2 Concepts: A Tutorial

Edouard Machery

2.1 Chapter Preview

The goal of this chapter is to review, in the form of a tutorial, the current state of the research on concepts in psychology (for more extensive discussion, see Murphy 2002, Machery 2009). I describe the four principal theories of concepts, highlighting how each of them explains some important characteristics of our capacities to categorize objects and draw inductions.[1] In section 2.2, I clarify the notion of concept, and I spell out the goals of a theory of concepts. In section 2.3, I describe the so-called classical theory of concepts before presenting the prototype theories of concepts in section 2.4. In section 2.5, I turn to the exemplar theories of concepts before describing the theory theories of concepts (also called "the knowledge view") in section 2.6. Finally, in section 2.7, I examine the relations between the existing theories of concepts.

2.2 What Are Concepts?

2.2.1 The Notion of Concept

It is often unclear what psychologists and other researchers interested in concepts—including philosophers, computer scientists, and scientists working in artificial intelligence (AI)—mean by "concept." It is then difficult to assess whether researchers who put forward different views about concepts genuinely disagree or are merely talking past each other, and it is also difficult to identify the criteria that are relevant for assessing any given theory of concepts. To remedy these unfortunate outcomes, it is important to clarify what most psychologists mean by "concept" and to regiment the use of this term to the extent that this is possible. In most fields of psychology and in related disciplines (e.g., cognitive neuroscience, AI), a concept of *x* (e.g., a concept of dog) is usually taken to be a body of knowledge about *x* (e.g., dogs) that is used by default in the cognitive processes that underwrite most higher cognitive competences when we make a judgment about *x* (e.g., a judgment about dogs). Thus, a concept of *x* is a subset of the knowledge about *x* we store in long-term memory; or, to put it differently, only part of our knowledge about *x* constitutes our concept of *x*. Which part? The part that is used by default when we categorize, when we draw an induction, when we make an analogy, when we understand sentences containing a lexeme expressing the concept of *x*—in brief, when we rely on what are commonly called our higher cognitive competences or capacities (categorization, induction, analogy-making, speech production, and understanding, etc.). This body of knowledge is used by default because it is used in a context-insensitive manner: It is retrieved from memory in every context. It springs to mind, so to speak, whenever we are thinking about its referent. An example might be useful to clarify these ideas. The concept of dog is a subset of our knowledge about dogs. It is retrieved from long-term memory in a context-insensitive manner, and it is used in the processes underwriting our higher cognitive competences. We use it to decide whether to classify something as a dog, to make inductions about dogs, to understand sentences containing the word "dog," and so forth.

2.2.2 Theories of Concepts

A theory of concepts in psychology attempts primarily to identify the properties that are common to all concepts. As psychologist Gregory Murphy (2002, 2) nicely puts it:

> The psychology of concepts cannot by itself provide a full explanation of the concepts of all the different domains that psychologists are interested in. This book will not explore the psychology of concepts of persons, musical forms, numbers, physical motions, and political systems. The details of each of these must be discovered by the specific disciplines that study them. . . . Nonetheless, the general processes of concept learning and representation may well be found in each of these domains. For example, I would be quite surprised if concepts of musical forms did not follow a prototype structure, did not have a preferred level of categorization, and did not show differences depending on expertise or knowledge.

Psychologists have been particularly interested in the following five properties of concepts. First, they have tried to determine the nature of the information that is constitutive of concepts. For instance, as we will see at greater length in sections 2.4 and 2.6, some psychologists—prototype theorists—hold that concepts consist of some statistical information about the properties that are typical and/or diagnostic of a class or of a substance (e.g., Hampton 2006; Smith 2002), while others—theory theorists—insist that concepts consist of causal and/or generic information (e.g., Tenenbaum, Griffiths, and Niyogi 2007). Second, psychologists want to determine the nature of the processes that use concepts. For instance, some psychologists have argued that these processes are based on some similarity computation (e.g., Hampton 1993), while others disagree (e.g., Rips 1989). Third, cognitive scientists develop hypotheses about the nature of the vehicles of concepts. To illustrate, neo-empiricists such as psychologist Lawrence Barsalou and philosopher Jesse Prinz contend that the vehicles of concepts are similar to the vehicles of perceptual representations (e.g., Barsalou 1999, 2008; Prinz 2002). Fourth, for about a decade, cognitive scientists have attempted to identify the brain areas that are involved in possessing concepts (for recent reviews, see Martin 2007; Mahon and Caramazza 2009). Finally, cognitive scientists have developed hypotheses about the processes of concept acquisition.

