Classification and Prior Assumptions about Category "Shape": New Evidence Concerning Prototype and Exemplar Theories of Categorization

Emmanuel M. Pothos
Department of Experiment Psychology
University of Oxford
South Parks Road, OX1 3UD, UK.
pothos@psy.ox.ac.uk

Nick Chater
Department of Psychology
University of Warwick
Coventry, CV4 7AL, UK.
n.chater@warwick.ac.uk

Abstract

According to prototype theories of categorization, the cognitive system makes the default assumption that a category, *C*, is a roughly convex region in an internal space. This suggests that the default assumption for the "negative" category, *not-C*, should be the complement of this region—i.e., the internal space, minus a convex "hole." These different prior assumptions suggest potentially radically different patterns of generalization in category learning. We show experimentally that such effects do occur. These results are compatible with prototype accounts of categorization, but seem incompatible with exemplar accounts. We consider potential empirical extensions of this research, and its wider theoretical implications.

Introduction

Category learning from examples appears to be of fundamental psychological importance, both in learning the structure of the world, and learning the meanings of words which refer to that structure. Learning categories from examples is a type of *inductive* inference. That is, the cognitive system must make the leap from the finite set of particular examples to a general characterization which applies to limitless numbers of future examples. But inductive inference is problematic because there are infinitely many general characterizations (corresponding to all the different ways of classifying the unseen items) that are compatible with the set of examples that have been encountered (e.g., Goodman, 1954, Watanabe, 1985). How does the cognitive system choose from this infinite range of alternatives?

All solutions to this problem involve proposing some "bias" in the cognitive system to favor certain generalizations rather than others. Such biases can take various forms ranging from strong nativism to strong empiricism. At the nativist extreme, the bias may take the form of a finite innate repertoire of prestored categories—then the problem of category induction reduces to the problem of deciding which innate categories are consistent with the examples encountered so far (see Fodor, 1980; Piatelli-Palmerini, 1989). At the empiricist extreme, the bias might be viewed as imposed merely by general properties of the learning mechanism. For example, exemplar models of categorization assume that category structure is given by similarity comparisons to the exemplars encountered. Therefore, the choice of similarity measure, and how it is used to determine the region of generalization, embody the bias in favor of a particular generalization (e.g., Nosofsky, 1988; Nosofsky, 1989).

Intermediate between these two extremes, and the focus of this paper, is the view that the cognitive system makes prior assumptions about the "shape" of categories in some internal space. Most notably, prototype accounts assume that categories correspond to roughly *convex regions* of space, whereas according to exemplar models categories can have arbitrary shapes depending on the locations of the examples encountered in the internal space (more formally, exemplar models are non-parametric category boundary estimators; see Ashby & Alfonso-Reese 1995). That prototype theories require convexity can be readily seen by noting that a convex category structure will necessarily include all measures of central tendency that might be of interest, whereas this would not typically the case for concave category structures. In this paper, we consider the evidence that such general assumptions about category "shape" are involved in learning categories from exemplars. Any such effects present considerable difficulties for exemplar theories of categorization.

The prototype assumption that a category, *C*, corresponds to convex regions in an internal space is represented by the filled circle in Figure 1. It follows that a category *not-C* should be assumed to have the shape of the complementary region, shown in Figure 2. This appears to lead immediately to a paradox. If we relabel *not-C* as *D*, then *C* becomes *not-D*. But switching to these terms appears to reverse the prior assumptions about category structure—now *D* (i.e., *not-C*) is assumed to have a convex shape, and *not-D* (i.e., *C*) the complementary shape. But, as noted before, any representation of a central tendency within a category structure requires convexity so that, in this view of categorization, *C* would be a possible category while *D* would not. Based on this observation, it seems reasonable to further suggest that the categorization system assumes natural language terms correspond to convex categories, and thus while *C* is a plausible *natural language predicate*, *D* is not. So the apparent symmetry between descriptions in

terms of *C*s to *D*s is illusory, if one is to maintain a prototype- driven view of categorization[1].

So suppose that under some circumstances, the categorization system will adopt the prototype assumptions about category structure. Then, describing a set of items as members or non-members of a category should affect the direction of generalization. This is because part of a prototype-based categorization process is the presupposition that categories have a convex spatial structure. Therefore, when the exemplars are presented as members of some category, generalization would extent to the convex space within the boundary defined by these exemplars. Crucially, when the same items are described as non-members of some category, the region they belong to would be assumed to be concave, so that the direction of generalization would be reversed.

