Finding the Pope in the pizza: Abstract invariants and cognitive constraints on perceptual learning

John E. Hummel and Philip J. Kellman

Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095-1563. **jhummel@psych.ucla.edu; kellman@psych.ucla.edu**

Abstract: Schyns, Goldstone & Thibaut argue that categorization experience results in the learning of new perceptual features that are not derivable from the learner's existing feature set. We explore the meaning and implications of this "nonderivability" claim and relate it to the question of whether perceptual invariants are learnable, and if so, what might be entailed in learning them.

Schyns, Goldstone & Thibaut argue that visual features governing object classifications can be created by categorization experience. This is an important idea (if not a completely novel one; see, e.g., Biederman & Schiffrar 1987), but the key unanswered question is where the new features come from.

Schyns et al. argue extensively that the features created during category learning are "not present in, or derivable from, the [existing] feature set (sect. 1.1, para. 4)." It is easy to understand why they would want to make this claim (henceforth, the *nonderivability* claim): if the features acquired during category learning are just concatenations (intersections and unions, or worse – simple weighted sums) of features that existed before category learning, then the phenomenon of "category-based feature learning" might be construed as a simple matter of "selection" or "weighting." Although Schyns et al. treat nonderivability as a stepping stone to the broader claim that category learning constrains feature perception, the issue of nonderivability is arguably the more important issue. Among other things, it relates to the notion of abstract *invariants,* which are important in shape perception and object recognition (Biederman 1987). Insight into whether and how invariants can be learned from experience would make a substantial contribution to our understanding of object perception, recognition, and categorization.

Clearer understanding of nonderivability is necessary in tackling this question. There are at least three senses in which some new feature might be nonderivable from the population of previously existing features in the system (although it is unclear which Schyns et al. intend). The most literal interpretation is that the new feature is not derivable (computable) *at all* from the existing features. This version of the claim is absurd: any feature that is detected by the visual system must be computed from some finite set of operations performed on the representations given by early visual processes.

The second and third interpretations of nonderivability are more interesting because they characterize, respectively, two different ways of detecting features in any computational system. The second interpretation is that the new feature is not a simple weighted sum (i.e., linear combination) of existing features. This interpretation is suggested by the discussion of XOR, a famous example of a function that cannot be computed by a linear system. On this version of the claim, the perceptual/categorization system can be viewed as analogous to a large, multilayered neural network whose units have nonlinear activation functions (such as a standard backpropagation net). New features in one layer of the system would be composed in a nonlinear way (e.g., as a weighted sum subjected to a threshold) from existing features. This approach to feature detection and learning is standard fare in artificial neural networks. On this interpretation of the nonderivability claim, Schyns et al.'s theory is (essentially) that category learning (e.g., in "higher" layers of the network) serves to guide feature learning (in lower layers). This would be interesting, but not earth-shattering.

The third – and most interesting – interpretation of nonderivability is that the newly discovered feature is an abstract invariant, which, although *computable from,* is not truly *definable in* the vocabulary of existing features. For example, no logical concatenation – conjunctive, disjunctive, or otherwise – of local retinal activations defines the invariant square. It is acordingly a mystery how the visual system discovers such invariants in the outputs of local features (such as edges). Where invariants are concerned, it is not the case that "novel visual features are certainly reducible to their retinal encodings" (sect. 2.7). The argument is similar to those arising in discussions of scientific reductionism (Putnam 1975). Put simply, squareness is both more and less than any finite set of retinal activation patterns. It is more because some new activation pattern might also be a square, and it is less because many of the attributes of retinal activation patterns have nothing to do with their squareness. "Square" is an abstract invariant. If this is what Schyns et al. mean by nonderivable, then their claim is that category learning directs the discovery of invariants, as Gibson (1969) suggested some time ago. To our knowledge, no one has demonstrated how such invariants are discovered. The question of how (and whether) nonderivables such as invariants can be learned is a computational/algorithmic one that demands a far more specific theory than the one presented in the target article.

Toward that end, it is important to appreciate that discoverable new features do not include all logically possible ones, as Schyns et al. seem to suggest. Rather, human cognition is organized (constrained) for the discovery and synthesis of overlapping patterns in space and time. For example, we are better at detecting and learning about spatial (and temporal) relationships among parts that are close together rather than widely separated, and we are much more sensitive to some kinds of shape attributes than others (compare locating first-derivative discontinuities in contours [corners] vs. third-derivative discontinuities). Some well-defined attributes are unlearnable or even undetectable (Julesz 1981). Many of the answers to the mystery of where new features come from will probably emerge from identifying constraints on the vocabulary of spatial and temporal properties and relations that make up the human endowment for perception and perceptual learning.

Can features be created on the fly?

Koen Lamberts

University of Birmingham, Birmingham BI5 2TT, England, United Kingdom. **k.lamberts@bham.ac.uk**

Abstract: It is argued that feature creation may not only depend on categorical distinctions that are made during category learning, but also on the choice set during subsequent categorization.

Schyns et al. argue convincingly that higher-level cognitive processes influence the lower-level features that are created and used. I largely agree with their analysis and with its conclusions. Although the influence of learning on low-level processes has been studied for a long time, it is good to see a systematic and rigorous exploration of the effects of category learning on feature creation. In this commentary, I will discuss one issue that has not been addressed in the target article.

My argument concerns the role of choice sets in category learning and categorization. Which features are functionally optimal will ultimately depend on two elements: (1) the categorical distinctions that need to be made during category learning, and (2) the set of category alternatives that are considered during subsequent categorization. Schyns et al. address only the first of these two elements, but I will argue that the second may be equally important if we want to understand how higher-level processes affect low-level feature creation.

Category learning in daily life differs in many respects from category learning in the typical laboratory experiment. In many category-learning experiments, there are only a few mutually exclusive alternative categories available. However, in daily life, the set of alternatives is usually much larger and often implicit.