As we have seen, concepts are used in the processes that underwrite our higher cognitive competences, such as induction, categorization, language production and understanding, and analogy making. By developing a theory of concepts—by explaining what kind of knowledge constitutes concepts, what kind of processes use concepts, and so on—psychologists hope to be able to explain, at least in part, how we classify objects into classes (events into event types, or samples as belonging to substances), how we draw inductions, how we make analogies, and so forth.

2.3 The Classical Theory of Concepts

2.3.1 Definitions

Until the 1970s, most psychologists held a simple view about the knowledge that constitutes a concept (e.g., Bruner, Goodnow, and Austin 1956; Conant and Trabasso 1964): Concepts were thought to be definitions (also called "rules"). According to the most common versions of this so-called classical theory of concepts, a concept of *x* represents some properties as being separately necessary and jointly sufficient to be an *x*. The concept *grandmother* is perhaps the best illustration of this approach to concepts: If people have a classical concept of grandmother, they hold that to be a grandmother it is necessary and sufficient to be the mother of a parent. Although proponents of the classical theory of concepts have done little work on the processes using concepts, it is natural to associate a simple model of categorization with this theory: When one decides whether an object is an *x* (or whether an event is an instance of an event-type or whether a sample is a sample of a given substance), one determines whether this object (event or sample) possesses the properties that one holds to be necessary and sufficient to be an *x*. The classical theory of concepts has not been used to explain how we draw inductions or how we make analogies.

Some psychologists have developed more complex versions of the classical theory of concepts. Instead of representing each property as necessary to be an *x*, a concept of *x* can consist of a representation of any Boolean combination of properties provided that this combination states a necessary and sufficient condition for being an *x*. In the following, (a) illustrates the simple versions of the classical theory of concepts, while (b) illustrates the more complex versions:

(a) Someone is a bachelor if and only if he is male, married, and adult.

(b) In baseball, a batted ball is a fair ball if and only if it settles on fair ground between home and first base or between home and third base, or is on or over fair territory when bounding to the outfield past first and third base, or touches first, second, or third base, or first falls on fair territory on or beyond first base or third base, or, while on or over fair territory, touches the person of an umpire or player.

The properties of being male, being unmarried, and being adult are each taken to be necessary, and together they are taken to be sufficient for being a bachelor. By contrast, the property of being a batted ball that settles on fair ground between home and first base or between home and third base is not necessary to be a fair ball. It is, however, necessary and sufficient that one of the disjuncts of (b) be satisfied for a ball to be a fair ball.

2.3.2 Research on Classical Concepts

Extensive research has examined how people learn classical concepts in experimental tasks. In these experiments (usually called "category learning experiments"; for a historical perspective, see Machery 2007a), participants are typically presented with artificial stimuli that constitute a category satisfying a classical concept, and they have to identify the rule that determines membership in this category. The researcher varies the presentation conditions (e.g., presence or absence of feedback; sequential vs. simultaneous presentation; presentation of noninstances in addition to instances) and nature of the rule (conjunction vs. disjunction, etc.), while measuring subjects' speed and accuracy of learning—which operationalize the difficulty of learning a category defined by a particular kind of definition.