Figures 3, 4, 5 and 6 illustrate the above points. Figures 3 and 4 show the same set of exemplars, arranged in a circle. In Figure 3, the items are labeled as "Chomps" and therefore the category would be perceived as positive; in Figure 4, the items are labeled as "Non-Chomps" and the category would be assumed to be negative. Figures 3 and 4 schematically represent the two conditions of the experiment that we report below. Crucially, generalization to the two test items, one inside the circle and one outside the circle, is predicted to be reversed.

Figures 5 and 6 show the direction of generalization that corresponds to the two different ways of understanding the category. If the category is assumed to be positive, the region of generalization will be convex, enclosing this circle of exemplars. Thus, the center of the circle will be considered a member of the category but not a point far away from the circle (see Figure 3). Conversely, if the category is perceived to be negative, then the most natural assumption is that the circle of points lie just outside the boundary of the complementary positive category (see Figure 4)[2]. This reverses the pattern of generalization on the previous assumption: The center of the circle will be rejected as a category member, and a point far away from the circle will be assumed to be a member.

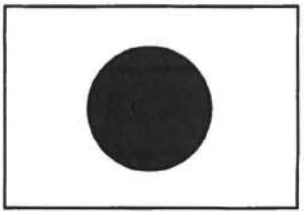

Figure 1. Prior assumption about the shape of a "positive" category.

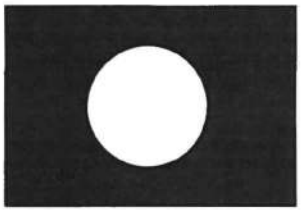

Figure 2: Prior assumption about the shape of a "negative" category.

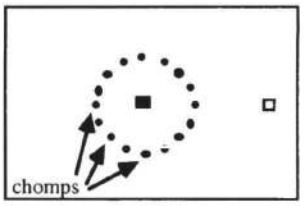

Figure 3: Training examples for a "positive" category. The exemplars, represented by filled blobs, are labeled "chomps". We assume that this lexically simple label allows the default assumption that the category is positive, and thus has a convex structure. The two squares correspond to two crucial test items. The filled square has the same category as the filled blobs; the unfilled square has the opposite category.

[1] Notice that prototype structure plausibly applies only to categories which include a small proportion of objects that a person may consider. This is because a category including many things will necessarily include objects which are so diverse that they cannot meaningfully be associated with the same prototype representation. Indeed, almost all lexical categories, such as *dog*, *table* and *plant* apply to a small proportion of objects; for instance, there are many more non-dogs than dogs. Thus, aside from the "shape" of a category, there is a related assumption concerning the "size" of that category. This corresponds to the *rarity* assumption that has proved important in reasoning research (e.g., Oaksford & Chater, 1994).

[2] Of course, the cognitive system might also assume that the complementary positive category lies somewhere outside the circle of negative exemplars. If the participants believe that exemplars are being chosen in order to help teach them the category, there may, however, be a presumption in favor of the assumption that the positive category is *inside* the circle of examples, because only on this viewpoint are the exemplars useful in constraining the *boundaries* of the category. We show below that, under some circumstances at least, people do make this assumption.

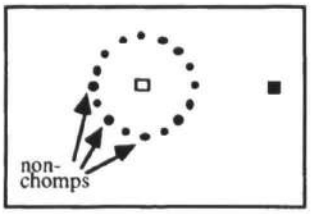

Figure 4: Training examples for a "negative" category. The exemplars, represented by filled blobs, are labelled "non-chomps." We assume that this label triggers the cognitive system to assume that the category is the complement of a convex category. As in Figure 3, the two squares correspond to two crucial test items. The filled square has the same category as the filled blobs; the unfilled square has the opposite category. The predictions about generalization from this category are the opposite of those predicted if the category is assumed to be positive (see Figure 3).

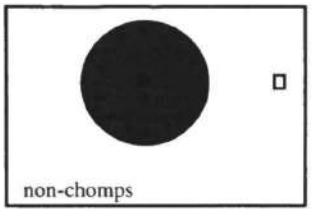

Figure 5: Induction using the prior assumption associated with a positive category. The filled blobs represent exemplars. The shaded area represents the region which is generalized from the examples. The unfilled square represents the inferred prototype for the category.

