Concept Acquisition by Autonomous Agents:
Cognitive Modeling versus the Engineering Approach

Paul Davidsson
Department of Computer Science
University of Lund
Box 118, S–221 00 Lund, Sweden
Paul.Davidsson@dna.lth.se

Abstract

This paper is a treatment of the problem of concept acquisition by autonomous agents, primarily from an AI point of view. However, as this problem is not very well studied in AI and as humans are indeed a kind of autonomous agent, the problem is also studied from a psychological point of view to see if the research in human concept acquisition can be of any help when designing artificial agents. However, the acquisition cannot be studied in isolation since it is dependent on more fundamental aspects of concepts. Consequently, these are studied as well. Thus, this paper will give a review of some of the research done in cognitive psychology and AI (and to some extent philosophy) on different aspects of concepts. Some proposals for how central problems should be attacked and some pointers for further research are also presented.

1 Introduction

In order to pursue goals and to plan future actions efficiently, an intelligent system (human or artificial) must be able to classify and reason about objects, behaviour and events. As a basis for this it needs *concepts*. Moreover, since the system's environment often cannot be totally predicted in advance, all necessary knowledge about the environment cannot be innate (humans) or preprogrammed (artificial systems). Thus, it must also be able to *acquire* concepts.

In some sense we can say that concepts refer to categories, where we by category mean a class of entities united by some principle(s). Such a principle may be rather concrete, like having similar perceptual characteristics, or more abstract, like having the same role in a theory or having similar functions.

The use of the term "concept" is rather ambiguous. In everyday language it usually refers to the name[1] (designator) of a category. However, the main concern of this paper is not the linguistic task of learning the names of categories.[2] Therefore, I will in what follows use the term in a different way, more in line with uses in Cognitive Psychology and in AI. In Cognitive Psychology the term "concept" sometimes refers to the mental representation of the category, and in AI the term often refers to the description of the intension[3] of a concept. In short, I will by "category" refer to a class of entities in the world and by "concept" refer to an agent's internal representation of this class.

There are two major approaches to the studying of computational concept acquisition (Michalski, 1987a) :

- *Cognitive modeling*, which strives to develop theories of the actual concept acquisition processes in humans (or animals). (i.e. Cognitive Psychology)
- *The engineering approach*, which attempts to explore all possible concept acquisition mechanisms, irrespective of their occurrence in living organisms. (i.e. AI)

The engineering approach has often been successful in acquiring artificial concepts in restricted domains, but in more realistic scenarios success has been limited. One such scenario, probably the most natural, general and

[1] A word in a *natural* language

[2] As a matter of fact, this paper is not about traditional linguistics at all (not deliberately at least). Thus, I make an assumption that concepts are independent of language (or at least can be studied independently of language). By language I mean here a natural language, not an internal (mental) language (language of thought, mentalese). The rejection of any assumption of language as a prerequisite for concepts has been made by a number of scientists, for instance Edelman (Edelman, 1989) who cites chimpanzees as an example of animals which lack linguistic abilities but can have and acquire concepts.

[3] That is, if we by intension mean something that includes criteria for determining category membership, but not properties that specify the concept's relations to other concepts (which sometimes is included in the concept of intension).

realistic, is a concept-learning *autonomous agent* acting in a real-world environment. It is from this perspective that I propose to view the problem of concept acquisition.

1.1 Autonomous Agents

An autonomous agent can be seen as a system that is capable of interacting independently and effectively with its environment via its own perceptors (transducers, sensors) and effectors in order to accomplish some given or self-generated task.[4] As I see it, an ultimate goal for AI is to construct intelligent autonomous agents. However, at the moment we are quite far from achieving this goal, partly because the problem of learning has not been solved. Learning is important because without it an agent is unable to adapt to unforeseen situations, and cannot take advantage of its experience in order to increase its performance. An important kind of learning is the acquisition of concepts. The sub-field of AI that studies learning is called Machine Learning (ML).

All computer-based autonomous agents have, more or less, the same basic architecture, which is illustrated in Figure 1.[5] The arrows in the figure symbolize data flows. The perceptors receive input from the environment and provide (sensor) data for the reasoner. The reasoner refines this data and stores it in the knowledge base (KB). When the reasoner is planning or doing other kinds of reasoning, it retrieves the appropriate information from the KB. When the reasoner has decided which action to perform it commands the effectors to carry out the action.

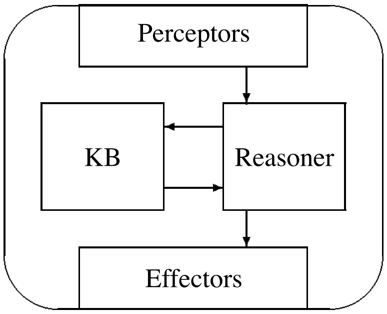

Figure 1: The basic architecture of a computer-based autonomous agent.

As mentioned above, this is a very rough description. Usually, the reasoner is further divided into several independent parts. Moreover, the reasoner may control the perceptors.

1.1.1 Scenarios for an Autonomous Agent

There are in principle two scenarios for an autonomous agent.

- The agent is alone in its environment.
- There are other agents in the environment.

An agent on an exploration mission on Mars is a (prototypical) instance of an agent being alone in its environment. In the case where other agents exist they can either be humans or machines (or both), as on a factory floor. When I say that other agents exist I suppose that our agent is able to communicate with them[6], otherwise it can be seen as being alone in its environment, for instance, an agent on a factory floor that cannot communicate with the other workers. Another possible scenario is a combination of these two scenarios, where the agent is trained by other agents in an initial stage and then put in an environment where it is alone. For instance, first trained at NASA and then sent away to Mars.

Which of these situations the agent is placed in has, of course, consequences for how concepts can be acquired.

1.1.2 Some Consequences of the Autonomous Agent Perspective

Most of our information about concepts, we get from some kind of observation. Since we are here dealing with autonomous agents that receive information directly from the environment we need another notion of observation than the one that is often used in AI (and in ML in particular), where an observation is a linguistic description in some specified language. It might be useful to follow Gärdenfors (Gärdenfors, 1992), who distinguishes three levels of describing observations.[7] The highest level is the already mentioned *linguistic* level, where observations are described in some language. The second level is the *conceptual* level where observations are not defined in relation to some language. Rather, they are characterized in terms of some underlying *conceptual space*, which consists of a number of *quality dimensions* (some of these dimensions are closely related to what is produced by our sensory receptors, like temperature, colour and size, while others are more abstract, like time). On the conceptual level an observation can be defined as "an assignment of a location in a conceptual space". The lowest level is the *subconceptual* level where observations are characterized in terms of the "raw" (not processed in any way) inputs from sensory receptors. This input, however, is too rich and unstructured to be useful in any conceptual task. It must be transformed to suit the conceptual

[4] Thus, humans (most of us, at least) can be regarded as autonomous agents.

[5] Except from some primitive reactive agents, such as Pengi (Agre and Chapman, 1987) and system based on neural networks, that do not have explicitly separated reasoners and knowledge bases.

[6] To reduce the complexity I will in what follows suppose that communication is done, more or less, on the agent's conditions (terms). Thus, I will not bother about such difficult topics as natural language understanding.

[7] A similar distinction is made by Harnad (Harnad, 1987).

or the linguistic level. To do this the subconceptual information must be organized and its complexity reduced. The work of Musgrove and Phelps (Musgrove and Phelps, 1990) is one of the few ML-approaches that treat observations on the subconceptual level. Harnad (Harnad, 1987) calls learning based on such perceptual observations *learning by acquaintance* and contrasts it with *learning by description* that bases the learning on observations on the linguistic level.[8]

This three-level perspective raises several more or less philosophical questions. For instance, is it convenient or even possible, to do any useful cognitive tasks (e.g. reasoning) on a sub-linguistic level, or do we need some kind of language?[9] If we assume that some kind of language is necessary for reasoning (as I tend to do), then observations must be described on some linguistic level in the reasoner. It seems clear that at some (early) stage in the perceptors observations must be described on the subconceptual level. Concerning the conceptual level descriptions, it is not that obvious where they should belong. The transitions from subconceptual to conceptual and from conceptual to linguistic descriptions might be done in the perceptors or in the reasoner or one transition in each.

