The Simplicity Principle in Human Concept Learning

Jacob Feldman[1]

Department of Psychology and Center for Cognitive Science, Rutgers University–New Brunswick, Piscataway, New Jersey

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 the concepts' intrinsic complexity; more complex concepts are more difficult to learn. This pervasive effect suggests, contrary to exemplar theories, that 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, what its general properties might be. The need for abstraction 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 a 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 yet fully understood.

Categories differ widely, of course, in the ease with which people can learn them from examples. Some categories—for example, *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, . . .—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 be effectively represented, it seems, only 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 I summarize later.

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 (known 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 nonmembers) 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; whereas 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, most of which were chosen from among the same few types. But notwithstanding this disagreement, one result of the domination of exemplar models in the psychological literature has been a deemphasis of 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 concept membership is determined by some combination of simple binary features. Each of the concepts extensively studied during the 1960s is conveniently depicted in a two-dimensional grid, in which each side represents one Boolean feature, and members of the concept are depicted by heavy dots at the appropriate vertices (see Fig. 1). For example, if the two features were size (small or large) and shape (square or circle), then the possible objects could be depicted by a grid in which the four vertices would represent, respectively, small squares, large squares, small circles, and large circles. Then, for example, the conjunctive concept *large squares* would be represented by a heavy dot at the large-square vertex.

The concepts extensively studied during the 1960s included the already-mentioned conjunctive and disjunctive types (see Figs. 1a and 1b), and several more exotic varieties. A famous study by Shepard, Hovland, and Jenkins (1961) went further by considering concepts with three features; each concept could thus be depicted in a three-dimensional cube (see Figs. 1c–1e). 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?

BOOLEAN COMPLEXITY

One answer to this question 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 already mentioned) it failed to explain the famous case of conjunction versus disjunction, two concepts that 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

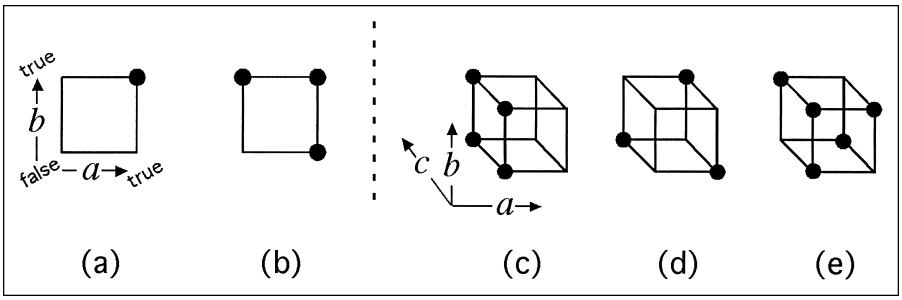

Fig. 1. Concepts illustrated as diagrams in feature space. Each axis represents one binary feature, so each vertex represents one possible combination of values of the features. For each concept, those combinations regarded as positive examples are indicated by heavy black dots. Concepts may be defined over two features (a, b), three features (c–e), or even more. Viewed this way, a concept may seem relatively simple or relatively complex (e.g., consider c vs. d). For each concept, there exists a complementary concept in which members and nonmembers have been interchanged (e.g., compare d and e). Because these two concepts have extremely similar logical structure, they are regarded as two versions, or parities, of the same type; the version with the smaller number of members is in up parity, and the other version is in down parity.

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, whereas complex or random strings require long programs. The most complex case is a string so lacking in pattern or order that there is 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. This view of complexity is now usually referred to as Kolmogorov complexity. Measuring Kolmogorov complexity turns out to be a mathematically sound way of capturing the intrinsic complexity of a string—the degree to which it is inherently unordered and unpatterned.

In the realm of Boolean concepts, the natural analogue 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*). This length is usually defined as the number (ignoring logical connectives) of variable names, called literals. For example, imagine that we are confronted by two example objects: a big apple and a small apple. This set 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* cannot be similarly reduced—it is 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? Ideally, in order to answer this question, one would study as comprehensive a set of concepts as possible. This has not always been done, however. As I mentioned earlier, with the notable exception of Shepard et al. (1961), studies in the 1960s 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 (see Feldman, 2003, for a comprehensive catalogue). In addition, each concept comes in twin types, one with a smaller (or equal) number of members than nonmembers, the other one complementary in its membership. I refer to these as concepts with *up* and *down* parity, respectively (e.g., compare the concepts illustrated in Figs. 1d and 1e). For example, the concepts *birds* and *nonbirds* are logically very similar, in that they both invoke the same categorical distinction; *birds* is the up-parity version, and *nonbirds* is the down-parity version.

