In press, *Current Directions in Psychological Science.*
Not quite final; please don't quote.

The Simplicity Principle in Human Concept Learning

Jacob Feldman[1]
Dept. of Psychology, Center for Cognitive Science
Rutgers University

Abstract

How do we learn concepts and categories from examples? Part of the answer might be that we induce the *simplest* category consistent with a given set of example objects. This seemingly obvious idea, akin to simplicity principles in many fields, plays surprisingly little role in contemporary theories of concept learning, which are mostly based on the storage of exemplars, and avoid summarization or overt abstraction of any kind. This article reviews some evidence that complexity-minimization does indeed play a central role in human concept learning. The chief finding is that subjects' ability to learn concepts depends heavily on their intrinsic complexity; more complex concepts are more difficult to learn. This pervasive effect suggests that, contrary to exemplar theories, concept learning critically involves the extraction of a simplified or abstracted generalization from examples.

Keywords: Categories, concepts, learning, simplicity, exemplars

Generalizing from experience is an essential aspect of everyday mental life. But when we make a finite number of observations of an enduring phenomenon, there is no strictly logical basis for forming any firm generalizations about it. Instead we must "induce," that is, make informed guesses, about what its general properties might be. The need for this is especially clear in the realm of *category* or *concept learning*, the process of learning categories from examples. Here we are given a few examples—say, a straight-backed chair, a plush armchair, and three-legged stool—from which we abstract or generalize to form an impression of the general category from which the examples were drawn (*chairs*). Despite centuries of inquiry into this problem, and decades of experimental research, the underlying mechanisms are not as yet fully understood.

Categories differ widely, of course, in the ease with which people can learn them from examples. Some categories—e.g. *chairs*—are easily guessed from few examples. At the other extreme, extremely disjoint or heterogeneous categories—say, an infinite set including a hat, a piano, the sun, the King of Sweden, etc.—are so incoherent and seemingly irregular that it seems *no* finite subset would suffice to communicate the essence of the category. Such a category can only be effectively represented, it seems, by simply *listing* its contents verbatim: no regularities or common trends hold sway. Such categories are "incompressible," and indeed are more difficult to learn from examples, as corroborated more formally by experiments summarized below.

Simplicity

The principle of *simplicity* or parsimony—that one should choose the simplest hypothesis consistent with the data—is one of the most ubiquitous in all fields of inference, including philosophy (as "Occam's razor"), in machine learning (under a variety of names, including the “Minimum Description Length principle”), and in visual perception (by the Gestalt term *Prägnanz* or the “minimum principle”). The principle seems particularly apt in the domain of concept learning, where it would dictate that we induce the simplest category consistent with the observed examples—the most parsimonious generalization available.

Yet, surprisingly, the idea of complexity-minimization plays very little role in contemporary theories of concept learning. Notwithstanding several early proposals (in particular Neisser & Weene, 1962), and some isolated strands in more recent literature (Medin, Wattenmaker, & Michalski, 1987; Pothos & Chater, 2001), the currently dominant models do not involve complexity-minimization in any way. One reason for this surprising neglect is the historical prominence of the dichotomy between conjunctive (*and*) and “disjunctive” (*or*) concepts, intensively studied in the 1960s (see Bourne, 1970). Conjunctive and disjunctive concepts are of equal complexity by almost any conceivable metric. Yet conjunctive concepts are easier for subjects to learn, suggesting a seemingly fundamental divergence between logical complexity and *psycho*logical complexity.

More recently, the neglect of complexity in concept learning has stemmed from the ascendancy of *exemplar theories* (e.g. Kruschke, 1992; Nosofsky, 1988). Exemplar theories model concept learning entirely via the storage of specific instances or exemplars, with new objects evaluated only with respect to how closely they resemble specific known members (and non-members) of the category. In such theories there is, by design, no representation of common tendencies in the stored exemplars; only properties of individuals are represented, without any overt generalization or abstraction. In a very literal sense, an exemplar model does not know that *water is wet*; it simply knows that some (or one, or all) stored examples of *water* have the property *wet*. Hence exemplar models may be thought of as at the most extreme philosophical contrast with complexity-minimization theories; while the latter emphasize the extraction of useful regularities, the former store examples without extracting any of the regularities that bind them together. In recent years, exemplar-based theories have achieved great empirical success (e.g. Kruschke, 1992; Nosofsky, 1988). This success has not been without controversy: for example some evidence suggests that human learners use exemplar-based strategies only early in learning, forming prototypes and generalizations later. Recently Smith and Minda (2000) have argued that the general empirical success of exemplar models is in part an artifact of the historical choice of concepts studied, many of which employ the same few concept types. But notwithstanding this disagreement, one result of the domination of exemplar models in the psychological literature has been a de-emphasis on the entire issue of complexity in concept learning. Occam plays no role in exemplar storage.