Moreover, it is not computationally tractable to process all the input data. The system needs to know what input information is relevant, some kind of mechanism that controls the focus of attention. This gives rise to a further question: on which level is it most appropriate (or even possible) to have such a mechanism? Is the “filter” between the subconceptual and the conceptual level sufficient or do we need further “filters” at higher levels?

Finally, it seems reasonable to suppose that an autonomous agent encounters only a few instances of some categories. Thus, it needs to learn a rather good concept representation based on relatively few instances. This is in contrast to trying to learn a perfect description from a vast number of instances, which is often the task in traditional AI, where it is often supposed that the learning system has a lot of examples to learn from.

1.2 Motivation and Disposition

Since humans obviously are autonomous agents that are able to acquire concepts by interacting with the real world, it may be a good idea to become inspired by the research done in cognitive modeling. However, when adopting this approach, there are some things we must keep in mind. The most important is perhaps that the task of creating an autonomous agent is not identical with cognitive modeling. For instance, if there are no advantages of adopting a particular feature of human concept acquisition then we have no reason to do so (humans are not optimal). On the other hand, a cognitive model does not have to be a true model of human cognition to be useful in AI. Thus, I will not care about the biological and psychological plausibility of the cognitive models presented in this paper. Moreover, it is important to remember that we only have a limited insight into the ways in which humans really acquire concepts. A problem from the AI point of view is that in experimental psychology, cognitive processes concerning concept acquisition are often not described algorithmically.

Since the topic of concepts is rather complex, I have chosen to divide it into four aspects:

1. The functions of concepts
2. The nature of categories
3. Representation of concepts
4. Concept acquisition processes

The primary focus of this paper is the acquisition of concepts, but this is clearly dependent on how concepts are represented. The representation, in turn, is dependent on what functions the concepts should serve. Moreover, the choice of representation is constrained by the nature of the actual categories. Thus, we have to examine these aspects also.

The goal of this paper is to review some of the research done in Cognitive Psychology on the four aspects and the work of the corresponding aspects in AI, and hopefully to compare them in a meaningful manner. Moreover, by doing this I hope to be able draw some conclusions for the construction of computer-based autonomous agents. However, already at this point the reader should be warned that these conclusions in some cases will be rather speculative (and maybe naïve).

2 The Functions of Concepts

Why does an agent need to have concepts in the first place? In this section I am going to try to answer this question, mainly by investigating what functions concepts have, or should have.

2.1 The Functions of Human Concepts

To begin with, we can state that concepts seem to be the very stuff out of which reasoning and other cognitive processes have as their basis. However, it is possible to distinguish several functions of human concepts, some of them are[10]:

- Stability functions
- Cognitive economical functions

[8] This distinction is probably inspired by Russell’s (Russell, 1912), but is not equivalent to his.

[9] Here “language” refers to a wider class than the class of natural languages, including internal (mental) languages. Thus, this question raises another still more fundamental question: What do we mean by a language?

[10] The list is inspired by works in cognitive psychology, linguistics and philosophy, in particular (Rey, 1983) and (Smith, 1988)

- Linguistic functions
- Metaphysical functions
- Epistemological functions
- Inferential functions

Concepts give our world *stability* in the sense that we can compare the present situation with similar past experiences. For instance, when confronted with a wasp,[11] we can compare this situation with a situation some years ago when we were stung by another wasp and consequently take the appropriate measures. Actually, there are two types of stability functions, intrapersonal and interpersonal. Intrapersonal stability is the basis for comparisons of cognitive states within an agent, whereas interpersonal stability is the basis for comparisons of cognitive states between agents.

By partitioning the world into categories, in contrast to always treating each individual entity separately, we decrease the amount of information we must perceive, learn, remember, communicate and reason about. In this sense we can say that categories (and thus concepts) promote *cognitive economy*. For instance, by having one representation of the category wasp instead of having a representation for every wasp we have ever experienced, we do not have to remember that the wasp we saw yesterday has a stinger (or that it has guts).

The *linguistic function* is mainly providing semantics for linguistic entities (words), so that they can be translated and synonymy relations be revealed. For instance, the fact that the English word "wasp" and the Swedish word "geting" have the same meaning enables us to translate "wasp" into "geting" and vice versa. Furthermore, it seems that it is the linguistic function together with the interpersonal stability function that make it possible for us to communicate (by using a language).

In philosophy, metaphysics deals with issues concerning how the world is, while epistemology deals with issues concerning *how we know* (believe, infer) how the world is. Thus, we might say that the *metaphysical functions* of a concept are those that determine what makes an entity an instance of a particular category. For example, we say that something actually is a wasp if it has a particular genetic code or something like that.[12] The *epistemological functions* then, are those that determine how we decide whether the entity is an instance of a particular category. For instance, we recognize a wasp by colour, bodyshape and so on.[13]

Finally, concepts allow us to *infer* non-perceptual information from the perceptual information we get from an entity, and to make predictions concerning it. Thus, we can say that concepts enable us to go beyond the information given. For instance, by perceptually recognizing a wasp we can infer that it is able to hurt us.

2.2 Functions of Concepts in Artificial Autonomous Agents

The functions of concepts are never really subject to discussion in the AI-literature. However, there is often an assumption made that the concepts acquired are to be used for some classification task. Thus, the function is mainly of an epistemological nature. The reason for this limitation is probably that AI researchers often do not study problems from an autonomous agent perspective. Consequently, they, in some respect, lose the wholeness of the problem. Therefore, it seems that I have to base the discussion on my own reflections.

The functions of intrapersonal stability and cognitive economy are of course important, but they are trivial in the sense that they emerge more or less automatically for the agent just by having concepts, independently of the choice of representation. By analogy with the stability functions, we can say that an agent can have both intrapersonal and interpersonal linguistic functions. Where the intrapersonal function is a rather weak one, implied only by the fact that the concepts have names internal to the agent. This function is, of course, also trivial in the same sense as above. But what about the interpersonal stability and linguistic functions? They are clearly not necessary in a one-agent scenario. However, if we are interested in a multi-agent scenario with communicating agents, the concepts must have also these functions.

However, it is the remaining three functions, the metaphysical, the epistemological and the inferential, that are the most interesting, and the ones I will concentrate on in the remaining part of this paper. Since an autonomous agent should be able to classify objects in ordinary situations, the epistemological function is necessary. The metaphysical functions can of course be useful for an agent to have, but in most cases it seems that it can manage without them. Finally, if the agent is to be able to reason and plan about objects it is necessary that it have at least some inferential functions.

3 The Nature of Categories

What can be said about categories in general? In this section I will try to answer this question, not taking into account how concepts are represented or acquired.

[11] Here, and in the following, "wasp" refers to the most common kind, the black- and yellow-striped wasp (or yellow-jacket).

[12] We use the word "metaphysic" in a more pragmatic way than in philosophy. In my notion that which makes an entity an instance of a particular category is decided by some kind of consensus amongst the agents in the domain. The example with the wasp is rather unfortunate though, since there exist several competing views of biological taxonomy. For instance, the cladistic view, which categorize species according to shared derived features, and the phenetic view, which categorize them on the basis of overall similarity.

[13] For human concepts this distinction is maybe not as clean-cut and unproblematic as described here, see (Lakoff, 1987), but nevertheless it suites my purposes very well.

3.1 The Nature of Human Categories

To begin with, we should make a distinction between categories that we normally use and *artificial* categories. Artificial categories are typically categories that are constructed for a particular psychological experiment,[14] whereas *natural* categories are those that have evolved in a natural way through everyday use. Artificial categories are constructed to be specified by a short and simple definition in terms of necessary and sufficient conditions, while this is not always possible with natural categories. Until quite recently cognitive psychologists have studied the different aspects of concepts using only artificial categories. We who investigate psychological theories in order to build machines that are able to learn concepts efficiently, find this state of affairs rather unfortunate, since machines learn artificial concepts relatively easily. However, during the last decades it has become apparent that this approach does not have much to say about how humans really acquire concepts. Therefore, some researchers have begun to use natural categories for their experiments. This movement towards a more ecologically sound approach is further elaborated in (Neisser, 1987).

Natural categories can have members that are either concrete, such as physical objects, or abstract, such as emotions. In what follows I will concentrate on concrete object categories. (This is probably difficult enough.) While dealing with autonomous agents trying to learn about their environment, it is also a quite natural initial assumption. As we will see later, humans use other types of categories that cannot be classified as natural, namely *derived* (Smith, 1986) or *ad-hoc* (Barsalou, 1986) categories. Moreover, natural categories can be divided into *natural kinds* and *artifacts* (Smith, 1986). However, let us begin by examining a more fundamental topic: the properties of objects.