Early work compared the learning of various types of definitions or rules. A robust result is that people more easily acquire conjunctive (*red and square*) than disjunctive (*red or square*) concepts (Bruner et al. 1956; Conant and Trabasso 1964). In addition, researchers showed that a concept is more easily learned from its instances than from its noninstances (Hovland and Weiss 1953). Finally, some researchers tried to determine a measure of conceptual complexity that would predict people's difficulty in learning more or less complex definitions or rules (Shepard, Hovland, and Jenkins 1961; Neisser and Weene 1962). Of particular importance for this latter project was a sequence of six concepts that are increasingly difficult to learn (Shepard, Hovland, and Jenkins 1961). Research on measures of conceptual complexity has in large part focused on explaining why learning these concepts is increasingly difficult.

Much of the recent work on definitions or rules has focused on finding a measure of conceptual complexity. Feldman's (2000, 2003) measure—minimal description length—has attracted much attention. He proposes that "the subjective difficulty of a concept is directly proportional to its minimal Boolean description length (the length of the shortest logically equivalent propositional formula)—that is, to its logical incompressibility" (Feldman 2000, 630). Minimal description length and another principle, called "parity" (viz., when two concepts have the same minimal description length, the concept with a smaller number of positive instances is easier to learn), explain the increasing difficulty of Shepard et al.'s sequence of concepts; it also explains half of the variance in the learning difficulty of 76 concepts developed by Feldman. From a psychological point of view, Feldman (2003, 227) interprets this result as follows:

> 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.

Recent work, however, has cast serious doubts on this proposal (e.g., Vigo 2006), and more complex hypotheses about conceptual complexity have been put forward (e.g., Feldman 2006). The problem with these hypotheses, however, is that their psychological significance is very unclear.

2.3.3 The Rejection of the Classical Theory of Concepts

Most psychologists have abandoned the classical theory of concepts since the 1970s (for some exceptions, see Nosofsky, Palmeri, and McKinley 1994; Ashby et al. 1998; Pinker and Prince 1999; Feldman 2000, 2003, 2006). Three main arguments have been put forward to justify this rejection (for a more extensive review, see Murphy 2002, chapter 2). First, some psychologists have argued that the classical theory of concepts cannot account for the vagueness of categorization—that is, for the fact that it is sometimes indeterminate whether an object is or is not a member of a class (e.g., Hampton 1993). For instance, it might be indeterminate whether some people, who have some but not much hair left on their head, are bald. However, albeit widespread, this argument is unconvincing: A conjunction of predicates might result in vague categorization judgments if the predicates are themselves vague. For instance, because "blue" is a vague predicate, it will sometimes be indeterminate whether something is a blue square, although the concept of a blue square is a classical concept.[2]

Second, suppose that a concept is defined by means of another. For example, people could represent the action of murdering as the action of killing intentionally that also meets some other conditions. Prima facie, this predicts that processing the concept of murdering would take longer than processing the concept of killing. However, Fodor et al. (1980) have shown that this is not the case: These two concepts are processed at the same speed.

Third, psychologists discovered in the 1970s several properties of our categorization decisions that are not explained by any version of the classical theory of concepts—particularly the so-called typicality and exemplar effects (see below).

2.4 Prototype Theories of Concepts

Prototype theories of concepts reject the idea that concepts represent some properties (or Boolean combination of properties) as being necessary and sufficient. They typically propose that concepts are prototypes, and that

Table 2.1
The prototype concept of vehicle (Hampton 1979, 459).

Vehicle
1. Carries people or things
2. Can move
3. Moves along
4. Has wheels
5. Is powered, has an engine, uses fuel
6. Is self-propelled, has some means of propulsion
7. Is used for transport
8. Is steered, has a driver controlling direction
9. Has a space for passengers or goods
10. Moves faster than a person on his own
11. Is human-made

a concept of *x* represents either the properties that are typical of category members, the properties that are diagnostic of them, or the properties that best weigh typicality and diagnosticity. A property is typical if the probability that a particular possesses this property if it belongs to the category is high, whereas a property is diagnostic if the probability that a particular belongs to the category if it possesses this property is high. So, for instance, a prototype of dogs could represent dogs as being furry, as barking, and so on.