Boolean concepts

A common test-bed for theories of concept learning has been the realm of Boolean concepts, in which membership is determined by some combination of simple binary features. The concepts extensively studied during the 1960s are each conveniently depicted in a 2-dimensional grid, in which each side represents one Boolean feature, and positive members of the concept are depicted by heavy dots at the vertices (Fig. 1). This includes the already-mentioned conjunctive and disjunctive types (see types a and b in the figure), and several more exotic varieties. A famous study by Shepard, Hovland, and Jenkins (1961) went further by considering concepts with three features, each of which can be depicted in a three-dimensional cube (see Fig. 1, right). As can be seen in the figure, such concepts exhibit a wider variety of structures, and they differ greatly in their degree of learnability. In the early 1970s, studies of this kind of artificial logically-defined concept waned, as interest turned to more graded and "fuzzy" models of concepts. Yet the known variations in subjective difficulty were never satisfactorily explained. What makes some concepts intrinsically more difficult to learn than others?

--- Insert Figure 1 about here ---

Boolean complexity

One answer is that learnability of concepts is determined by their intrinsic complexity. This hypothesis had in fact been suggested by Neisser and Weene (1962), but was poorly received—in part because (as discussed above) it failed to explain the famous case of conjunction vs. disjunction, which are equally complex but differ in learnability. Moreover, the idea may have also failed to catch on because the fundamental mathematical ideas necessary to make the idea of "complexity" completely clear had not as yet been developed. Only a short time later, however, three mathematicians (Chaitin, Kolmogorov, and Solomonoff, working independently) put the mathematics of complexity—and simplicity—on a firm foundation. They proposed that complexity is, in essence, *incompressibility*. More specifically, they showed that the complexity of any string of symbols can be understood as the length of the shortest computer program that expresses the string (see Li & Vitányi, 1997). Simple strings can be expressed by short programs, while complex or random strings require long programs. The most complex case is a string so lacking in pattern or order that there's no better way to encode it than simply to *quote* it verbatim, so that the shortest program is about as long as the string itself; such a string is thus maximally random or incompressible. Measuring complexity this way, now usually known as Kolmogorov complexity, turns out to be a mathematically sound way of capturing the intrinsic complexity of the string—the degree to which it is inherently unordered and unpatterned.

In the realm of Boolean concepts, the natural analog of Kolmogorov complexity is *Boolean complexity*, defined as the length of the shortest logical expression that is equivalent to the set of positive examples (called the *minimal formula*), usually counted in terms of the number of variable names (ignoring logical connectives). As an example, imagine that we are confronted by two example objects: a big apple and a small apple. This can be thought of as a "logical formula:" *big apple or small apple*. This expression is logically equivalent to the shorter formula *(big or small) apple*, which is, in turn, equivalent to the even smaller formula *apple* (assuming that everything is either big or small). This maximally compressed form has only one variable reference in it, so the concept has Boolean complexity 1. By contrast, the concept *big apple or small orange* can't be similarly reduced—it's not equivalent to any shorter expression—so it has Boolean complexity 4 (it mentions four variables: big, apple, small, and orange). The same reduction trick can be applied to any Boolean concept, of any length. After the concept has been compressed as much as possible, the length of the shortest formula gives a measure of the concept's intrinsic complexity.

A comprehensive experiment

So how does Boolean complexity match up to the subjective difficulty of concepts? In order to answer this question, one would like to study as comprehensive a set of concepts as possible. This has not always been done. As mentioned, with the notable exception of Shepard et al. (1961), studies in the 1960s had almost exclusively considered bivariate concepts, which are severely limited in variety, forming a poor basis for generalization. Boolean concepts come in a limited variety of intrinsic "shapes" in Boolean space. For a given number of features and number of positive examples, there are really only a finite number of logically distinguishable forms (Feldman, 2003, gives a comprehensive catalog). In addition, each concept comes in two twin types, one with a smaller (or equal) number of positives than negatives, the other one complementary (i.e. with positive and negative labels switched). I refer to these as *Up* and *Down* parity respectively (for example compare concepts d and e in Fig. 1).

