 4, category A and B have opposite values on the same stimulus dimensions. For example, some subjects were instructed on both Lake Victoria Shrimp, with the defining feature “A high quantity of ACh neurotransmitter,” and Madagascar River Shrimp with the defining feature “A low quantity of ACh neurotransmitter.” Similarly, features \( F_1 \) and \( \sim F_1 \) were opposing values on the same dimension (“Long-lasting flight response” vs. “Short-lasting flight response”), and so on for the remaining two stimulus dimensions. Participants in a control condition were provided with identical category information except that the causal relations between features were omitted.

The question addressed in Experiment 2 concerns how the presence of causal relations changes subject’s judgments regarding an object’s membership in one of two possible categories. Gelman (2003) has suggested that essentialized categories should exhibit the phenomenon of boundary intensification in which category boundaries become more extreme, or more dichotomous, than they would otherwise be. This effect is somewhat analogous to the categorical perception of speech sounds. For example, the
sounds \(d\) and \(t\) differ from one another on a single dimension (voice onset time). But when voice onset time of those sounds is varied experimentally along a continuum, they are nevertheless perceived “categorically” as either as a \(d\) or a \(t\) but rarely a blend of the two (Lisker & Abramson, 1979). Applied to the current experiment, “categorical perception” would work to make the category boundaries between categories \(A\) and \(B\) in Figure 4 more extreme because the observable features are causally linked to an underlying core property.

Once again, it is worthwhile to work out a concrete example. First, suppose that in the absence of causal knowledge category \(A\)’s three features (\(F_1, F_2,\) and \(F_3\)) are viewed as each occurring with probability 75% in members of \(A\), but they have no causal relations with defining feature \(D\). Similarly, category \(B\)’s three features (\(\neg F_1, \neg F_2,\) and \(\neg F_3\)) occur with probability 75% among members of \(B\), and are causally unrelated to \(\neg D\). This corresponds to \(m = 0\) and \(b = .75\) for both causal models Figure 4. If one then observes an object \(O\) with features \(F_1, F_2,\) and \(\neg F_3\), \(P(A|O) = (.75)(.75)(.25) = .141\) and \(P(B|O) = (.25)(.25)(.75) = .047\). Assuming that the two categories are equally probable beforehand (i.e., \(P(A) = .50\)), Bayes’ law tells us that the probability that \(O\) is an \(A\), \(P(A|O)\), is .75.

Now consider the case in which the categories’ features are thought to be generated by their defining features (\(D\) or \(\neg D\)) via causal mechanisms that operate with 50% reliability, that is, \(m = .50\). In this case, each individual feature is now generated with probability \(m + b - mb\), that is, .875 (assuming independence once again). If one again observes an object \(O\) with features \(F_1, F_2,\) and \(\neg F_3\) then \(P(A|O) = (.875)(.875)(.125) = .096\), \(P(B|O) = (.125)(.125)(.875) = .014\), and thus the probability that \(O\) is an \(A\) is now .875. That is, one should be more confident that the object is a category member with causal relations (.875) than without them (.75). The probability that some object \(O\) is a member of category \(A\) as a function of the number of \(A\) features it possesses is presented in Table 3 for the case where there are causal relations (\(m = .50\))
and when there are not \( (m = 0) \). Table 3 demonstrates the phenomenon of boundary intensification: For all objects, one’s confidence that it is a member of its most likely category increases when its features are linked to an underlying defining feature.

To test these predictions, Experiment 2’s categorization test presented objects with values on all three observed dimensions (as opposed to the single features presented in Experiment 1). Participants were asked to choose which of the two categories the object belonged to (e.g., Lake Victoria Shrimp or Madagascar River Shrimp), and then to also rate their confidence in that judgment. The eight possible objects that can be formed from three binary dimensions were each presented twice. Thirty-six undergraduates participated in this experiment, and each was instructed on one category pair with causal knowledge and another (control) pair without it (the order of presentation of these two pairs of categories was balanced).