In an attempt to achieve a more exhaustive survey of concepts than in earlier studies (Feldman, 2000), I considered every distinguishable Boolean concept that can be defined with three or four binary features and between two and four positive examples (up versions), as well as their complements (down versions). The experiment sought to estimate the psychological learnability of each of these concept types, determining for each concept type the proportion of objects correctly classified after a learning session of fixed duration (see Fig. 2). The results are summarized in

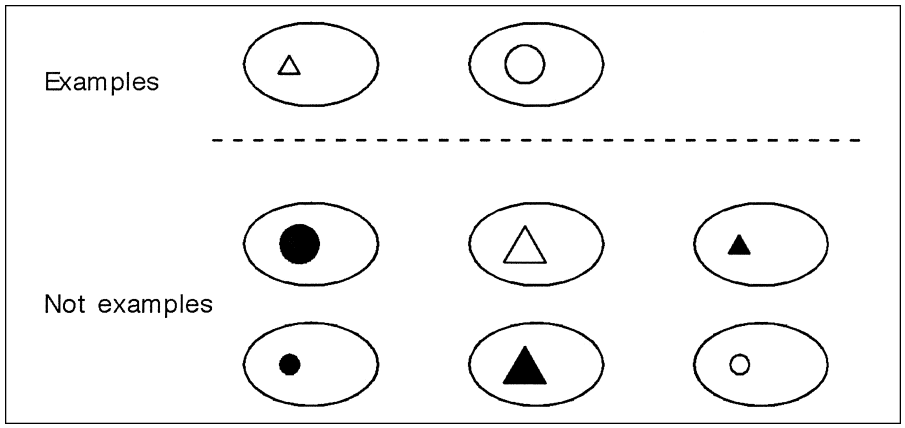

Fig. 2. A sample learning screen as viewed by subjects in the concept-learning experiment (Feldman, 2000). Subjects studied each screen for a given amount of time and were then asked whether each object had been presented as an example or a nonexample. The concept shown has three features (triangle vs. circle, small vs. big, white vs. black) and two positive examples, and is in up parity.

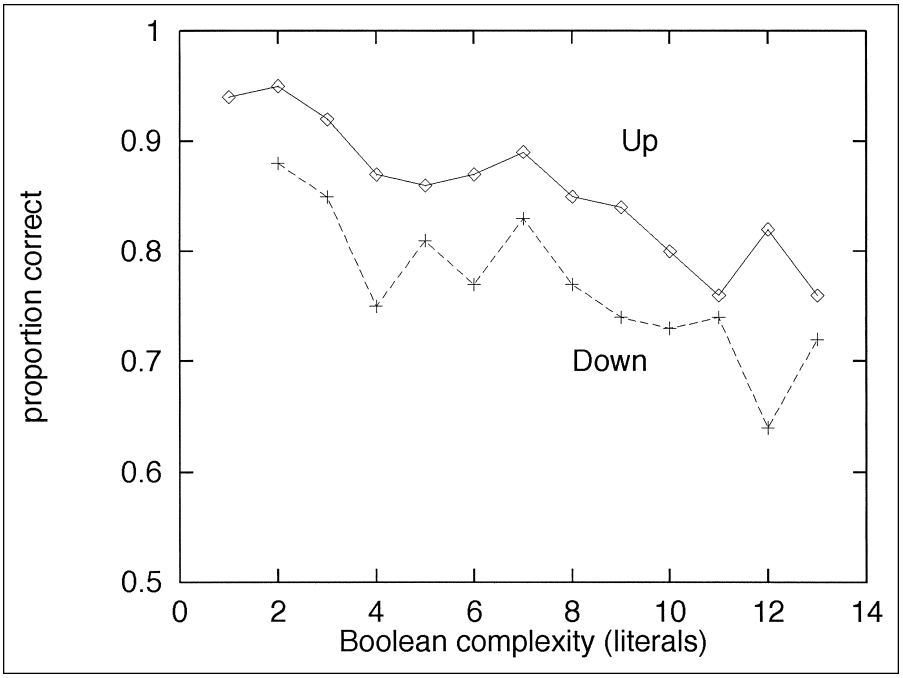

Fig. 3. Human performance on Boolean concepts plotted as a function of their Boolean complexity. Results are shown separately for the up and down versions of each concept. (From Feldman, 2000.)

Figure 3, which shows the effects of both Boolean complexity and parity. As the figure shows, success in learning steadily decreased as complexity increased, and the up versions had a roughly constant advantage over the down versions.

Thus, more complex concepts are indeed harder to learn. Altogether, Boolean complexity and parity accounted for more than half the variance in the data ($R^2 = .5017$). Two prominent exemplar models (those of Kruschke, 1992, and Nosofsky, 1988) did not fare as well; each accounted for only about a quarter of the variance ($R^2 = .2062$ and .2881, respectively). Intriguingly, each of these exemplar models, like the human subjects, exhibited worsening performance as complexity increased. Thus, even though complexity minimization is not an overt part of their design, they are sensitive to complexity epiphenomenally (i.e., as a side effect). But their inferior fit to the data shows that they are not as severely affected by complexity as human learners are; 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 reexamination. Human learning involves a critical element of compression or complexity minimization that is not present in exemplar models.