So in an attempt to achieve a more exhaustive survey, the study reported in Feldman (2000) considered *every* distinguishable Boolean concept that can be defined with three or four binary features and between two and four positive examples (in the Up version) as well as their complements (Down versions). The experiment sought to estimate the psychological learnability of each of these concept types, giving for each concept the proportion successfully learned after a learning session of fixed duration (see Fig. 2). The results are summarized in Fig. 3, which shows the effects of Boolean complexity and parity. As can plainly be seen in the figure, success in learning steadily decreases as complexity increases, with a roughly constant advantage for Up over Down versions.

---Insert Figure 2 about here ---

---Insert Figure 3 about here ---

Thus more complex concepts are indeed harder to learn. Altogether, Boolean complexity and parity account for more than half the variance in the data (R2 = .5017). Two prominent exemplar models (those of Kruschke, 1992 and Nosofsky, 1988) don't fare as well; each accounts for only about a quarter of the variance (R2 = .2062 and .2881, respectively). Intriguingly, each of these exemplar models, like the human subjects, exhibit worsening performance as complexity increases. Thus even though complexity-minimization is not an *overt* part of their design, they are sensitive to complexity "epiphenomenally" (that is, as a side-effect of their performance). But their inferior fit to the data shows that they don't suffer from the effects of complexity as severely as do human learners; they learn complex concepts too easily, and penalize complexity too lightly. Thus it seems that the heavy emphasis on exemplar storage in current theories is in need of re-examination. Human learning involves a critical element of compression or complexity-minimization that is not present in exemplar models.

Discussion

The main result—the complexity effect—points to a kind of simplicity principle governing human learning. As we study a set of examples, we attempt to encode them in as compact a manner as possible. The more effectively the examples can be compressed—the lower the complexity—the more successful this strategy will be, and the more effectively the examples will be retained. Thus human learners do indeed seek the simplest generalization possible, as Occam dictated.

The other result, the parity effect, suggests that subjects have some kind of complexity-independent preference for looking at concepts through their positive examples. Indeed others had noticed the same tendency long before (see Feldman, 2000 for references). The novelty here is to see this as a factor orthogonal to complexity, because seeing it this way changes the way older results are viewed—specifically the old conjunction/disjunction dichotomy. Conjunction and disjunction are actually the same type in the mathematical classification: conjunction is the Up version and disjunction the Down version. For example, the complement of the conjunction *small apple* can be expressed as the disjunctive concept *non-small or non-apple*. Thus the critical difference between conjunctive and disjunctive types does not, after all, involve complexity, but an orthogonal variable. The complexity effect is inconspicuous when comparisons are restricted to such simple bivariate forms. But when a more exhaustive range of concept types is tested, a substantial complexity effect turns out to be driving much of the variance in subjective conceptual difficulty.

Rules vs. exceptions

The idea of complexity minimization also sheds some light on how rule-formation and example storage might relate and co-exist. The dichotomy between these two styles of learning pervades cognitive science (see Hahn & Chater, 1998, for discussion). Within concept learning, some theories have explicit combined them, with one component for extracting rules, and another component for storing examples that don't fit into the rule scheme. The idea of complexity-minimization brings the essential distinction between rules and exceptions into sharper focus.

Some concepts, by their nature, reduce to a very simple rule that covers all their members (like *red things*). At the other extreme, some concepts are totally irreducible (like the one containing a hat, a piano, the sun, and the King of Sweden), meaning that their complexity is as high as it can be. As discussed above, a maximally complex concept's minimal formula consists essentially of a verbatim list of its members. In between these extremes are some concepts whose minimal formulae have a component (literally, a disjunct) that covers most of the objects, plus one or more additional objects (more disjuncts) that *aren't* covered by the "main rule." An example might be a collection of 27 red things plus a banana. The additional object or objects (e.g. the banana) are "exceptions," in that they are not covered by the main part of the rule. But they are also *part* of the rule, in that they are mentioned in the full statement of the minimal formula describing the concept. As conceptual complexity increases, concepts' optimal representations increasingly resemble explicit lists of “exceptions.”

This observation helps clarify just what the word “exception” really means. What is the intrinsic difference between rule-bound and exceptional parts of a concept? The answer is that exceptions are objects that need to be represented verbatim—listed explicitly—*even in the maximally compressed representation of the concept.* Any such object is “intrinsically” exceptional in the context of that concept. And the complexity of a concept determines how intrinsically exceptional the concept is—how much of it consists of irreducible items that need to be stored by rote.