The results are presented in Figure 5 for both the causal and control conditions. Figure 5 shows the probability that the object was classified as a member of category A as a function of the number A features it possessed. Of course, in both conditions subjects’ classifications were sensitive to how many features O possessed which were characteristic of A or B: It was likely to be classified as an A if it possessed mostly A features and as a B if it possessed mostly B features. But the figure shows that this effect was more pronounced when those features were described as causally related to the underlying defining features D or \( \sim D \). This effect manifested itself as a significant interaction between condition (causal vs. control) and number of features. The same pattern of results were reflected in confidence ratings.

These results demonstrate how the boundary between categories can become more extreme when features are causally related to an underlying defining property. According to a generative view of classification, causal relations constrain the generation of features, so that the category becomes more homogenous, and thus less accepting of exemplars that are discrepant with respect to those causal laws. The result
is a sharpening of category boundaries, in which objects are perceived more “categorically”.

Experiment 3: Classification As Prospective versus Diagnostic Reasoning

The first two experiments are examples of how classification can be taken to be a case of causal reasoning in which observable features are diagnostic of core properties. However, classification does not always involve reasoning *backward* (i.e., diagnostically), but sometimes involves reasoning *forward*, prospectively, to a core property. Consider the examples in Figure 6A which shows two causal networks each involving the HIV virus. In the left panel of Figure 6A, the category “being HIV positive” (which is itself not directly observable) can be diagnosed in term of the symptoms it produces, such as lymphoma, sarcoma, and pneumonia. However, the right panel of Figure 6A illustrates how HIV can also be inferred by reasoning forward from its possible causes, such as blood transfusions, intravenous drug use, or participating in unsafe sex. That is, to the extent that an individual has these properties, the likelihood that he or she has HIV increases.

Because the generative approach to classification subsumes causal reasoning as a basis for determining category membership, it applies equally well to both diagnostic and prospective reasoning. Moreover, the HIV example is well-suited to illustrating how classification can involve specifically *causal* reasoning, because it exemplifies the asymmetries that obtain when reasoning from multiple possible effects to a cause versus from multiple possible causes to an effect. To demonstrate these asymmetries, first consider the (unrealistic) case that each of the causal relations shown in Figure 6A are deterministically sufficient, that is, a cause produces its effect(s) with 100% reliability \((m = 1)\). The left panel of Figure 7A presents the probability that HIV will generate a given case \(O\) as a function of the number of HIV symptoms (lymphoma, sarcoma, or pneumonia) that it has. (I continue to use “\(O\)” to represent an “object” that
displays a set of features, in this case a patient with a set of symptoms.) Analogously, the right panel of Figure 7A presents the probability of HIV as a function of the number of its causes present. Figure 7A indicates that when causal relations are assumed to be deterministic, the probability of HIV is zero when any of its symptoms are absent (because if HIV is present, it should generate all its symptoms). In contrast, the probability of HIV is certain when any of its causes are present. This is the case because each cause generates HIV with certainty.

Figure 7B presents the more realistic situation in which the causal relations linking HIV with its causes and symptoms are probabilistic \( m < 1 \). Figure 7B presents the (logarithm) of the probability of HIV as a function of the number of symptoms (left panel) or causes (right panel) present. As a result of the causal relationships being probabilistic rather than deterministic, the probability of HIV now increases monotonically as the number of symptoms increases. Note that the utility of plotting these probabilities in log coordinates is that it demonstrates how evidence for HIV increases as a function of the number of symptoms: Because in the generative model evidence consists of individual probabilities which multiply, plotting the overall probability in log coordinates reveals an additive (i.e., linear) relationship (left panel of Figure 7B). In contrast, the right panel of Figure 7B reveals that the relationship between HIV and its causes in nonlinear, such that adding the first possible cause of HIV produces a larger increase in the probable presence of HIV than adding additional causes. Nevertheless, those causes do not invariably lead to HIV, so the presence of additional causes continues to increase the probability that HIV is present.