3.1.1 Properties

It is commonly accepted that the basis for the representation and categorization of an object is the properties[15] that characterize the object. Some properties are *perceptual*, in the sense that they (in some sense) are directly available from the perceptual system, while others are more abstract (functional, for instance). Furthermore, some features are structural (for instance, a table has legs). However, what is considered a feature is relative. Some features can be thought of as categories themselves (not necessarily a concrete object category, though). As a matter of fact, we have a kind of tree-hierarchy of categories (where the leaves are just features).[16] An example of such a hierarchy is shown in Figure 2. (Observe that only a small part of the hierarchy is shown. Apples obviously have more than three properties.)

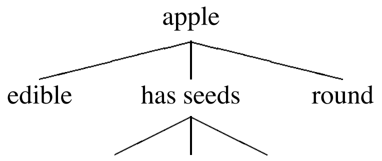

Figure 2: Part of the property-hierarchy of the category "apple".

Some of the leaves (the branch end-points, edible and round) of the resulting tree structure are perceptual features (round).[17] These are the features we normally use to determine which category an object belongs to. They are sometimes said to constitute a considerable part of the *identification procedure* (Smith and Medin, 1981) and are closely related to the epistemological function. Moreover, the features on the second level (edible, has seeds, round) are called the *core* of the concept and are together with the rest of the features more connected to the metaphysical and inferential functions.[18] It is also important to notice that these hierarchies imply that most categories are, in some way or another, dependent on other categories.

Given the distinction between perceptual and abstract features we can say that concepts allow us to go beyond the information given and thus make possible the inferential function. Because, once we have assigned an object to a class on the basis of its perceptual features, we can infer or predict its non-perceptual features.

3.1.2 Natural Kinds and Similarity

It seems natural to assume that categories emerge as a consequence of the correlational structure of the environment, where the properties of the instances of a category make them stand out as a natural class, distinct from other categories. For instance, take the situation where you see an elephant for the first time. Because of its distinct (perceptual) features, you create a new category. Moreover, if you see another elephant you decide that it belongs to the same category because of the features it shares with the first one. Quine (Quine, 1969) has termed this type of category *natural kinds*. Rosch and her colleagues (Rosch et al., 1976) also emphasized that natural categories emerge in this way, assuming

[14] In some sense also some scientific categories, such as mathematical categories, are artificial.

[15] Sometimes a distinction is made between *quantitative* properties (dimensions), such as temperature in °C, and *qualitative* properties (features), such as hot and cold. Since it is possible to transform a qualitative property into a quantitative and vice versa, I will in most cases not hold on to this distinction, and just speak of properties (sometimes however calling them features or attributes).

[16] This structure is rather messy though. It is mostly the structural features (has seeds) that can be thought of as a category. (Actually, it is "seed" that are the category.) The functional features (edible) and perceptual features are best thought of as just features.

[17] Whether round is a perceptual feature or not may be open for discussion, but let us suppose that it is.

[18] Or like Smith (Smith, 1988) (p.29) puts it: "When reasoning we use the cores, when categorizing we use the identification procedures."

that the environment constrained the categorizations, in that human knowledge could not provide correlational structure where there was none at all.

However, it is a rather strong metaphysical claim to argue that there exist objective categories in the world. We must remember that all human categorization depends (at least partially) on human physiology: observations on the perceptual level are furnished by the sensory projections of objects, whereas observations on the linguistic level are furnished by symbolic statements about objects that in turn are furnished by sensory projections of these objects. A more sensible and somewhat less strong, rather epistemological, claim is that categories "through human perception" stand out as natural categories.

The natural kind categories seem to depend on a notion of similarity, where similarity is a relation between two objects. Similar objects are grouped together to form a natural kind. This state of affairs forces us to analyze the concept of similarity and how it can be measured. The theoretical treatment of similarity has been dominated by two kinds of models: *geometric* and *set-theoretical*. [19]

In geometric models (see, for instance, (Shepard, 1974)) objects are represented as points in some coordinate space. The similarity between two objects is then measured (defined) by the metric distance between them. However, pure geometric models are inadequate for several reasons, for instance:

- The measure is only meaningful if the selected attributes are relevant for describing perceived object similarity. (Michalski and Stepp, 1983)
- All selected attributes are given equal weight (Michalski and Stepp, 1983) (or an arbitrary weight, as I would like to put it).
- It is more appropriate to represent some features as qualitative. (Tversky, 1977)

The geometric models seems related, in some way or another, to Gärdenfors' conceptual spaces. I will try to make this relation explicit. It is possible to see the subconceptual level as a (low-level) feature space of a high dimensionality. Thus, it can be said to correspond to a "pure" geometric model. A conceptual space can then be seen as the resulting space when the two first problems have been taken care of, corresponding to a "refined" geometric model.

In set-theoretical models objects are represented as collections of features. The most well-known set-theoretical model is Tversky's (Tversky, 1977) contrast model. It expresses the similarity between two objects as a linear combination of the measures of their common and distinctive features. However, set-theoretical models (Tversky's at least) have, more or less, the same problems as geometric models. They do not specify how relevant attributes are selected. The attributes are weighted, but how this is done is only loosely specified. That the feature must be weighted is implied (I think) by the theorem of the ugly duckling (Watanabe, 1969).[20] Moreover, any two objects can be arbitrarily similar or dissimilar by changing the weights. Finally, it is probably true that it is more appropriate to represent some features as quantitative.

As we have seen there are problems with "pure" similarity models, especially with the selection of relevant features. Schank and his colleagues (Schank et al., 1986) go one step further by stating that a "simple" theory for specifying the relevant features is impossible. Mainly because the relevance of features depends on the goals of the agent having the concept. They conclude (p. 640): "The process of determining which aspects of instances to be generalized are relevant must be based on an *explanation* of why certain features of a category took on the values they did, as opposed to other values that might a priori have been considered possible."

Thus, it seems that not all categories that humans normally use arise in the purely *bottom-up* fashion (Holyoak and Nisbett, 1988) described above. This suggests that even the weak claim that categories "through human perception" stand out as natural categories may be too strong, not covering all natural categories. For instance, Rosch herself (Rosch, 1978) argues (taking back her earlier claim) that some types of attributes present problem for these claims. For instance, there exist attributes that appear to have names not meaningful prior to knowledge of the category (e.g. seat - chair). Moreover, there exist functional attributes that seemed to require knowledge of humans, their activities, and the real world to be understood (e.g. "you eat on it" - table). From these examples she concludes: "That is, it appeared that the analysis of objects into attributes was a rather sophisticated activity that our subjects (and indeed a system of cultural knowledge) might well be considered to be able to impose only *after* the development of the category system." Moreover, she states that attributes are defined so that the categories once given, would appear maximally distinct from one another.

Similarly, Murphy and Medin (Murphy and Medin, 1985) have claimed that people's intuitive theories about the world guide the representational process. They made the demand upon categories that they must exhibit something called *conceptual coherence*. A coherent category is one "whose members seem to hang together, a grouping of objects that makes sense to the perceiver."

Thus, the problem with a purely "syntactical" model of similarity is that it ignores both the perceptual and the theory-related constraints that exist for, at least, a

[19] It seems that geometric models tend to treat all features as quantitative, whereas set-theoretic models tend to treat features as qualitative.

[20] This theorem, that is formally proved, shows that whenever objects are described logically, no two objects can be inherently more similar than any other pair. In other words, for similarity to be meaningful, the predicates describing an object must be censored or weighted.

certain kind of categories.[21]

3.1.3 Derived Categories

As pointed out earlier natural kind categories arise in a bottom-up fashion. In contrast, *top-down* category formation is triggered by the goals of the learner.

The categories formed in a top-down manner are often characterized in terms of functional features, whereas bottom-up categories are characterized in terms of their structure.

As Corter (Corter, 1986) points out, the two types of categories seem to be characterized by different kinds of features and feature relationships. Bottom-up categories tend to group instances that share co-occurring properties (they are "similar"), whereas top-down categories often consist of disjunctive groupings of different types of objects that may not share many properties (they do not have to be "similar"). For instance, the category "things-in-my-apartment" may include such things as records, books, chairs, apples, and so forth.