This argument plainly suggests that exemplar models might be especially well-suited to storing highly complex concepts. Very complex concepts can’t be captured by extracting their common regularities, as prototype theories do; by definition, such concepts don’t *have* any common regularities. Rather the most efficient way to store them is verbatim, item by item, exactly as exemplar models do. This is a direct consequence of their high complexity—in fact it’s essentially the *definition* of “high complexity” in Kolmogorov’s sense. This point underscores the validity of Smith and Minda (2000)’s argument that the historical choices of concepts for study, which have favored especially complex ones, has unintentionally tilted the scales in the direction of exemplar models.

Conclusion

I have argued above that some kind of simplicity principle is an essential component of human learning. However complexity-minimization may be carried out in any number of different “codes” or representation languages. Complexity measurements taken in one code tend in the limit to be highly correlated with those taken in other codes (see Li & Vitányi, 1997), so the empirical success of one code doesn’t necessarily prove that it's the true code. The code used in the Feldman (2000) minimal formulae (based on conventional logical operators), and the associated minimization techniques (based on heuristics such as factoring) are not particularly psychologically plausible; their basic role was simply to establish the prima facie role of complexity, not to validate one particular code. Hence an essential goal for future research is to identify the underlying "cognitive code" actually employed by human learners.

A more sophisticated and psychologically-motivated answer to this question is proposed in more recent work proposing a "concept algebra" (Feldman, 2001). The basic idea here is to express inductive concepts in terms of the "regularities"—that is, patterns in the observed examples—that they obey (Feldman, 1997). Representations of concepts are then built by algebraic combinations of these atomic concepts. Complexity can be measured, as usual, by the size of the most compact representation of a given concept in the algebra. As befitting the more psychologically-motivated choice of atomic concepts, this algebraic complexity measure more predicts human performance on the Feldman (2000) dataset more accurately than does Boolean complexity (or any other known model). Another important step will be to extend the algebra to cover concepts defined over continuous features (Fass & Feldman, 2002).

Another important direction of future research is to understand the details of processing, including neural processing, by which complexity-minimization is actually carried out in the brain. There have been a number of recent advances in understanding the neural mechanisms of concept learning, but these have yet to be integrated with the newer understanding of complexity-minimization. This integration may represent the flowering of one of the oldest ideas in cognitive science: that organisms seek to understand their environment by reducing incoming information to a simpler, more coherent, and more useful form.

References

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, *77*(6), 546–556.

Fass, D., & Feldman, J. (2002). Categorization under complexity: a unified MDL account of human learning of regular and irregular categories. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing 15.* Cambridge: MIT Press.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical* Psychology, *41*, 145–170.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Feldman, J. (2001). *An algebra of human concept learning.* (Under review)

Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, *47*(1), 98–112.

Hahn, U., & Chater, N. (1998). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, *65*, 197–230.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its* applications. New York: Springer.

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1987). Constraints and preferences in inductive learning: an experimental study of human and machine performance. *Cognitive Science*, *11*, 299–339.

Neisser, U., & Weene, P. (1962). Hierarchies in concept attainment. *Journal of*

Experimental Psychology, *64*(6), 640–645.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory* and Cognition, *14*(4), 700-708.

Pothos, E. M., & Chater, N. (2001). Categorization by simplicity: a minimum description length approach to unsupervised clustering. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 51–72). Oxford: Oxford.

Shepard, R., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(1), 3–27.

Suggested readings:

Chater, N. and Vitányi, P. Simplicity: a unifying principle in cognitive science? *Trends in* Cognitive Sciences, 2003, 7(1) 19–22.

Feldman, J. (2000) Minimization of Boolean complexity in human concept learning. *Nature*, **407**, 630–633.

Li, M. and Vit´anyi, P. (1997) *An introduction to Kolmogorov complexity and its* applications. New York: Springer.

Sober, E. (1975) *Simplicity*. London: Oxford University Press.

Author Note

I am grateful to Lyle Bourne and Josh Tenenbaum for thoughtful conversations. Preparation of this manuscript was supported by NSF SBR-9875175. Portions were presented as the 2002 George Miller Award address at the meeting of the American Psychological Association in August, 2002.

Footnotes

[1] Address correspondence to Jacob Feldman, Dept. of Psychology & Center for Cognitive Science, Rutgers University - New Brunswick, 152 Frelinghuysen Rd., Piscataway, NJ 08854.