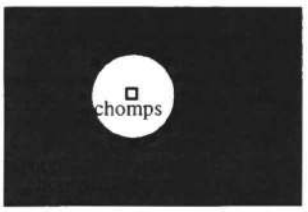

Figure 6: Induction using the prior assumption associated with a negative category. The filled blobs represent the same exemplars as before. The shaded area represents the region which is generalized from the examples. The unfilled square represents the inferred prototype for the negation of the category. Notice that the square is therefore *not* a member of the same category as the exemplars.

So the prototype account of category structure raises the possibility that there that may be a reversal in generalization, depending on whether a category is assumed to be positive or negative. Notice that any such effect appears to be beyond the scope of an exemplar-based account of categorization. This is because exemplar accounts do not make assumptions about category structure—they simple generalize according to the similarity of test items to the exemplars already presented. Therefore, whether the category is considered to be positive or negative should have no effect on generalization, according to an exemplar account.

Experimental Test

We presented an imaginary category in a way which we hoped would encourage the cognitive system towards assuming a prototype structure for that category. We tried to bias subjects towards presupposing an abstract representation of the category by using a simple category label and other experimental manipulations. On these assumptions, we expected that the direction of generalization would depend on whether the training exemplars were described as being members or non-members of the category: According to our theory, in the former case generalization would extend to a convex region in some spatial representation of the items, whereas in the latter case the direction of generalization would be reversed.

Subjects

Participants were University of Oxford students who either volunteered to participate or were paid one pound. Paid subjects participated in another completely unrelated experiment. They were all naive regarding the theory behind this experiment.

Design

A between subjects randomized design was employed where in one condition subjects were told that the items they saw in the training part were not members of an imaginary category (referred to as the category of Chomps) while in the other condition the items in the training part were described as members of that category. Ten participants were tested in each condition[3]

Materials

The stimuli used were arrangements of little black squares in a ten by ten imaginary grid so that about half the grid was in black. A line dividing the grid into two halves was an axis of symmetry for the symmetrical stimuli (but this was not pointed out to the subjects). The items used in the training part, referred to as near symmetric patterns (*NS*), were all nearly symmetric but for one defect (that is, moving one black square in one of its nearest neighbor positions would make the pattern perfectly symmetric about its axis of symmetry). Figure 7 shows a typical *NS* pattern. Care was taken to ensure that the defects were randomly distributed in

[3] The data of one subject were replaced; her generalization set included exemplars from all the types of patterns we used so that her results do not bear on the questions addressed in this experiment.

the four quadrants of the grid. In the test part, there were three kinds of stimuli: Symmetric patterns (*CS*) that corresponded to the *NS* patterns of the training part, additional symmetric patterns (*S*) that were not related to the *NS* patterns, and also random patterns (*R*), which were created by randomizing the distribution of black squares. The stimuli were printed individually on A4 paper in black ink. The size of the grid was about 6.4 x 6.4 cm^2 and the size of the little squares 0.65 x 0.65 cm^2.

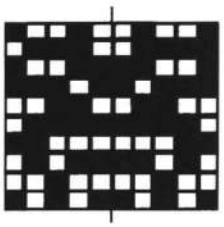

Figure 7: An exemplar in the training phase of the experiment.

Procedure

At the beginning of each experimental session subjects were given one sheet of printed instructions where they were simply told they were about to receive a folder with a set of items, and that they would have to examine these items as carefully as possible. It was noted that there was no time limit for going through the items. They were also told orally, after going through the instructions, that there was no particular order in which they had to study the items and that they were free to review items they had already studied. Finally, it was noted that these items would be available to them in the second part of the study as well, but no further information was given about the test part. There were eight *NS* patterns in the training part, seven of which were unique. Once subjects reported that they had seen the items, they were then presented with a second sheet of printed instructions where they were told in one condition (the Non-Chomps condition) that: "There is a category to which we shall be referring as the category of Chomps. The items you have just seen are all non-Chomps, that is they were all not members of that category. You will shortly receive another folder with items. Please sort the new items into two piles, one for Chomps and another for non-Chomps." In the other condition (the Chomps one) the instructions were modified so that the training items were described as Chomps. It was again emphasised that there was no time limit, that they could sort the items in whichever order they liked, and that they were allowed to make corrections (that is, they could change their mind about which items would be classified as Chomps and which ones as non-Chomps). The training items were also available to them in case subjects wanted to consult them. In the test part, there were the eight *CS* patterns, another eight *S* ones and also eight *R* patterns. Once subjects had sorted the items but before actually identifying either pile, they were reminded which category the training items belonged to, and were asked to indicate the items they thought were Chomps and non-Chomps. All subjects were tested individually and the experiment lasted for approximately ten minutes.