To test these predictions, participants in Experiment 3 were instructed on a single category (as in Experiment 1). One group of subjects were presented with the common cause structure shown in the left side of Figure 6B in which one feature (D) was the defining feature and was described as the cause of the three observable features. Another group was instructed on the common effect structure in the right side of Figure
6B in which D was the defining feature and was described as being caused by the three observable features. There was also two control groups which were identical to the common cause and common effect conditions, respectively, except for the presence of the three causal relationships. All groups then performed a categorization test which presented objects with values on all three observable dimensions. As in Experiment 1, participants were asked to rate how likely the object (e.g., a shrimp with a given set of features) was a member of the category (e.g., Lake Victoria Shrimp). The eight possible objects that can be formed from three binary dimensions were each presented twice. 144 undergraduates were assigned in equal numbers to the four conditions.

The results are presented in Figure 7C. To allow comparison with the predictions shown in Figure 7B, the logarithm of the categorization ratings have been taken. The left panel of Figure 7C indicates that, as predicted, in the common cause condition the (logarithm) of the ratings were a linear function of the number of effect features present. In contrast, in the common effect condition those ratings exhibited a nonlinearity in which the presence of one potential cause of D produced a larger increase in the ratings as compared to adding a second or third cause (right panel of Figure 7C).

Besides illustrating the predicted asymmetry between the common cause and common effect networks, the results also demonstrate how participants treated the causal relationships as probabilistic, because each additional cause (or effect) produced an increase in the probability that the defining feature D was present. In other words, the results correspond to the probabilistic predictions (Figure 7B) rather than the deterministic ones (Figure 7A). Also note that features in the common-cause and common-effect conditions were more diagnostic of category membership (each additional feature resulted in a larger increment in category membership) as compared to the control conditions. That is, as in the first two experiments, categorizers use causal knowledge to infer an underlying defining feature.
Experiment 4: Non-Essentialized Categories and Theoretical Coherence

The generative view of categorization presented here borrows much from the view of essentialism described by Medin and Ortony (1989). As here, Medin and Ortony observed that category membership is often based on unobservable properties. And, as here they proposed that underlying properties not only establish category membership, they "are best thought of as constraining or even generating properties the might turn out to be useful in identification." (p. 185). But by subsuming causal reasoning as a basis for determining category membership, Experiment 3 has just shown how the generative view can also account for cases in which an observable feature is the cause of, rather than caused by, an underlying property. Moreover, the generative view is not restricted to only essentialized categories. This is important, because it is likely that not all categories that human know are essentialized to the same extent (or at all).

For example, I have used diseases as a paradigm case of essentialized categories. However, the causes of many diseases were unknown at an earlier stage of scientific knowledge, and as a result the categories for diseases were initially organized around their characteristic features (i.e., symptoms) rather than an underlying cause. In addition, while research has shown that biological kinds may be strongly essentialized for adults (as Rips’, 1989, transformed bird illustrates) this may not be true of all individuals—for example, they may be less essentialized for young children (Keil, 1989). Finally, although underlying causal properties might be important for complex artifacts (e.g., automobiles, computers), simple artifacts like pencils and wastepaper baskets appear to be defined more in terms of their perceptual properties and/or function (although see Bloom, 1998; Rips, 1989; Matan & Carey, 2001 for more essentialist-based construals of artifacts).

An advantage of a generative view of classification is that it makes predictions not just for essentialized categories, but also for categories which are “essentialized”
only to a degree (i.e., have underlying properties that provide strong but not defining
evidence for category membership), or not all. For example, consider the simple causal
network in Figure 8 which involves only two category features, C and E, and in which C
is the cause of E. Note that the assumption is that neither C nor E are defining of
category membership, that is, neither C nor E will appear in all category members.