There are various prototype theories (Hampton 2006). The simplest theories (e.g., Hampton 1979; see table 2.1) assimilate prototypes to lists of typical properties. More complex theories (e.g., Smith et al. 1988; see table 2.2) are related to frame theories (Barsalou 1992) in that they distinguish attributes from values. Attributes (e.g., colors, shapes) are kinds of properties: They determine that the members of a category possess a property of a particular kind. For instance, apples are represented as having a color. Values (e.g., red, green, brown) are the properties possessed by the category members. The weight of an attribute represents the importance of this attribute for deciding whether an object is a category member, whereas the weight of a value represents the subjectively evaluated frequency of this particular value among members.

The two theories of prototypes briefly described represent prototypes by means of schemas, whereas other prototype theories represent prototypes

Table 2.2
The prototype concept of apple (Smith et al. 1988, 490).

Apple			
Attributes		Values	
Color	1	Red	27
		Green	3
		Brown	—
Shape	0.5	Round	25
		Cylindrical	5
		Square	—
Texture	0.25	Smooth	24
		Rough	4
		Bumpy	2

as points in multidimensional spaces (Gärdenfors 2000). These two ways of characterizing prototypes differ in how similarities between prototypes and other representations (e.g., the representations of the objects to be categorized) are computed (for discussion, see Storms 2004).

2.4.1 Categorization and Category Learning

In contrast to the classical theory of concepts, prototype theories of concepts are associated with relatively precise models of the processes underlying various cognitive competences, including categorization (Hampton 1993; Smith 2002), induction (Osherson et al. 1990; Sloman 1993), and concept combination (Smith et al. 1988). As an illustration, I review Hampton's (1993) model of categorization before reviewing some of the phenomena that prototype theories are taken to explain (see also Murphy 2002, chapter 2; Hampton 2006; Machery 2009, chapters 4–7).

Hampton's model consists of a prototype model of concepts, a similarity measure, and a decision rule. This prototype model of concepts is similar to the one by Smith et al. (1988) described above. Following Hampton (1993, 73–74), the similarity measure, $S(x,C)$, of an instance x to a category C is defined in terms of valuations $w(x,i)$, each of which is the weight of the value (e.g., red) possessed by x for attribute i of the prototype (e.g., color). A particular similarity measure is defined by some specific way of aggregating the weights $w(x,i)$ for all relevant attributes. For example,

$$S(x,C)=\sum_{i} w(x,i). \tag{2.1}$$

This means that prototype models typically assume that categorization judgments are influenced by the properties taken independently from one another. Their configuration does not matter. Or, to put the point differently, in these models, categorization cues are independent.

Hampton's decision rule for categorization is a simple deterministic rule,

$$S(x,C)>t \Rightarrow x \in C, \tag{2.2}$$

where *t* is a criterion (or threshold) on the similarity scale. Nondeterministic decision rules can also be used, and this rule can be modified to explain how people decide whether to categorize an object in one of two categories.

Thus, Hampton's model of the categorization process involves a matching process between representations—namely, the prototype and the representation of the object to be categorized—as well as a linear measure of the similarity between the prototype and other representations. These are trademark characteristics of prototype models of cognitive processes. Hampton's model also assumes that the same process of similarity evaluation underlies both typicality judgments (how typical an object is of its category) and categorization judgments. Typicality ratings are supposed to be monotonically related to similarity.

This type of model accounts for the typicality effects identified at the end of the 1960s and in the 1970s (Posner and Keele 1968, 1970; Rips, Shoben, and Smith 1973; Rosch and Mervis 1975; Hampton 1979). Typicality—the extent to which an object possesses the properties that are typical of a category—has repeatedly been shown to have an extensive influence on people's performances in a range of cognitive tasks. Typicality can be measured objectively for artificial categories (e.g., Posner and Keele 1968, 1970; Rosch and Mervis 1975, experiments 5 and 6); it can be measured by asking people to list the properties of instances of the relevant categories (Rosch and Mervis 1975, experiments 1–4; Storms 2004); or it can be estimated by asking people to judge how good an example a particular object is ("typicality judgments") (e.g., Rosch and Mervis 1975, experiments 1–4; Hampton 1979, 1981).