The main result of this experiment—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 remembered. 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, other researchers had noticed the same tendency long before my experiment (see Feldman, 2000, for references). Noticing that parity and complexity make independent contributions to learning changes the way older results—specifically, the old conjunction/disjunction dichotomy—should be viewed. Conjunction and disjunction are actually the same concept type in the appropriate mathematical classification: Conjunction is the up version and disjunction is the down version. For example, the complement of the conjunction *small apple* can be expressed as the disjunctive concept *nonsmall or nonapple*. Thus, the critical difference between conjunctive and disjunctive concept types does not, after all, involve complexity, but parity. 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 VERSUS EXCEPTIONS

The idea of complexity minimization also sheds some light on how rule formation and example storage might relate and coexist. The dichotomy between these two styles of learning pervades cognitive science (see Hahn & Chater, 1998, for discussion). Some theories of concept learning have explicitly combined them, including one component for extracting rules and another component for storing examples that do not 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 I discussed earlier, a maximally complex concept's minimal formula consists essentially of a verbatim list of the concept's members. In between these extremes are some concepts whose minimal formulas have a component (literally, a disjunct) that covers most of the objects plus one or more additional objects (more disjuncts) that are not 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 in fact 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 such 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. Such concepts cannot be captured by extracting their common regularities; by definition, maximally complex concepts do not *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 is essentially the *definition* of maximal complexity in Kolmogorov's sense. This point underscores the validity of Smith and Minda's (2000) argument that many of the highly complex four-dimensional concepts studied in the 1970s and 1980s unintentionally tilted the scales in the direction of exemplar models.

CONCLUSION

I have argued that some kind of simplicity principle is an essential component of human learning. However, complexity minimization may be carried out through any number of different ways of encoding (i.e., codes or representation languages). Complexity measurements taken in one code tend to be highly correlated with those taken in other codes (see Li & Vitányi, 1997), so the empirical success of one code does not necessarily prove that it is the true code. The code I used (Feldman, 2000) in my minimal formulas (based on conventional logical operators), and the associated complexity-minimization techniques, 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.

In more recent work (Feldman, 2001), I have proposed a more sophisticated and psychologically motivated code. My proposal is based on the idea that inductive concepts are expressed 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 by the size of the most compact representation of a given concept in the algebra. Given that the choice of atomic concepts is psychologically motivated, is it not surprising that this algebraic complexity measure predicts human performance in my concept-learning experiment (Feldman, 2000) more accurately than does Boolean complexity (or any other known model). Another important step will be to extend the algebra beyond features with a finite number of distinct values to cover concepts defined over continuous features (which have an infinite spectrum of values—e.g., shape and color; see Fass & Feldman, 2002, for steps in this direction).

Another important direction for future research will be to uncover 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 principle 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.

Recommended Reading

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences, 7*(1), 19–22.
Feldman, J. (2000). (See References)
Li, M., & Vitányi, P. (1997). (See References)

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

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

Note

1. Address correspondence to Jacob Feldman, Department of Psychology and Center for Cognitive Science, Rutgers University–New Brunswick, 152 Frelinghuysen Rd., Piscataway, NJ 08854.

References

Bourne, L.E. (1970). Knowing and using concepts. *Psychological Review, 77*, 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* (pp. 35–42). Cambridge, MA: 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*. Manuscript submitted for publication.

Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology, 47*, 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*, 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*, 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*, 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, England: Oxford University Press.

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*, 3–27.

When Good Pain Turns Bad

Linda R. Watkins[1] and Steven F. Maier

Department of Psychology and Center for Neuroscience, University of Colorado at Boulder, Boulder, Colorado

Abstract

Classically, pain is viewed as being mediated solely by neurons. However, recent research has shown that activated glial cells (astrocytes and microglia) within the spinal cord amplify pain. These nonneuronal cells play a major role in the creation and maintenance of pathological pain. Glia become activated by immune challenges (viral or bacterial infection) and by substances released by neurons within the pain pathway. Activated glia amplify pain by releasing proinflammatory cytokines. Taken together, research findings suggest a novel approach to human pain control that targets glia. In addition, it is likely that such glial-neuronal interactions are not unique to pain, but rather reflect a general rule of sensory processing.

Keywords

astrocytes; microglia; spinal cord; proinflammatory cytokines; hyperalgesia

One might envision that life would be lovely without pain. However, people born with a congenital insensitivity for pain bear witness that this is not so. Such people lean on hot stoves and realize it only upon smelling their burning flesh, fail to pull away from sharp objects, and are unaware of bone breaks, infections, or internal injuries, which become life threatening as a result. They learn only with great difficulty how to survive in a world full of danger.

Pain is good. Normal, everyday pain serves key biological functions. First, pain is a warning device, helping to prevent tissue damage. Pain signals carried by sensory nerves to the spinal cord trigger protective reflexes to rapidly withdraw your body from danger. In turn, spinal cord neurons relay the pain message to the brain to organize adaptive behaviors, such as swatting an offending bee. Second, pain serves a recuperative function. After injury, pain motivates you to tend to the wound, and to enter a period of inactivity and behavior that will promote healing. Thus, normal pain is highly adaptive for survival.

PAIN IS DYNAMIC

But there is more to pain. Pain is arguably the most dynamic of the senses. It is not passively relayed from the periphery of the body to the brain. Rather, it is powerfully modulated at the first synapse, at which sensory nerves relay pain in-