Nevertheless, it is possible to define the probability with which the network
generates particular objects (i.e., particular combinations of C and E). The network has
an \( m \) parameter that specifies the probability that the causal mechanism generates E
when C is present, and a \( b \) parameter that specifies the probability that E is brought
about by some unspecified background cause. The parameter \( c \) specifies the probability
that C will be present. Table 4 presents how these parameters together specify the
probability with which a category (A) with this causal network will generate the four
combinations of C and E. The probability that C and E will both be absent, \( P(\sim C \sim E \mid A) \),
is the probability that C is absent \( (1 - c) \) times the probability that E is not brought about
by any background causes \( (1 - b) \). The probability that C is absent but E is present,
\( P(\sim CE \mid A) \), is \( (1 - c) \) times the probability that E is brought about by some background
cause, \( b \). The probability that C is present but E absent, \( P(C \sim E \mid A) \), is \( c \) times the
probability that E is not brought about the causal mechanism and not brought about by
the background cause, \( (1 - m)(1 - b) \). Finally, the probability that C and E are both
present, \( P(CE \mid A) \), is \( c \) times the probability that E is brought about by the causal
mechanism or brought about by the background cause \( (m + b - mb) \). (Again, in these
equations I assume that the two potential causes are independent.)

Table 4 also presents the probabilities that the four cases will be generated when
\( c = 0.67 \), \( m = 0.80 \), and \( b = 0.20 \). This example illustrates how a generative theory can be
applied to categories missing a defining feature, as the probability of C (\( = P(C \sim E) +
P(CE) = 0.67 \)) and E (\( = P(\sim CE) + P(CE) = 0.63 \)) are both less than one. But it also illustrates
how the theory predicts that combinations of features make for better or worse category
members. In particular, the two objects in which C and E are both present or both absent are the most probable (.563 and .264, respectively). In contrast, an object where one is present and the other absent is very improbable (.066 for ~CE and .107 for C~E). In fact, these objects, which have one feature present, are both less probable than the one with both features absent. This result reflects the empirical observations one would expect when a causal law holds between two variables—namely, these variables should be correlated with one another. In other words, a generative view of classification predicts that objects will be good category members to the extent they exhibit theoretical coherence, that is, whether they are consistent with or corroborate a category’s causal laws (causes and effects either both present or both absent).

To test these predictions, half the participants in Experiment 4 were instructed on one of the six experimental categories with the causal network in Figure 9A in which four features were related by two causal links. The other half learned an identical control category missing the two causal relations (Figure 9B). In both conditions, each feature was described as occurring in “most” category members. A categorization test then followed which presented objects with values on all four dimensions, and participants rated how likely the object was a member of the category. The 16 possible objects that can be formed from four binary dimensions were each presented twice. Thirty-six undergraduates were assigned in equal numbers to the causal and control conditions.

The results are presented in Figure 10. For simplicity, Figure 10 presents the ratings only for objects that are maximally coherent (causes and effect either all present or all absent) or maximally incoherent (both causes present but both effects absent, or vice versa). In the control condition, categorization ratings were of course a monotonic function of the number of features: The object missing all four features received the lowest rating, the one with all four features receiving the highest rating, and objects possessing two features received an intermediate rating. In contrast, the results in the
causal condition showed a strong effect of causal knowledge. For example, incoherent objects received a significantly lower category rating as compared to the same items in the control condition. In fact, in the causal condition the incoherent objects (each with two features) received a significantly lower rating than items missing all four features. This results illustrates how the corroboration of causal laws can in fact override the importance of the number of characteristic features that an object displays.

These findings support the claim that subjects judged an object’s category membership as a function of the likelihood it was generated by the category’s causal laws. Other studies have also demonstrated the importance of whether combinations of features exhibit theoretical coherence. Ahn, March, Luhmann, and Lee (2002) have shown that exemplars are viewed as more typical when they manifest correlations between theoretically related pairs of features (e.g., an animal that lives underwater should also have gills) (also see Malt & Smith, 1984). And, I have shown that adults are not only sensitive to whether pairs of features exhibit coherence, but whether an entire collection of features linked together in more complex networks manifest the correlations between features that such networks produce (Rehder, 2003a; 2003b). Moreover, effects such as these are not limited to adults: Barrett, Abdi, Murphy, and Gallagher (1993) found that first- and fourth-graders were more likely to classify a bird as a member of a novel category if it manifested an expected correlation (e.g., between the bird’s memory capacity and brain size) than if it broke that correlation.