Barsalou (Barsalou, 1986) suggests that many of the top-down categories, which he calls ad-hoc categories, do not have the same static nature as bottom-up categories. While bottom-up categories generally are believed to be represented by relatively permanent representations in long-term memory,[22] he states that "many ad-hoc categories may only be temporary constructs in working memory created once to support decision making related to current goal-directed behaviour." As an example of a ad-hoc category he takes "activities to do in Mexico with one's grandmother". Other top-down categories, like "food", are relatively permanent though.

3.1.4 Artifact Categories

Not all natural categories are natural kinds. A natural division can be made between *species* (natural kinds) and *artifacts*. Rosch's examples above, "chair" and "table", (which certainly are natural categories) are typical artifacts. Characteristics for artifacts are that they are made by humans to have a certain function, implying that they should be characterized in terms of their functional features. However, it seems that the instances of most artifact categories also have structural, and thus perceptual, similarities. Moreover, some objects made for one purpose may be used for another purpose, it is possible for instance to use a chair as a table. Thus, we can say that artifact categories differ from natural kinds in that they seem to arise both in a bottom-up and a top-down fashion.

3.1.5 Taxonomies

Categories (natural categories at least) are also hierarchically organized in a different sense than "the property hierarchies" described above, namely, in taxonomies.[23] A part of a taxonomy is illustrated in Figure 3.

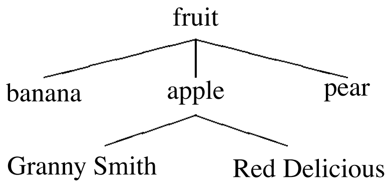

Figure 3: Part of a taxonomy of fruits

Taxonomies also serve an important function by promoting cognitive economy. How this is possible is demonstrated by Figure 4 and Figure 5. In Figure 4 we have a part of the fruit-taxonomy augmented with some features of the categories.

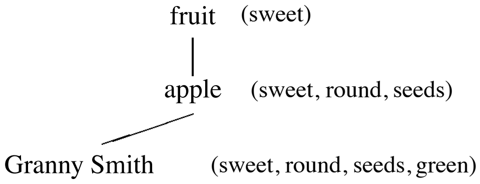
fruit (sweet)
apple (sweet, round, seeds)
Granny Smith (sweet, round, seeds, green)

Figure 4: Part of a part of a taxonomy of fruits augmented with features.

By noticing that categories on one level inherit the features from the (parent) category on the level above we can reduce the amount of information that we must store on each level. This is illustrated in Figure 5.

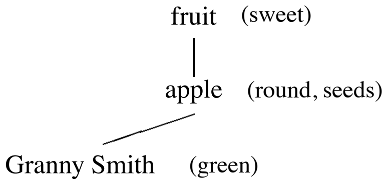

Figure 5: Part of a part of a taxonomy of fruits augmented with features (optimized).

Rosch et al. (Rosch et al., 1976) argue that there exists in these taxonomies a "basic level". They write: "Basic categories are those which carry the most information, possess the highest cue validity[24] and are thus, the most differented from one another ... Basic-level categories

[21] However, a perceptual system must have some built-in constraints that determine what will count as an attribute and the salience (weight) an attribute will have. Thus, in an autonomous system the perceptual constraints are determined by its perceptors.

[22] The representations can, of course, be modified but they are permanent in the sense that there always exists a representation of the category.

[23] This is a rather strong idealization, since some categories do not belong to any taxonomy at all while others belong to several.

[24] The cue validity of a feature F with respect to a category C is the validity with which F is a predictor of this category. The cue validity of an entire category may be defined as the summation of the cue validities for that category of each of the attributes of the category.

possess the greatest bundle of features ... Basic objects are the most inclusive categories which delineate the correlational structure of the environment." In our taxonomy of fruits (Figure 3) bananas, apples and pears constitute the basic level.

The basic level has some interesting properties that have consequences for both the epistemological and the inferential function. Since the basic level is the one we prefer (is the easiest) for categorization, the epistemological function is, in some sense, maximized at this level. Also the inferential function is maximized at the basic level. The basic categories have "the greatest bundle of features" (perceptual and non-perceptual) and many of the features are distinctive, permitting us to infer a substantial number of properties without much perceptual effort. In contrast superordinate categories (fruit) have relatively few properties and hence cannot enable us to make that many inferences. Although subordinate categories (Granny Smith) have many properties they have so few distinctive properties that they are more difficult to categorize perceptually. Finally, we should note that which level that actually is the basic level is context dependent, in the sense that it is the most appropriate in most situations but not all.

3.2 The Nature of AI Categories

In traditional AI, categories are often presumed to be artificial in the sense that all aspects of the category can be summarized by a short and simple definition in terms of necessary and sufficient conditions. This, of course, makes things a lot easier than they really are.

Most work in AI is concerned with bottom-up concept formation (similarity-based learning, SBL), although exceptions exist, for example explanation-based learning (EBL) (DeJong and Mooney, 1986; Mitchell et al., 1986).[25] For SBL both geometric and set-theoretic models have been used. Geometric models are often used by *conceptual clustering* systems, such as CLUSTER/2 (Michalski and Stepp, 1983; Stepp and Michalski, 1986), whereas *supervised* learning systems, such as version spaces (Mitchell, 1977), often use set-theoretic models.

Michalski and Stepp (Michalski and Stepp, 1983) propose an approach for measuring similarity in geometric models that, besides the objects (seen as points), takes into account other objects and "the set of concepts which are available for describing" the objects "together". They call this measure *conceptual cohesiveness*.

In AI the problem of selecting relevant features is often solved by letting the user select them. The choice of features is called the *bias*[26] of the learning system. However, sometimes the learning system has to select among the user-selected features. Several, more or less statistical, approaches for the selection of relevant attributes have been proposed, for instance, multidimensional scaling (Kruskal and Wish, 1978) and neural networks (Gärdenfors, 1992).[27] Another approach could be to use the algorithms presented in (Matthews and Hearne, 1991). A related task is what is sometimes labelled *constructive induction*. In (Subramanian, 1989) this task is described as: "finding a compact and simple concept representation given a labeled description of the instances of the concept". Moreover, "The chief problem is the generation of high-level features from the lower level features that characterize the instances." (Rendell, 1989) provides an overview of this problem.

Experiments have been conducted that use explanations to select relevant attributes when doing top-down concept formation (EBL). However, the success has been limited, probably due to the difficulties in specifying the appropriate background knowledge.

Taxonomies and their properties are rather well studied both in AI and in computer science in general. Take object-oriented languages, such as Smalltalk and Simula, where the classes are members of taxonomies and where features are inherited from super-classes, for instance. The topic of taxonomies in AI and computer science is further elaborated in (Jansson, 1987). However, among the existing concept learning systems it is only the conceptual clustering systems (Fisher and Langley, 1985) that actually construct taxonomies. Some of these systems, for instance (Hanson and Bauer, 1989; Fisher, 1988), also try to include basic-level aspects.

3.3 Conclusions

In contrast to traditional AI where artificial categories are often used, an autonomous agent in a real-world environment has to deal with the same kind of categories (natural and derived) as humans do.

As we have seen, objects have properties of different types. Some properties, common to all objects of the category,[28] are characteristic or discriminant, these can be used for metaphysical and epistemological classification (e.g. for the category "human", the genetic code and "walking upright" respectively). Some properties are common to all objects in the category although not characteristic or discriminant, these can be used to make inferences (e.g. having a kidney). The more or less useless properties that are not common to all objects in the category (sometimes called irrelevant properties) (e.g. hair colour) are then left over. Moreover, it seems that

[25] However, categories are not *formed* in EBL. The categories are formed beforehand and a high-level description of them is given as input to the learner. The task is only to transform the (abstract) high-level characterization into a low-level (often perceptual) characterization.

[26] Actually, this is one of several types of bias. Other types are, for instance, (Utgoff, 1986), the space of hypotheses that the system can consider, the order in that hypotheses are to be considered, and the criteria for deciding when a hypothesis is good enough.

[27] Multidimensional scaling and neural networks are used more to reduce the number of attributes, than to actually find the relevant attributes. However, these tasks seem closely related.

[28] Do not take "all" too literately. It may be the case that universal regularities do not exist, implying that reasoning about categories must be probabilistic in nature.

some features are represented more naturally as qualitative and some as quantitative. Thus, it would be nice if it were possible for the agent to have both types.

