Classifying with Essentialized Categories

Bob Rehder (bob.rehder@nyu.edu)
ShinWoo Kim (shinwoo.kim@nyu.edu)
Department of Psychology, New York University
6 Washington Place, New York, NY 10003 USA

Psychological essentialism states that certain categories are assumed to have an underlying hidden reality (or "essence") that defines objects' identity (Gelman, 2003; Medin & Ortony, 1989). Everyday classification, on the other hand, must be based on the features of objects that are observable. How do we reconcile these facts? One way is to assume that essential features cause observable ones, and that classification involves reasoning backwards from observable features to their hidden cause. Three experiments tested classification with essentialized categories to determine whether causal inference underlies classification.

Method

In each experiment, 24 subjects learned two categories. For example, some subjects learned about Kehoe Ants and their features (high amounts of iron sulfate, hyperactive immune, thick blood) and Argentine Ants and their features (high amounts of metallic sodium, fast digestion, short life span). Both categories were essentialized, because each had one feature (iron sulfate and metallic sodium) that was described as occurring in all category members and no nonmembers. The other features were described as occurring in 75% of their respective category members. Each category also possessed interfeature causal relations. Pairs of ants, shrimp, cars, computers, stars, and molecules were tested.

Fig. 1 presents the causal relations in Expts. 1-3. For example, in Expt. 1's Category A the essential feature E_A (iron sulfate) was described as causing A_1 (hyperactive immune) but not A_2 (thick blood); in Category B the essential feature E_B caused B_2 but not B_1. After learning subjects were presented with pairs of features, one from each category (e.g., A_1B_1), and asked to choose whether the item was an A or B. We predicted that features would be more diagnostic of category membership when they could be used to reason backwards to their underlying essence.

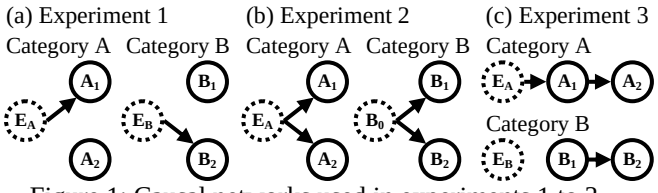

Figure 1: Causal networks used in experiments 1 to 3.

Results and Discussion

As predicted, features were more diagnostic of category membership when causally related to an underlying category essence. Test item A_1B_1 was classified more often as an A (81%), presumably because E_A can be inferred from A_1 whereas E_B cannot be inferred from B_1. Similarly, item A_2B_2 was also classified more often as a B (88%). We also tested items in which the presence of a feature was explicitly denied. $\sim A_1 \sim B_1$ (normal immune system and digestion) was classified as a B 65% of the time (because $\sim A_1$ implies $\sim E_A$) and item $\sim A_2 \sim B_2$ was classified as an A 71% of the time (because $\sim B_2$ implies $\sim E_B$). Classifiers appear to reason backwards from observed features to essential ones to establish category membership.

An alternative interpretation is that A_1 and B_2 were more diagnostic because they participated in a causal relation, not because they were used to reason to an essence. Expt. 2 addressed this possibility. The underlying features (E_A & B_0) were described as causing both observed ones (Fig. 1) but, rather than being essential, B_0 was described as having a 75% base rate. Both test items A_1B_1 and A_2B_2 were classified more often as an A (67%), supporting the claim that classifiers were reasoning causally to the underlying cause, and that an essential feature (E_A) is more diagnostic than a merely probable one (B_0). In Expt. 3, each category had an essential feature, but A_2 was causally linked (indirectly) to E_A but B_2 was not linked to E_B (Fig. 1). Test item A_2B_2 was classified more often (73%) as an A (despite that A_2 and B_2 are involved in the same number of causal links), apparently because one can infer E_A from A_2 but not E_B from B_2.

Expts. 1-3 support the claim that classification with essentialized categories can involve causal inference from observed to unobserved essential features. We do not claim that causal inference occurs in all acts of classification, because categories vary in the degree to which they are essentialized. Some might be *partly* essentialized in that observable features still provide their own *direct* evidence for category membership (in addition to the *indirect* evidence they provide via causal inference to an essence). The degree to which categories are essentialized might vary with domain and conceptual development (Rehder, in press). But when categories are explicitly essentialized, the current results show that humans readily engage in causal reasoning in service of classification.

References

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

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

Rehder, B. (in press). Essentialism as a generative theory of classification. In A. Gopnik & L. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford, UK: Oxford University Press.