Besides illustrating the importance of theoretical coherence, another goal of Experiment 4 was to demonstrate how a generative view of classification can be applied to categories without a defining feature. Note, however, that the importance of coherence is not limited to just nonessentialized categories—even for categories based on an essence, coherence among observable features will contribute to category membership. This fact can be illustrated with the bird category shown in Figure 1. If one is presented with an unfamiliar animal (e.g., an ostrich) which shouldn’t be able to fly
(e.g., because it is too heavy relative to its wingspan), it may be considered likely to be a bird only if it doesn’t fly (despite the fact that flight is usually highly diagnostic of birds). This might be the case because a large, small-winged animal which somehow flies is actually more likely to be some kind of artifact instead (with an invisible propulsion system to explain its otherwise unexplained ability to fly).

Discussion

An enduring problem in the field of categorization has to been to account for both the undisputed fact that everyday categorization is based on observable properties, and the fact that categories have an underlying reality that goes beyond that which is perceptually available. By itself, the claim that categories possess both core properties (which define category membership) and observable ones (which serve as the basis for identification) leaves unexplained the conditions under which people will override perceptual information and rely on core properties instead. In this chapter, I have presented a solution to the categorization field’s mind-body problem by specifying the interaction between defining and observed features in terms of generative causal relations. On this account, objects’ observable features serve as evidence for category membership because they imply the presence of the defining feature. But people know that such causal inferences are no longer justified when an object’s features are transformed through external intervention. As a result, in such cases they will fall back on the object’s core properties to establish category membership.

One goal of this work of course has been to contrast a generative view of classification with one that merely distinguishes between core and observable properties. But an advocate of the core/observable position might ask why the relation between these property types need be conceived of as causal at all. To make this point, they need only refer to the existence of certain categories whose relation between core and observable properties in manifestly noncausal. For example, if a late-night TV
movie opens on a scene with people doing the hula underneath palm trees, you might guess that they’re on an island, but this guess is unlikely to be based on your belief that an island’s defining properties (small body of land surrounded by water) causes its observable ones (hula dancing and palm trees). By analogy, one might wonder about the necessity of causal relations even for biological kinds. For example, Rips’ (1989) transformed bird/insect may have still been considered a bird because there was no reason to think, given the scenario described, that its underlying defining properties had been changed by the transformation. Loosely speaking, they (the defining properties) were there before, there is no reason to believe that they’re not there now, and so why should one think that the category membership of the object has changed?

However, this argument leaves unexplained numerous empirical findings reported in this chapter. First, it leaves unexplained why symptoms might be more diagnostic of a disease when they are causally linked to the disease. This prediction was tested in Experiment 1, which found that in fact observable features provided stronger evidence for category membership when they were stipulated as being caused by a defining feature. It also fails to explain the phenomenon of boundary intensification, in which, categories become more homogenous and less tolerant of discrepant category members. This prediction was tested in Experiment 2, which found that in fact judgments of category membership were more extreme when causal knowledge was provided. Finally, it leaves unexplained why there might be asymmetries depending on whether a defining feature causes, or is caused by, the category’s observable features. This prediction was tested in Experiment 3, which in fact demonstrated the predicted asymmetry.