Results

We wanted to investigate the hypothesis that there are situations where classification is based on some prototype representation of the category and that this leads the cognitive system to make specific presuppositions about the shape of the category. The only assumption we made about the arrangement in some internal psychological space of the items we used was that symmetric patterns occupied some convex region (as the category "symmetric patterns" can clearly have a prototype structure), while non-symmetric patterns are represented in the surrounding concave areas. Therefore, the prediction of our hypothesis was that subjects would be selecting *S* and *CS* patterns as Chomps in the Chomps condition (category of training exemplars assumed to be positive so that generalization would occur within the convex space of symmetric patterns) and that they would be selecting the *S* patterns as Chomps in the Non-Chomps condition as well (category of training items assumed to be negative; generalization extends to the concave region surrounding the *S* patterns). As can be seen from Figure 8, the pattern of classification was completely reversed across the two conditions, thus confirming our expectations. The chi-square test we used to assess qualitatively these differences was found highly significant (on one degree of freedom, $\chi^2=9.9$, $p<0.005$). Figures 9 and 10 further illustrate the observed pattern of generalization.

Figure 8: Generalization patterns in the Chomps and Non-Chomps experimental conditions[4].

[4]Generalization to Random Patterns means that all the *R* patterns were selected as consistent with the training items and, likewise, Generalization to Symmetric Patterns means that all the *S* and *CS* patterns were selected (except in the Chomps condition where in some cases only a subset of the *S* and *CS* patterns was selected; however, this does not affect our results qualitatively since we are interested in the direction of generalization).

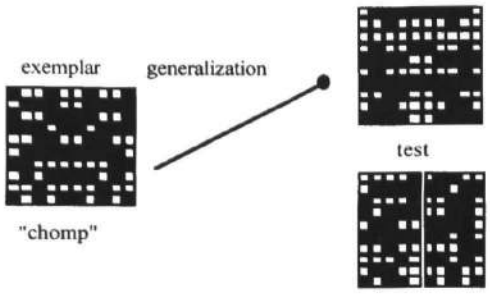

Figure 9: Generalization when exemplars are labelled as "chomps", suggesting a positive category.

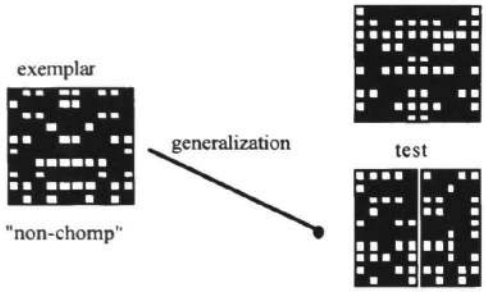

Figure 10: Generalization when exemplars are labelled as "non-chomps", suggesting a negative category. This gives the reverse pattern of generalization to that shown in Figure 9.

Cognitive Penetrability of the Categorization System

The advent of the cognitive school of psychology has been partly associated with the realization that computational modeling, that is explanations based on rules and mental representations, is a fruitful way to describe much of our cognition. However, as Pylyshyn (1981a) noted, among others, if left unconstrained computational models have simply too many free parameters; potentially they can be modified to fit any set of data (refer also to Anderson's mimicry theorem, Anderson, 1978). To resolve the issue, Pylyshyn suggested that "The coherence of such a view [a computational view of cognition] depends on there being a principled distinction between functions whose explanation requires that we posit internal representations and those that we can appropriately describe as merely instantiating causal physical or biological laws" (Pylyshyn, 1980) so that the latter processes would provide the fixed points, the "cognitive constants" (Pylyshyn, 1981b) in computational models. Cognitive penetrability was put forward as the criterion according to which we are to distinguish between the two types of processes, where a cognitively penetrable mechanism is one that is affected by tacit knowledge, background beliefs etc. (e.g. Pylyshyn, 1980). Of relevance here is our demonstration that the logical structure of category labels affects prior expectations of category structure. This raises the question of the extent to which classification learning performance may be potentially affected by other sources of knowledge, such as background knowledge of the category domain. Perhaps classification research has been so tractable to exemplar-based and other formal models precisely because the stimuli used are "meaningless" (as are their labels) so that such knowledge cannot be brought to bear (see Pickering & Chater, 1995 for related discussion). Further demonstrating the cognitive penetrability of categorization mechanisms would raise questions concerning the generality of current formal accounts.