Rips, Shoben, and Smith (1973) found that typical category members are classified more quickly and more accurately than atypical category members (see also Hampton 1979; for review, see Murphy 2002, chapter 2):

Participants respond more quickly to “a robin is a bird” than to “an ostrich is a bird.” Similar results are obtained when the stimuli are presented visually, for instance, when participants are shown a picture or a drawing of the object to be categorized, such as a drawing of a robin (Murphy and Brownell 1985). Similar findings are also found with artificial categories (Rosch and Mervis 1975, experiments 5 and 6).

Typicality with respect to a category predicts the likelihood of being considered a member of this category (Hampton 1979). A similar result has been found in linguistics. Labov (1973) has shown that, in American English, artifacts are called “mug” or “bowl” to the extent that they are similar to a prototypical shape.

Typicality also affects concept learning. Using artificial stimuli, Posner and Keele (1968, 1970) have shown that, following the acquisition of a concept, the most typical member of the category is sometimes more likely to be classified as a category member than the category members seen during training, although this most typical member has not been seen during training. In experiments with artificial categories, participants learn the category membership of typical items faster than the category membership of atypical items (Rosch and Mervis 1975). Participants also more easily learn to classify items in a category if they are trained with typical items than if they are trained with atypical items.

The findings reviewed so far are consistent with the prototype theories of concepts.[3] Since the representation of a target is supposed to be matched with a prototype during categorization, theories of prototype-based categorization expect typicality to affect categorization. Because concept learning consists in forming a prototype, prototype theories also expect typicality to affect concept learning.

The idea that typicality effects support prototype models of concepts has been challenged from several directions. First, Armstrong, Gleitman, and Gleitman (1983) have argued that typicality effects do not show anything about conceptual structure because they are also found with concepts that satisfy the classical theory of concepts (for critical discussion, see Machery 2009, chapter 6).

Second, Barsalou (1985) has shown that typicality judgments (“how good a bird is this robin?”) are not merely influenced by typicality (robins are typical birds), but also by how frequently a category member is encountered as a category member (e.g., how frequently robins are encountered and viewed as birds) and by how similar a category member is to an ideal member of a category (how similar a robin is to an ideal bird). These findings raise a problem for prototype theorists because these theorists support prototype theories by appealing in part to the fact that typicality, as measured by typicality judgments, predicts performance in experimental tasks (for critical discussion, see Hampton 1997; Machery 2009, chapter 6).

Third, exemplar theories (see section 2.5) can account for many typicality effects (Medin and Schaffer 1978). It is thus unclear whether the typicality effects found in the 1960s and 1970s support prototype theories over exemplar theories. More recent research suggests that whether a prototype or an exemplar is learned in category learning experiments depends on the category structure (number of category members presented during training, similarities between category members, dissimilarities between various categories) and on the stage of category learning (Smith and Minda 1998, 2000; Minda and Smith 2001; Nosofsky 2000; Nosofsky and Zaki 2002; Smith 2002; Zaki and Nosofsky 2007).

2.4.2 Induction

In addition to the tasks related to categorization and category learning, typicality effects are also found in categorical induction tasks (Murphy 2002, chapter 8; Sloman and Lagnado 2005; Machery 2009, chapter 7). In such tasks, people have to infer whether the members of a category (the target category) possess a property on the basis of being told that the members of another category or of other categories (the source category or categories) have this property. For instance, participants might be asked whether sparrows have sesamoid bones given that robins have sesamoid bones, or, equivalently, how good the following inference is:

(a) Robins have sesamoid bones.
____
Hence, sparrows have sesamoid bones.