Another important component of the generative view is that causal relations between defining and observable features are typically viewed as probabilistic rather than deterministic. A probabilistic view of causality explains why, in Experiment 1, features more directly caused by the defining feature served as stronger evidence for
category membership as compared to less directly caused ones. It also explains why, in Experiment 3, an increase in the number of features resulted in a monotonic increase in category membership ratings for both common-cause and common-effect structures. But more importantly, a probabilistic view of causality explains the fundamentally probabilistic nature of categorization judgements with real-world categories. For example, research has demonstrated that people are often very uncertain about category membership (e.g., McCloskey & Glucksberg, 1978, found that just over 50% of subjects judged a peanut to be a vegetable; just under 50% judged a leech to be an insect). And although there are clearly individual features (e.g., flying) that are highly diagnostic of membership in certain categories (e.g., birds), such features are often not present in all category members (e.g., ostriches). But according to a generative version of essentialism, the uncertainty of whether a leech is an insect arises because the features of a leech provides weak inferential support for the defining features of insects as compared to, say, a mosquito. Likewise, flying provides evidence for bird category membership, but because flying is generated by the bird category’s causal model (Figure 1) with less than certain probability, it admits of the possibility of birds that don’t fly. According to the generative view, cases of fuzziness in category membership arise not necessarily because of the absence of defining features, but rather due to inferential uncertainty, that is, from the fact that sets of observed features vary in the extent they provide evidence of a defining feature.

The categories tested in the first three experiments were intended to be experimental analogs of fully essentialized categories, categories with an unobservable defining feature that causes observable features. But an advantage of a generative view is that it can accommodate categories that are less than fully essentialized. The results from Experiment 4 illustrate how objects will be considered better category members to the extent they make sense, that is, to the extent they are coherent in light of the causal laws the category is thought to possess.
An important advantage of the fact that the generative view can be applied equally well to both essentialized and non-essentialized categories is that it can accommodate categories for which underlying causal features have not yet been identified—as in earlier scientific epochs in which the causes of many diseases were unknown. It can also accommodate the developmental shift that apparently occurs in which many categories (e.g., biological kinds) increase in the extent to which they are essentialized. For example, Keil (1989) conducted a transformation experiment similar to Rips’ in which children were shown a picture of a raccoon and then told that doctors painted the raccoon black, and then added a white stripe down its back and a “sac of super smelly yucky stuff.” Whereas second- and fourth-graders judged that the animal was still a raccoon (illustrating again that category membership can be based on more than just appearance), kindergartners categorized it as a skunk, that is, perceptually. And although there is evidence that children as young as four-years-old might be biased to weigh “insides” of objects heavily in categorization (Gelman & Wellman, 1991; also see Hirschfeld, 1996, Diesendruck, 2001), it is frequently assumed that categories are initially organized around perceptual information and are augmented with more conceptual information over time (Keil, 1989; 1994).

This shift in the organization of a category—based either on scientific progress or cognitive development—can be expressed in terms of an evolving set of causal models, an example of which is shown in Figure 11. First consider what might be a category’s initial state in Figure 11A. In this early stage, knowledge of the category consists of only its observable features (closed circles) and how those features covary with a category label (depicted as an additional binary variable in Figure 11A). As a result, evidence that an object is a member of this category is simply determined by whether or not it has these features, a relationship depicted in Figure 11A by dotted arrows. Note that although dotted arrows represent a relation between features and the category label which is inferential (one infers a category label from features) it is non-causal. Because the
category likely possesses a family resemblance structure (in which features vary in the extent to which they are correlated with the category label but no one feature covaries perfectly), the function which computes the probability of the category label will usually involve a weighted combination of the number of features present.

In the second stage of the category’s development (Figure 11B), its representation has been elaborated not only with additional features, but also with the presumed presence of an underlying cause that generates several of the observed features. Knowledge of this underlying cause might arise from explicit education (i.e., being told). It might also arise because children are causal determinists who postulate the presence of hidden causes to explain what they observe (Gelman, 2003; Gelman, Coley, & Gottfried, 1994). In some instances the cause might be attributed to something external to the category itself (often the case with artifacts); in others, the knowledge of the cause might be so vague that it functions as a placeholder only (Medin & Ortony, 1989). But regardless of the source or nature of the cause, at this point it only provides the child with an explanation for what he or she observes. Classification itself is still determined by the observable features alone.