Somewhat carelessly one might say that bottom-up categories arise due to curiosity, whereas top-down categories arise due to problem solving activities. Information about bottom-up categories is to a great extent derived from perceptual observations of the environment, whereas information about top-down comes from more abstract observations. Thus, a passive agent (that perhaps just tries to make a description of its environment) could manage with only bottom-up categories, whereas a problem solving agent also needs top-down categories.

As we have seen above, bottom-up category formation has been rather well studied in AI. However, some problems remain to be solved. Top-down category *formation* on the other hand is hardly studied at all. Unfortunately, we do not get much help from the psychologists either. They have pointed out that there are categories that are formed in a top-down manner, but they do not give us a hint as to how the formation takes place.

There is a problem with artifact categories in that they seem to be both bottom-up and top-down categories, where the top-down-ness, and the problem, is due to the emphasis on the function of the artifact objects. The recognition of possible functions of an object from perceptual observations seems like a very hard problem.[29] However, one approach to this problem is presented in (Vaina and Jaulent, 1991). On the other hand, it would require a very large knowledge base to be able to form artifact categories in a top-down manner, and we still do not know how this should be done.

The simplest solution may be to form artifact categories in a bottom-up fashion, making the assumption that perceptual similarity is enough. Thus, having bottom-up categories as the only permanent categories, and maybe constructing temporal top-down (ad-hoc) categories when convenient in problem solving tasks (where it seems more likely that the right background knowledge is available).

Finally, we need to structure the categories into taxonomies to promote cognitive economy and inferential functions. However, since this is a rather well studied topic, it is probably wise to concentrate on other problems.

4 Representation of Concepts

The classes of objects which we call categories, whose members exist externally, must of course be internally represented in some way. In this section I will discuss the following questions; How do humans represent concepts? How do present AI systems represent concepts? How should AI systems represent concepts?

4.1 Human Representation of Concepts

In (Medin and Smith, 1984) three views of concepts are presented: the *classical*, the *probabilistic* and the *exemplar*. These views are to a great extent theories about representation of concepts.

4.1.1 The Classical View and It's Problems

According to the classical view all instances of a concept share common features that are singly necessary and jointly sufficient for defining the category. Thus, it would be possible to represent a concept by these features. Categorization would then be a matter of straightforward application of this "definition".

However, there are some problems with this view (according to, for instance (Smith and Medin, 1981) and (Smith, 1988)):

- Natural categories are, in contrast to artificial categories, often not representable by necessary and sufficient features.
- Even if a category can be defined as above we tend to not use this definition.
- There are unclear cases of category membership.
- It is generally believed that some exemplars of a category are more typical than others.
- We often think more concretely than the situation demands.

The fact that some categories do not have a classical definition is sometimes called the ontological problem (Amsterdam, 1988). A nice and famous example, mentioned by Wittgenstein, is the category "game". As we shall see below, this issue is related to the metaphysical function of concepts.

Assuming that a classical definition exists for a category, it is interesting to notice that instead of using the classical definition we often use non-necessary features to characterize a category or to categorize objects of the category. As we shall see below, this issue is closely related to the epistemological function of concepts.

An example of unclear category membership is that it is hard to decide for some objects whether they are a bowl or a cup. In this example there is a relation between an object and a category but the same problem arises between levels in a taxonomy (subcategory-category relations). For instance, is tomato a fruit or a vegetable, or is a rug a piece of furniture?

[29] Mainly because, "[one] must have some knowledge that is capable of mediating between the features at the two levels; that is, to determine whether an abstract feature is perceptually *instantiated* in an object, one must have recourse to ancillary knowledge about the relation between abstract and perceptual features." (Smith and Medin, 1981) (p.19). Moreover, the functions must, of course, be known in advance, preprogrammed or learned (which seems to be an even harder problem). It goes without saying that by letting the agent have access to observations on the linguistic level, where the function is explicitly given, the problems with functional properties disappear. However, assuming that the agent has access to such observations seems too generous for most applications.

Prototype usually refers to the best representative(s) or most typical instance(s) of a category as opposed to the treatment of categories as equivalence classes.[30] For instance, it has been shown that (at least for the experiment subjects) robins are prototypical birds whereas penguins are not.

It seems that we often think about specific objects when we actually refer to a category. For instance, if someone says that he had to see a dentist, we often think of a specific dentist.

It has been suggested that instead of the strong demand that category shall have a classical definition, the instances of a category should only have to have a sufficient amount of *family resemblance* (Wittgenstein, 1953; Rosch and Mervis, 1975). Family resemblance usually refers to the number of features that are shared by members of a category. It can be viewed as a measure of typicality. Typical members of a category share many attributes with other members of the category (and few with members of other categories).

Thus, it seems clear that the classical view cannot explain all aspects of human concepts. In response to this, the probabilistic and the exemplar view have been presented as views being more realistic and consistent with empirical findings.

4.1.2 The Probabilistic View

According to the probabilistic view, concepts are represented by a summary representation in terms of features that may be only probable (or characteristic) of category members. Membership in a category is graded rather than all-or-none. Better members have more characteristic properties than the poorer ones. An object will then be categorized as an instance of some category if, for example, it possesses some critical number of properties, or sum of weighted properties, included in the summary representation of that category.

Thus, rather than applying a definition, categorization is a matter of assessing similarity.

4.1.3 The Exemplar View

Those in favor of the exemplar view argue that categories may be represented by (some of) their individual exemplars, and that concepts thus are represented by representations of these exemplars. A new instance is categorized as a member of a category if it is sufficiently similar to one or more of the category's known exemplars. Thus, also in this case categorization is a matter of assessing similarity rather than applying a definition.

There are several models consistent with the exemplar view. One such model is the *proximity* model which simply stores all instances. An instance is categorized as a member of the category which contains its most similar stored exemplar. Another model is the *best examples* model. It only stores selected, typical instances. This model assumes that a prototype exists for each category and that it is represented as a subset of the exemplars of the category. Another possible alternative is that the prototype is a non-existing "average" instance that is derived from the known instances.

4.1.4 Combining the Probabilistic and Exemplar View

Another possibility is that the representation of a concept contains both a probabilistic summary representation and exemplars (Smith and Medin, 1981). It seems reasonable that when the first instances of a category are encountered we represent it in terms of these instances. And when further instances are encountered we apply abstraction processes to them to yield a summary representation.

It seems that this approach has some interesting features that relates to *non-monotonic reasoning* (Ginsberg, 1987). Consider a point in time where a person has both a summary and a exemplar representation of the concept "bird", where the summary representation contains the feature "flies" (as very probable). How should the representation be updated when the person is confronted with a penguin? It would not be wise to alter the old summary representation too much because the fact that a random bird flies is very probable. A better solution is probably to store the penguin as an exemplar as can be done in a combined representation. However, there are many details to work out before we have a complete theory about a such combined representation.[31]

4.1.5 Comments

We must remember that the existence of prototypes does not have any clear implications for the construction of models of human concept representation, processing and learning. Thus, prototypes do not specify such models, only impose constraints on them. Actually, I think that prototypes are only a problem for the classical view if it states that categories are equivalence classes. Clearly, there exists categories that have a classical definition but still have prototypes. For instance, some triangles are more typical than others.

Another reflection is that the classical view seems to try to capture the intension of concepts whereas the exemplar view (at least partially) describes the extension.

[30] "Prototype" is ambiguous though, it has also been used to refer to a description of a category that is more appropriate to some members than it is to others.

[31] The possible connection between prototype-based representations and non-monotonic reasoning has been pointed out in (Gärdenfors, 1990). It is suggested that concepts at the conceptual level are represented as convex regions in a conceptual space. When an individual is first known as being a bird, it is believed to be a prototypical bird, located in the center of the region representing birds. In this part of the region birds do fly. If it then is learned that the individual is a penguin, the earlier location must be revised so that the individual will be located in the outskirts of the "bird-region", where most birds do not fly. However, my reflection concerns the acquisition of the representation, whereas in Gärdenfors' case the representation is already learned. Moreover, the combined approach is on the linguistic level and not restricted to convex regions.

4.2 Representation of Concepts in AI

Traditionally in AI, categories are treated as equivalence classes that can be described by necessary and sufficient conditions. Thus, AI has adopted a rather strong version of the classical view. Some of the representation languages that have been used are: *logic-based notation* (Michalski, 1980) , *decision trees* (Quinlan, 1986) and *semantic nets* (Winston, 1975).