Several findings show that typicality influences people's inductions. Consider first "the similarity effect." A conclusion that is inferred from a single premise is judged to be stronger to the extent that the source category is judged to be more similar to the target category (Rips 1975; Osherson et al. 1990). Thus (a) is a better inference than the following one:

(b) Robins have sesamoid bones.
____
Hence, penguins have sesamoid bones.

Consider also "the typicality effect" (Rips 1975). A conclusion that is inferred from a single premise is judged to be stronger to the extent that the source category is typical of the target category (if the target category includes the source category) or of the category that includes both the target category and the source category (if the target category does not include the source category). Consider, for instance, the following inferences:

(c) Robins have sesamoid bones.
Hence, birds have sesamoid bones.

(d) Penguins have sesamoid bones.
Hence, birds have sesamoid bones.

Inference (c) is judged to be stronger than inference (d) because robins are a more typical kind of bird than penguins.

Two well-known models of the processes involved in induction explain the similarity and typicality effects (as well as other effects) by assuming that we retrieve from memory the prototypes of the source categories and of the target category (Osherson et al. 1990; Sloman 1993). For the sake of space, I review only Osherson et al.'s (1990) similarity-coverage model. In this model, the strength of the induction is a function of the average similarity between the source categories and the target category and of the coverage of the source categories, defined as the average similarity between the source categories and either the typical subclasses of the target category—when the target category includes the source categories—or the typical subclasses of the lowest-level category that includes both the source and target categories—when the target category does not include the source categories. Similarity is determined by matching the relevant prototypes. The similarity effect falls out from the similarity component in the model. The typicality effect is a consequence of the coverage component of the model because the typicality of a category *x*, such as robins, with respect to a more inclusive category *y*, such as birds, is correlated with the similarity between the prototype of *x* and the prototypes of the typical subclasses of *y*.

2.5 Exemplar Theories of Concepts

Exemplar theories (Brooks 1978; Medin and Schaffer 1978; Nosofsky 1992) reject the idea that, when people acquire a concept, they abstract some statistical information about the represented class (e.g., information about typical or diagnostic properties). Rather, they propose that people store representations of particular category members (a representation of this kind is called "an exemplar"), and that they use these representations to make categorization judgments, to draw inductions, and so on. So, for these theories, a concept of dogs consists in a set of representations of particular dogs (say, a representation of Fido, a representation of Rover, etc.), which are used in the cognitive processes underlying our higher cognitive competences. Medin and Schaffer (1978, 209–210) have well captured the gist of the exemplar theories:

> The general idea of the context model [the name of their model] is that classification judgments are based on the retrieval of stored exemplar information. . . . This mechanism is, in a sense, a device for reasoning by analogy inasmuch as classification of new stimuli is based on stored information concerning old exemplars. . . . Although we shall propose that classifications derive from exemplar information, we do not assume that the storage and retrievability of this exemplar information is veridical. If subjects are using strategies and hypotheses during learning, the exemplar information may be incomplete and the salience of information from alternative dimensions may differ considerably.

Because concept acquisition does not require abstraction (or, at any rate, requires less abstraction) according to exemplar theories of concepts, learning turns out to be simpler on these views. On the other hand, because cognizing involves retrieving from long-term memory numerous singular representations (exemplars) and using them in cognitive processes (e.g., in the process underlying categorization), whereas prototype theories propose that cognizing involves retrieving and using a single representation, cognitive processing is more computationally intensive according to exemplar theories. Another difference between prototype and exemplar theories is that prototype theories assume that categorization judgments—judgments to the effect that something is an *x*, for instance a dog or a table—and recognition judgments—judgments identifying an individual as an individual, e.g., the judgment expressed by "This is John"—involve two distinct kinds of representation (respectively, prototypes and representations of particulars), whereas exemplar theories propose that both types of judgments involve a single kind of representation (i.e., exemplars).

Most exemplar theories have been developed in a spatial framework (e.g., Nosofsky 1992; but see Storms 2004). Exemplars are represented as points in a multidimensional space, whose dimensions represent the