In the third stage (Figure 11C), the category has begun to undergo an essential shift in which the category label is now directly dependent on the underlying cause. Importantly however, this shift is not complete, because the category label is still directly functionally dependent on observable features as well. At this stage, observable features contribute to category membership in two ways. The first is that they directly imply category membership (as in the first two stages). The second contribution is indirect, because from observable features one infers the likely presence of the underlying cause, which then increases the probable presence of the category label yet further. Because of this second inferential path from observable features to the category label, it is at this point that categories begin to become less variable, and, as a result, undergo boundary intensification.
The essential shift is completed in Figure 11D which presents a fully essentialized category in which category membership is directly dependant on the underlying cause alone. At this point observable features still imply the category label, but do so indirectly by implying the underlying cause, in the same way that one can infer a disease from its symptoms. With increased knowledge, one might also come to learn about more features, and more causal links between features and the underlying cause. In addition, one might learn that some of those feature are causes of the underlying defining feature rather than being caused by it (e.g., one learns that unsafe sex is a potential cause of contracting HIV).

I suggest that it is in this manner that categories shift from being primarily perceptually based to having an essentialized underlying causal structure. Notably, at each stage classification proceeds on basis of observable features, as it must. But what changes is the nature of the inference itself. What starts as a noncausal inference from features to a category label (as assumed by most current theories of categorization), turns into a causal inference from observable to unobservables (and then to the category label).

Of course, not every category necessarily progresses through each of the four stage shown in Figure 11. For example, the presence of a unique word for a type of object may result in a child’s category representation progressing immediately from the first stage to the third (or fourth) (Coley, Medin, & Atran, 1997). Conversely, note that for many categories it is an open question of whether the process of essentialization is ever fully completed, even for adults. For example, Hampton (1995) has demonstrated that even when biological categories’ so-called defining properties are unambiguously present (or absent), characteristic features continue to exert an influence on judgments of category membership (also see Braisby, Franks, & Hampton, 1996; Malt, 1994; Malt & Johnson, 1992).

But regardless of whether a category is essentialized in full, in part, or not at all,
this chapter has demonstrated how classification can be construed a process which estimates the likelihood that an object is generated by the category’s causal model. For centuries categorization theorists have wrestled with the problem of the relationship between the observable and the unobservable when discussing human knowledge of categories—no less eminent a philosopher than John Locke distinguished nominal essences (observables which form the basis of classification) from real ones (categories’ true underlying nature). If for a moment we treat Locke as a psychologist instead, we can see that he may have been on the right track all along:

...nature, in the production things, always designs them to partake of certain regulated established essences, which are to be the models of all things to be produced. (Locke, 1690/1974, pp. 289-290)
References


Author Note

Bob Rehder, Department of Psychology, New York University.

Correspondence concerning this article should be addressed to Bob Rehder, Department of Psychology, 6 Washington Place, New York, NY 10003 (email: bob.rehder@nyu.edu).
Footnotes

1 The hedge “all else being equal” (which appears twice in this paragraph) refers to the fact that strength of the evidence that $S_i$ provides for the disease $D$ will of course also depend on the evidence that it provides for some other disease, and also the prior probability of $D$ and those other diseases. This issue is discussed further in Experiment 2. The purpose of the discussion to this point is simply to establish how the likelihood of $D$ given a symptom, $P(D | S_i)$, is going to vary as a function of the probability that $D$ generates that symptom, $P(S_i | D)$.

2 For example, in terms of causal-model theory, each feature $i$ in Figure 11A would have a $b_i$ parameter representing the probability that it occurs in category members. The probability that a given object was generated by that category, $P(O | C)$, would then be

$$P(O | C) = \left[ \prod_{i \in P} b_i \right] \left[ \prod_{i \in A} (1 - b_i) \right]$$

where $P$ is the set of features present in $O$ and $A$ is the set of feature absent in $O$. Alternatively, category membership could be based on $O$’s similarity to the category’s prototype. Note that these two alternatives would be equivalent if a multiplicative similarity rules was used (Nosofsky, 1992; Smith & Minda, 2000), because they would both involve multiplying rather than summing evidence.