4.2.1 Non-traditional Representation

Within the last few years, some experiments with non-classical representations have been done. Some researchers are inspired by the exemplar view and some by the probabilistic view.

Let us begin with those who are influenced by the exemplar view. Kibler and Aha (Kibler and Aha, 1987) have experimented with both the proximity model where all instances are stored and selected examples model where a subset of the instances are stored. Systems that use this kind of representation often use some version of the nearest neighbor algorithm to classify unknown instances. That is, a novel instance is classified according to its most similar known instance. Musgrove and Phelps (Musgrove and Phelps, 1990) have chosen to have a singular representation of the average member (not necessarily an instance) of the category, which they call the prototype. Nagel (Nagel, 1987) presents a best examples model that, in addition to the prototype(s), stores transformations that transforms less typical instances to a prototype. Learning systems which use specific instances rather than abstractions to represent concepts have by Aha and his colleagues (Aha et al., 1991) been labeled *instance-based*. They also provide a theoretical analysis of such algorithms.

Followers of the probabilistic view are, for instance, de la Maza (de la Maza, 1991) who calls his type of representation *augmented prototypes*. Fisher's (Fisher, 1988) *probabilistic concept tree* represents a taxonomy of probabilistic concepts.

An important reflection is that, at least, the exemplar view seems to demand some kind of similarity measure. But, as we have seen, similarity is not an unproblematic topic. Moreover, the probabilistic representations seem to have trouble with atypical instances. Therefore, it would be interesting to experiment with implementations of a combination of the probabilistic view and the exemplar view, which seem to handle such instances quite well. Moreover, since a combination does not have to store as many instances as an exemplar representation, it requires less memory and it probably categorizes faster, since fewer comparisons between instances are needed.

A quite different approach to non-traditional concept representation is taken by Michalski and his colleagues (Michalski, 1987b; Bergadano et al., 1992). Their representation has two components, the *base concept representation* (BCR) and the *inferential concept interpretation* (ICI). The BCR is a classical representation that is supposed to capture typical and relevant aspects of the category, whereas the ICI should handle exceptional or borderline cases. When categorizing an unknown object, the object is first matched against the BCR. Then, depending on the outcome, the ICI either extends or specializes the base concept representation to see if the object really belongs to the category. This approach is similar to Nagel's, but she uses a prototype as the BCR, not a classical definition.

4.2.2 Subsymbolic Representation

Recall Gärdenfors' three levels of observation from the first section. In the same way that observations can be described on different levels, it is possible to represent concepts on different levels. The methods of representation described above are all on the linguistic (symbolic) level. A method of representing (and acquiring) concepts on a lower level is to use neural networks. These were initially meant to be cognitive models of the brain at the level of neurons.

Pylyshyn (Pylyshyn, 1984) has distinguished three levels of cognitive modeling. The lowest level is concerned with the physiological mechanisms underlying thought. The highest level is concerned with the content of thought, the aspects of the world that are encoded in the mind. Between these levels are the mechanics of how a representation is formed without regard to the content of the representation. Newell (Newell, 1990) refers to this level as the symbol manipulation level. Thus, neural networks belong to the lowest of these levels.

In the last years there has been a growing optimism about the capability of neural networks, both as cognitive models (e.g. the works of Grossberg (Carpenter and Grossberg, 1986) and of Edelman (Edelman, 1989)) and as tools for pattern recognition (e.g. backpropagation networks (Rumelhart et al., 1986)). However, one must keep in mind that neural networks that can be simulated on a computer, as most neural networks can, are of course at the most Turing-machine-equivalent. They might be better suited (more efficient or easier to program) than symbolic algorithms (computers) for some problems, but are not a more powerful tool in general.

Clearly, neural networks are adequate for cognitive modeling of the physiological mechanisms underlying thought, but since they do not represent knowledge explicitly, which seems crucial for the implementation of the metaphysical and inferential functions, they do not seem suitable for such purposes. The functions that the subsymbolic methods will be able to handle seem, for the moment, limited to tasks like perceptual categorization.[32] Thus, there is a possibility that they might be able to implement the epistemological function. But for at least three reasons I will not discuss

[32] Even though the opposite opinion is sometimes held, see for instance (Balkenius and Gärdenfors, 1991).

them further in this paper. First, it is difficult to introduce background (a priori) knowledge.[33] Second, it is difficult to exploit and reason about both the learned knowledge and the learning process. Third, they learn too slowly in the sense that they need many instances to learn a fairly good representation and since the weights often are randomly chosen, the behaviour of the net is unpredictable in early stages and cannot be used for classification.[34]

For a more detailed discussion about possibilities and limitations of connectionist models in general, see (Smolensky, 1988).

4.3 Conclusions

So, how should autonomous agents represent concepts? Let us analyze this question in terms of the functions that the concepts should be able to serve.

Some categories can be characterized by a classical definition (necessary and sufficient conditions). However, such a definition is often based on features that under normal circumstances are non-perceptual, such as atomic structure, genetic code or functionality. Thus, these definitions are not adequate for perceptual classification,[35] and consequently not appropriate representations for supporting the epistemological function. Instead, the implementation of the epistemological function seems to demand some kind of prototypical (or maybe subsymbolic) representation.

The implementation of the metaphysical function, on the other hand, demands by definition a classical definition. However, it seems almost impossible for a non-communicating autonomous agent to learn such a definition, since it is limited to perceptual observations. To find the metaphysical definition can be seen as a very sophisticated version of classical *induction* (as studied in philosophy), since we do not only have to induce one rule but possibly several. Moreover, we need to know that these rules are necessary and sufficient.

To implement the inferential function we must have some "encyclopedic" knowledge about the category and its members. This knowledge can probably be seen as a collection of universal or probabilistic rules. Kirsh (Kirsh, 1986) has called this collection "a package of associated glop". The acquisition of this knowledge seems like a hard learning problem, involving classical induction.

Traditional work on concept representation in AI has assumed that a single and simple structure (such as a logic-based description, a decision tree, or an instance-based description) could capture all the relevant aspects of a concept. However, the above discussion makes clear that this is not possible except for in very restricted domains. We need a richer composite representation that is structured according to the functions of the represented concept.

Supported by the research reviewed in this paper I propose the structure illustrated in Figure 6 as a reasonable representation of concepts by autonomous agents. The dashed boxes in the figure indicate optional fields.

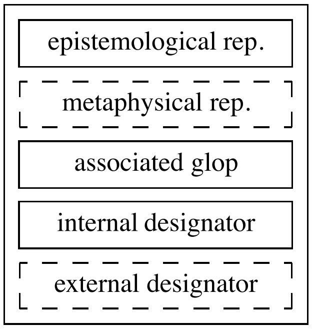

Figure 6: Composite Concept Representation

All parts of the representation are not always necessary or even adequate. Metaphysical representation only exists for some concepts and might, moreover, be irrelevant for an autonomous agent, and external designators are only necessary for communicating agents.

Let us illustrate the idea of composite representation by using the category "wasp". The kind of information that the epistemological representation may include is that wasps are black and yellow striped, cylinder-shaped, approximately two centimeters long and half a centimeter in diameter, that they hum, have two wings, and so on. As concluded above this information is best represented by some kind of prototypical representation, probabilistic or instance-based or a combination of these, or possibly a neural net. The metaphysical representation may include information about the genetic code of wasps expressed in a logic-based notation or maybe by a decision tree. The kind of encyclopedic knowledge that the associated glop would include is for instance: can hurt other animates with its sting, lives in collectives, and so on. This kind of information is probably best expressed in a logic-based notation. Internal designator: organism.animate.xxx[36] External designator: wasp.

There is of course no sharp distinction between what information is included in these representations. Thus, there may be redundant information. For example, in addition to being an essential part of the epistemological representation, the fact that wasps have wings is a quite natural part of the encyclopedic knowledge represented

[33] There have been experiments with introducing symbolic knowledge into "knowledge-based" neural networks, see for instance (Towell et al., 1990). However, as I see it these networks are rather symbolic than subsymbolic representations since every node represents something. This implies moreover that the knowledge in these kinds of nets are not distributed, which is one of the characteristic features of neural networks.

[34] Moreover, I simply like the symbolic level more.

[35] However, in traditional AI it is very common to try to make a classical definition of a category based directly on the perceptual data.