Future Work

This research suggests that categorization may be influenced by linguistic information about the prior structure of categories. An interesting question is to consider whether such effects occur more generally. For example, consider the case of disjunction. Suppose that exemplars form two distinct groups (as in Figures 11 and 12). If the category is given a lexically simple label, we would expect that the default prototype structure is assumed (see Figure 11). Thus, a test item, that lies outside the two clusters of exemplars, but between them, will be judged to be a member of the category (indeed, it may be judged a particularly good member of the category, because it is near the central tendency of the category). By contrast, suppose that a person is cued that the category has a disjunctive structure, e.g., by being given a verbal label such as *"chomps or blibs"*. This may cue the assumption that the category shape corresponds to *two* convex regions in an internal category space. Using this prior assumption leads to the generalization shown in Figure 11, where the test item, which was previously prototypical, is now not considered to be a member of the category at all. This suggests that, whether a test item is classified as a member of the same category as the exemplars, may depend on the prior structure that is assumed. Whether this result is observed is an interesting direction for future research. In a more general context, following Murphy and Medin (1985), the question of what makes categories coherent may be partly answered by discovering that the cognitive system makes default assumptions about category structure, under different circumstances.

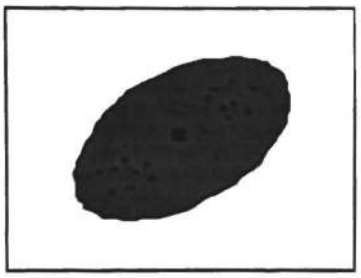

Figure 11: Generalization assuming that the category corresponds to a single convex region.

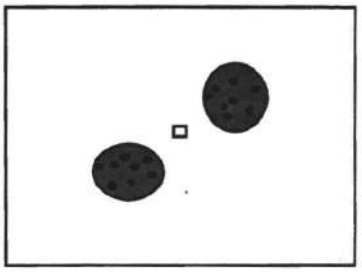

Figure 12: Generalization assuming that the category corresponds to two convex regions—i.e., assuming that the category has a disjunctive structure.

Conclusions

We have outlined a novel methodology for studying human categorization, where prior assumptions concerning category "shape" may be manipulated by changing the logical structure of the category to be learned. We have shown that this method of manipulating prior assumptions about category structure, using negation, can reverse generalization behavior. This result creates considerable difficulties for theories of categorization which do not allow prior assumptions to influence classification learning, such as exemplar models. Further study of the manipulation of prior expectations, using a variety of methods, therefore appears to be an exciting direction for future research on categorization.

Acknowledgments

The first author was supported by the UK Medical Research Council (reference number: G78/ 4804), St. Peter's College, Oxford, and the A. S. Onasis foundation (reference: Group S-076/1996-97).

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, 85, 249-277.

Ashby, G. F. & Alfonso-Reese, L. A. (1995). Categorization as Probability Density Estimation.. *Journal of Mathematical Psychology* 39, 216-233.

Fodor, J. A. (1980). The present status of the innateness controversy. In J. A. Fodor (Ed.) *Representations*, (pp. 257-316), Cambridge, MA: MIT Press.

Goodman, N.: (1954). *Fact, fiction and forecast*, London: Athlone Press.

Murphy, G. L.., Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review*, 92, 289-316.

Nosofsky, R. M. (1988). Similarity, Frequency, and Category Representation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14(1), 54-65.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Journal of Experimental Psychology: Perception and Psychophysics*, 45(4), 279-290.

Oaksford, M. & Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review*. 101, 608-631.

Pickering, M. & Chater, N. (1995). Why cognitive science is not formalized folk psychology. *Minds and Machines*, 5(3), 309-337.

Piatelli-Palmerini, M. (1989). Evolution, selection and cognition. *Cognition*, 31, 1-44.

Pylyshyn, Z. W. (1981a). The Imagery Debate: Analogue Media Versus Tacit Knowledge. *Psychological Review*, 16-45.

Pylyshyn, Z. W. (1981b). Psychological explanations and knowledge dependent processes. *Cognition*, 267-274.

Pylyshyn, Z. W. (1980). Computation and cognition: issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3, 111-169.

Watanabe, S (1985). *Pattern recognition: human and mechanical*. New York: Wiley.