Of course, in Figure 11 we are once again considering only the evidence of an object $O$’s category membership with respect to a single category. The extent to which a feature provides evidence for category membership will also depend on the likelihood that that feature is associated with other categories. (See Experiment 2 for discussion.)
Table 1
Feature probabilities for the causal network of Figure 2.

| Feature | P(S₁|D; mᵢ, bᵢ) | P (S₁|D; mᵢ = .80, bᵢ = .20) |
|---------|----------------|-------------------------------|
| S₁      | m₁ + b₁ - m₁b₁ | .84                           |
| S₂      | m₂P(S₁|D) + b₂ - m₂b₂P(S₁|D) | .74                           |
| S₃      | m₃P(S₂|D) + b₃ - m₃b₃P(S₂|D) | .67                           |
Table 2
Features and causal relationships for the Lake Victoria Shrimp experimental category.

<table>
<thead>
<tr>
<th>Features</th>
<th>Causal Relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>A high quantity of the ACh neurotransmitter causes a long-lasting flight response. The duration of the electrical signal to the muscles is longer because of the excess amount of neurotransmitter.</td>
</tr>
<tr>
<td>$F_1$</td>
<td>A long-lasting flight response causes an accelerated sleep cycle. The long-lasting flight response causes the muscles to be fatigued, and this fatigue triggers the shrimp’s sleep center.</td>
</tr>
<tr>
<td>$F_2$</td>
<td>An accelerated sleep cycle causes a high body weight. Shrimp habitually feed after waking, and shrimp on an accelerated sleep cycle wake three times a day instead of once.</td>
</tr>
</tbody>
</table>
Table 3
Hypothetical predictions for Experiment 2.

| Number of A Features | P \( (A | O; m = 0; b = .75) \) | P \( (A | O; m = .50; b = .75) \) |
|----------------------|----------------------------------|----------------------------------|
| 0                    | .047                             | .003                             |
| 1                    | .250                             | .125                             |
| 2                    | .750                             | .875                             |
| 3                    | .964                             | .997                             |
Table 4
Object probabilities for the causal network of Figure 7.

<table>
<thead>
<tr>
<th>Object (O)</th>
<th>$P(O \mid A; c, m, b)$</th>
<th>$P(O \mid A; c = .67, m = .80, b = .20)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sim C \sim E$</td>
<td>$(1 - c) (1 - b)$</td>
<td>.264</td>
</tr>
<tr>
<td>$\sim CE$</td>
<td>$(1 - c) (b)$</td>
<td>.066</td>
</tr>
<tr>
<td>$C \sim E$</td>
<td>$(c) [(1 - m)(1 - b)]$</td>
<td>.107</td>
</tr>
<tr>
<td>$CE$</td>
<td>$(c) (m + b - mb)$</td>
<td>.563</td>
</tr>
</tbody>
</table>
Figure 1

Bird DNA

Sings
Wings
Light
Eats Seeds
Flies
Nests in trees
Figure 2
Figure 4

Category A

D → F₁ → F₂ → F₃

Category B

~D → ~F₁ → ~F₂ → ~F₃
Figure 5

![Graph showing choice (A) against number of A features for causal and control conditions.](image-url)
Figure 6

A

Common Cause

HIV+

Sarcoma

Pneumonia

Common Effect

Blood Transfusion

Intravenous Drug Use

Unsafe Sex

B

Common Cause

D

F₁

F₂

F₃

Common Effect

D

F₁

F₂

F₃
Figure 7
Figure 8
Figure 9
Figure 10

```
<table>
<thead>
<tr>
<th>Category</th>
<th>Membership</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td>Causes and Effects</td>
<td>Both Absent</td>
</tr>
<tr>
<td>1010</td>
<td>Causes but no Effects</td>
<td>Effects but no Causes</td>
</tr>
<tr>
<td>0101</td>
<td>Effects but no Causes</td>
<td>Causes and Effects</td>
</tr>
<tr>
<td>1111</td>
<td>Causes and Effects</td>
<td>Both Present</td>
</tr>
</tbody>
</table>

Features in Object O

Category Membership Rating

- Causal
- Control
```
Figure 11