[36] The choice of the internal designator is entirely up to the system, it should be as convenient and effective as possible for the system.

in the associated glop. However, the fact is probably not represented in the same way in these representations. It may be rather implicit in a prototype for the epistemological representation and more explicit in a logic-based notion for the associated glop.

This composite structure enables concepts to serve all the functions listed before. The epistemological and metaphysical representations support the epistemological and metaphysical functions respectively. The associated glop supports the inferential function. The internal designator supports the intrapersonal stability, whereas the external designator supports both the interpersonal stability and the linguistic function.

Depending on the situation, the composite concept representation is *accessed* (or retrieved) in different ways. External "stimuli" in the form of direct perception of objects access the concept via the epistemological representation. If, on the other hand, the external stimuli is on the linguistic level, as when communicating with other agents, the concept is accessed via the external designator. Finally, if the stimulus is internal, like in the case of reasoning, the concept is accessed via the internal designator.

In this example taxonomical knowledge is expected to be stored outside the actual concept-structure. Another possibility is to store such knowledge inside the concept-structure.

As we have seen, the AI community has already studied all of the well developed psychological models of concept representation. The only psychological model that has not been implemented yet (at least as far as I know) is the combined exemplar and probabilistic model. Even though it has not been studied in any depth in cognitive psychology either, it might be a candidate, at least from the AI point of view, for the epistemological representation of concepts.

In the following I will concentrate on the learning of the epistemological representation and to some extent the metaphysical representation. The learning of the associated glop, on the other hand, has more in common with traditional induction than concept acquisition, and is therefore not discussed in the rest of this paper.

5 Concept Acquisition

Finally, we have reached the stage were we are able to discuss how concepts, these internal representations of external categories, can be acquired.

5.1 Human Concept Acquisition

According to Atkinson et al. (Atkinson et al., 1987), humans learn about categories in two different ways:

- by being explicitly taught
- by learning through experience

Unfortunately, the authors do not further elaborate this distinction, and I have not found any other discussions concerning this topic, so I will try to elaborate it myself. As I take it, it is possible to be explicitly taught about categories on both the linguistic level (learning by description) and on a sublinguistic (perceptual) level (learning by acquaintance). Examples of learning on the linguistic level are when you learn something reading a book or being told something by some kind of teacher. It seems likely that we learn the metaphysical aspects of concepts in this way. As an example of being explicitly taught on the perceptual level we have the situation when a teacher shows an exemplar of a category (ostensive definition).[37]

When you learn from experience, there is no teacher available to help you with the classification. For instance, if you are confronted with an instance of a category you know rather well, but this instance is different in some aspect from those you have experienced, you might nevertheless "guess" what category it belongs to and, thus, learn something about the category. Another situation is when you are confronted with an instance of a category you know nothing about. You may then form a new category based on that instance. Thus, there are two cases of learning from experience, it can either be learning something about a known category or about an unknown category. It is important to notice that the input when learning through experience is often on the perceptual level.

There is yet another way of learning about categories that is, in a way, orthogonal to the others, namely, learning by experimentation. It could be performed by actually making experiments or, maybe more common, by asking questions. Asking questions belongs to the linguistic level whereas learning by actual experiments seems to belong to the perceptual level. This type of learning bears resemblance to scientific discovery.

It is important to remember that in real life we do not acquire a concept in just one of these ways. Instead, we use them all interchangeably. Which kind of learning that is appropriate in a particular situation is, of course, to a great extent determined by the environment.

There are several other restrictions that the environment imposes on the concept acquisition process. For instance, it must be *incremental*, since we do not encounter all instances of a category at one point in time. Instead, we encounter an instance now and then, incorporating it into our "bulk of knowledge of concepts". Thus, concepts are acquired in a gradual fashion, by interacting with the environment over time. Moreover, we do not learn one concept at a time, concepts are rather acquired in *parallel*.[38]

[37] The explicitness in the last example is weaker than in the examples of linguistic level learning. Thus, it would be more appropriate to place this type of learning between the two categories above.

[38] Here we refer to the normal, rather *passive*, concept acquisition process. However, in some situations we adopt a more *active* strategy, where we concentrate on one concept at the time.

As Schank et al. (Schank et al., 1986) point out, any dynamic and autonomous theory of concept acquisition must specify at least three processes:

1. Deciding when to create a new concept.
2. Deciding when to modify a concept.
3. Deciding what part of the concept to change.

5.1.1 Theories of Concept Acquisition

As pointed out in the introduction of this paper, all our knowledge about categories cannot be innate. However, it is possible, and even plausible, that some knowledge about categories is innate. Different researchers emphasize this to different degrees. Fodor's theories (Fodor, 1975), for instance, rely heavily on innate knowledge.

If all concepts are not innate then some of them must be acquired in some way. How this is done has, of course, been the subject of research in cognitive psychology. The three most predominant psychological theories of human concept acquisition are:

- The association theory
- The hypothesis testing theory
- The exemplar strategy

The *association* theory as described in (Solso, 1991) seems rather outdated, with its roots in stimulus-response psychology. It holds that the learning of a concept is a result of (1) reinforcing the correct pairing of a stimulus with the response of identifying it as a concept, and (2) non-reinforcing (punishment) the incorrect pairing of a stimulus with a response of identifying it as a concept. This theory seems only to cover the case of being explicitly taught something about the category on the perceptual level. Moreover, it is extremely vague and thus consistent with most theories. However, it is interesting to notice its resemblance with the backpropagation algorithm for teaching neural nets.

The theory of *hypothesis testing* states that "we hypothesize what properties are critical for determining whether an item belongs to a category, analyze any potential instance for these critical properties, and then maintain our hypothesis if it leads to correct decisions." (Atkinson et al., 1987) Thus, it assumes that the category can be characterized by a classical definition, and it seems to assume that all instances of the category are concurrently available for analysis. These assumptions are too strong for most learning situations. The theory does not specify when to create a new concept. Moreover, it is non-incremental and only learns one concept at the time. In my opinion, the hypothesis testing theory is a sort of model of some kind of learning by experimentation, such as when a scientist is doing experiments.

Finally, the *exemplar strategy* simply states that when encountering a known instance of a category a representation of it is stored. This theory is consistent with the exemplar view of representation and thus inherits its limitations (for instance, covers only the epistemological aspects). However, several questions remains open, for example: How many, and which, instances should be memorized? Moreover, the strategy is only specified for learning from preclassified instances. However, it seems possible to extend the theory to include learning from experience, but then, *when* to create a new concept must be specified. Advantages with the exemplar strategy are, that it is incremental in nature and that it accounts for the acquisition of many concepts at the time.

5.2 AI Methods for Concept Acquisition

The concept acquisition process of an autonomous agent is of course restricted by the environment in the same way as a human is. Thus, from the earlier discussion we can conclude that for artificial autonomous agents the concept acquisition must also be incremental, concepts must be acquired in parallel, and several methods must be employed simultaneously.

5.2.1 AI Methods for Concept Acquisition Using Traditional Representation

In AI, several ways of learning about categories have been studied. The most studied are:

- Direct implanting of knowledge
- Learning from examples
- Learning by observation
- Learning by discovery
- Learning by deduction

Direct implanting of knowledge is the extreme, almost trivial, case of concept acquisition in which the learner does not perform any inference at all on the information provided. It includes learning by direct memorization of given concept descriptions and the case when the descriptions are programmed into the computer. The latter can, from the perspective of an autonomous agent, be seen as a way of incorporating innate, or a priori, knowledge about concepts into the agent. In *learning by instruction* (learning by being told), which is rather similar to direct implanting of knowledge, the learner acquires concepts (explicitly described on the linguistic level) from a teacher, database, textbook or some other organized source. This form of learning, in contrast to direct implanting of knowledge, requires selecting the relevant information and/or transformation of this information to an usable form.