Figure Captions

Figure 1. Concepts illustrated as diagrams in “feature space.” Each axis represents one binary feature, so each vertex represents one possible combination of values of the features. Heavy black dots indicate those combinations regarded as positive examples of the concept. Concepts may be defined over two features (left) or three (right) or even more. Viewed this way, concept may seem relatively simple or relatively complex (e.g. consider concept c vs concept d). Concepts may come in Up parity (e.g. concept d) or complementary Down parity (e.g. concept e).

Figure 2. A sample “learning” screen as viewed by subjects in the Feldman (2000) experiment. The subject studied such a screen for a fixed interval, and was then tested on all of the objects in random order. The concept shown has three features and two positives, and is in in Up parity.

Figure 3. Human performance on Boolean concepts plotted as a function of their Boolean complexity, showing the increasingly impaired learning as complexity increases, and the approximately constant advantage of Up over Down versions of each concept. These two factors account for more than half the variance in the data (R^2 = .5017), much more than exemplar models applied to the same data.

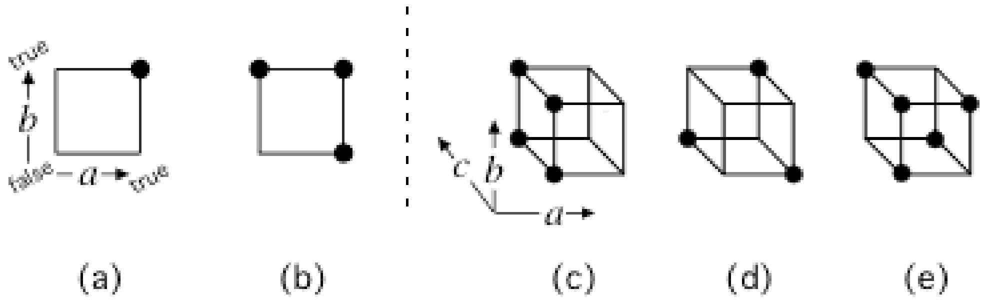


Figure 1

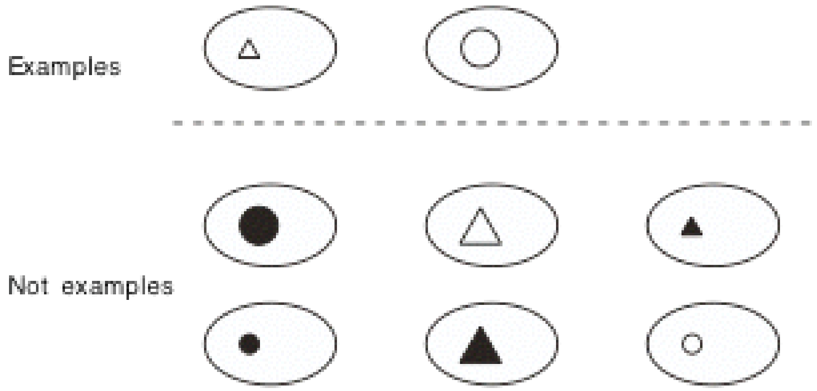
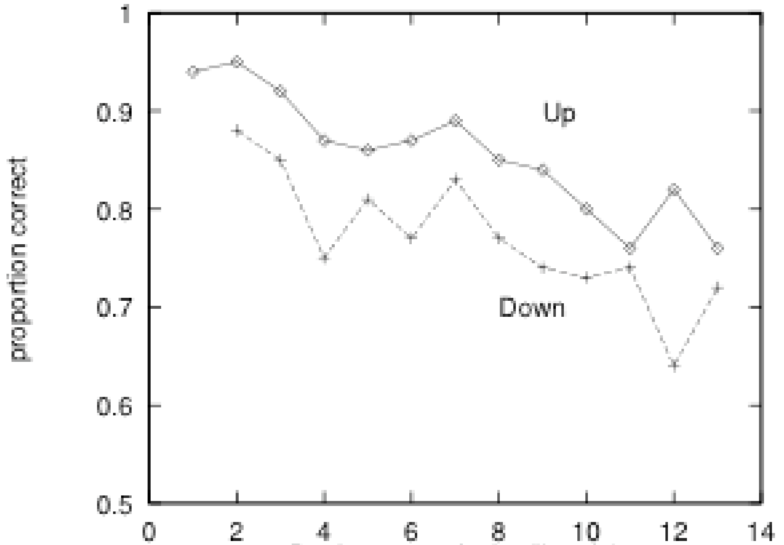


Figure 2



1
0.9
0.8
0.7
0.6
0.5
0
2
4
6
8
10
12
14
proportion correct
Boolean complexity (literals)
Up
Down

Figure 3