Learning from examples is by far the most studied type of learning in AI and can be seen as learning by being explicitly taught. In this kind of learning the learner induces a concept description from preclassified examples (and, in most cases, counterexamples) of the category that are provided by some kind of teacher. Since there is a teacher present to guide the learning process, this type of learning is sometimes called *supervised learning*. Thus, it is the teacher who decides when to create a new concept. The task for this type of learning can be seen as finding a definition (description) consistent with all positive examples but no negative examples, if there are any, in the training set. Most of the systems learning from examples can be viewed as carrying out a search through a space of possible concept descriptions. This space can be partially ordered, with the most general description at one end and the most specific at the other. The most general description has no features specified, corresponding to the set of all possible instances, whereas the most specific have all features specified, corresponding to instances. There are basically two strategies for searching the space of concept descriptions. In the general-to-specific strategy, one begins with the most general description as the hypothesis of the correct concept description, and as new instances are encountered, more specific descriptions (hypotheses) are produced. In the specific-to-general strategy, one begins with a very specific description, typically a description of the first instance, moving to more general descriptions as new instances are observed. Some systems use one or the other of these strategies, while more sophisticated systems, like Version Spaces (Mitchell, 1977), combine the two strategies. Since there is no inherent non-incrementality in this approach, it seems possible to make systems based on this approach that learn incrementally.[39]

A different kind of learning-from-examples systems are the so called top-down induction of decision trees (TDIDT) systems (Quinlan, 1986). These systems need both positive and negative instances of the category to be learned, with each instance represented as a list of attribute-value pairs. The output is a decision tree that can be used to decide if an instance is a member of the category or not. TDIDT systems begin with the root of the tree and create the decision tree in a top-down manner, one branch at a time. At each node they use an *information theoretic* evaluation function to determine the most discriminating attribute. The evaluation function is based on the number of positive and negative instances associated with the values of each attribute. An advantage of TDIDT systems is that they carry out very little search, relying on the evaluation function instead. However, a serious limitation of these systems is their non-incremental nature. To incorporate new instances, the tree has to often be recomputed from scratch.

The learning-from-examples systems typically learns just one concept at the time, without considering other known concept descriptions. An exception to this is AQ11 (Michalski and Larson, 1978; Michalski and Chilausky, 1980) by Michalski and his colleagues, which learns multiple concepts. Another exception is a system by Gross (Gross, 1988) which incrementally learns multiple concepts. The current concept description that is learned is constrained by the descriptions of the other concepts. However, this system can also be described as learning by experimentation, since it selects the next instance to be analyzed from a given description space itself. This instance is then classified by an oracle. The introduction of an oracle being able to classify every possible instance makes the learning easier and less realistic.

In *learning by observation* the learner forms categories itself, through direct interaction with the environment. Thus, it can be seen as learning through experience. Since it is the environment, not a teacher, that provides the examples, this type of learning is sometimes called *unsupervised learning*. Typically, the learner is given a number of entities (that are not preclassified) described by a number, n, of features. Based on their features it groups the entities into categories (aggregation). This is often done by treating the instances as points in a n-dimensional space and employing statistical methods (cluster analysis and numerical taxonomy), augmented with a preference criterion concerning the concept description language. Thus, it is the learner that decides when to create a new concept. When the aggregation is done, the system creates descriptions of the categories (characterization). This is done much in the same way as the systems that learn from examples. These types of systems are commonly called *conceptual clustering* systems. Some of the most well-known are CLUSTER/2 (Michalski and Stepp, 1983; Stepp and Michalski, 1986) and RUMMAGE (Fisher and Langley, 1985). Notice that all conceptual clustering systems form concepts in parallel. Moreover, they structure the created concepts into taxonomies, while other types of systems usually learn concepts at a single level. CLUSTER/2 and RUMMAGE learn in a non-incremental fashion, but incremental systems exist, like UNIMEM (Lebowitz, 1986).

All of these conceptual clustering systems use some kind of similarity measure, which depends on some kind of distance metric, for the aggregation task. As pointed out earlier, such a metric has several disadvantages, for instance, there exists no natural distance metric since it is dependent on the relative scaling of the axes of the space (which is arbitrary). Moreover, a distance metric can take into account totally irrelevant features. An interesting clustering (aggregation) technique that does not use a distance metric is presented in (Matthews and Hearne, 1991). The clusterings are instead optimized on the prediction of feature values, which the authors believe is the intended function of the clustering. Thus,

[39] However, some systems, version spaces for instance, have at some stages in the learning process several competing hypotheses. Having several hypotheses makes it difficult to use the concept and requires more memory space. However, the memory requirements are substantially less than for systems that must memorize all instances, such as Winston's (Winston, 1975).

this approach aims at maximizing the utility of the clustering.

Learning by discovery is also a type of unsupervised learning. However, systems that learn by discovery are more active in their search for new categories than systems learning by observation. They exploit their domain, sometimes by experiments, rather than passively observe it. The most famous system of this kind is Lenat's AM system (Lenat, 1976; Lenat, 1977). Another well-known system is GLAUBER (Langley et al., 1983). AM works in the domain of mathematics and searches for and develops new "interesting" categories after being given a set of heuristic rules and basic concepts. It uses a "generate-and-test" strategy to form hypotheses on the basis of a small number of examples and then tests the hypotheses on a larger set to see if they appear to hold. Surprisingly, the AM system works very well. From a few basic categories of set theory it discovered a good portion of standard number theory. However, outside this domain AM does not work very well. Two of the reasons are that there are difficulties in specifying heuristics for other less well-known domains, and that in the implementation of AM implicit knowledge about number theory was built-in. Moreover, even though AM initially performed well in the domain of number theory, its performance decreased after a while and it was not able to discover any new interesting categories. This was due to the static nature of the heuristics, which did not change when the system's knowledge about the domain increased, resulting in a static system. Thus, for such a system to be more dynamic, it must also be able to reason and manipulate with the heuristics.

In *deductive learning*, the learner acquires a concept description by deducing it from the knowledge given and/or already possessed (background knowledge). The most investigated kind of deductive learning is *explanation-based learning* (EBL) (Mitchell et al., 1986; DeJong and Mooney, 1986) which transforms a given abstract concept description (often based on non-perceptual features) to an operational description (often based on perceptual features) using a category example (described by operational (perceptual) features) and background knowledge for guidance.

The standard example of EBL is about the concept "cup". In this example the abstract concept description includes the facts that a cup is an open, stable and liftable vessel. Moreover, the background knowledge includes information such as: if something is light and has a handle then it is liftable, if something has a flat bottom then it is stable, and so on. Given this and an example of a cup in terms of more perceptual features (such as, light, has a handle) and the operationality criterion that the concept description must be expressed in terms of the perceptual features used in the example, the EBL-system produces a description of the concept "cup" that includes the facts that a cup is light, has a handle, has a flat bottom, and so on.

This form of learning is clearly a kind of top-down learning, since the learning is triggered by the goals of the learner. It can, as has pointed out earlier, be seen as just a reformulation of concept descriptions, since the abstract description is given. Thus, no new categories are created.

5.2.2 AI Methods for Concept Acquisition Using Non-traditional Representation

As mentioned in the previous section, there have been some experiments involving non-classical representations during the last years. However, these experiments have been limited to learning from examples and learning by observation.

Kibler and Aha (Kibler and Aha, 1987) describes three algorithms that learn from examples, using an exemplar representation of concepts. The *proximity* algorithm simply stores all training instances. The *growth* (additive) algorithm stores only those training instances that would not be correctly classified. These two algorithms are incremental in contrast to the third, the *shrink* (subtractive) algorithm. Instead, the shrink algorithm begins by placing all the training instances into the concept representation, and then continues by testing each instance in turn to see if it would be correctly classified by the remaining instances. In (Nagel, 1987) Nagel presents another system that learns incrementally from examples, using an exemplar representation. When a positive instance is presented to the system, the system will try to find a sequence of transformations that transforms the instance into a prototypical instance. The new transformations are then stored as a part of the concept description to be used for assimilating new instances.[40] De la Maza's PROTO-TO system (de la Maza, 1991) also learns incrementally from examples but uses a probabilistic representation. It groups the instances according to their categories and then builds a prototype (some kind of an average member) for each category. The prototypes are then augmented, weighting each attribute in order to form a probabilistic representation.

The PLANC system by Musgrove and Phelps (Musgrove and Phelps, 1990) learns from observation by a clustering algorithm that first applies multidimensional scaling to reduce the dimensionality of the input data. When the clusters are detected, their members are used to produce the prototype (a hypothetical average member). The system uses an exemplar representation and is non-incremental. A system that learns from observation incrementally is Fisher's COBWEB (Fisher, 1987; Fisher, 1988), which builds a probabilistic concept tree. As an evaluation measure of clusterings in the aggregation task, COBWEB uses *category utility* instead of a distance metric. It was originally developed by Gluck and Corter (Gluck and Corter, 1985) as a means of pre-

[40] How the prototypes are learned in the first place is not described in the material that, for the moment, is available to me.

dicting the basic level in human taxonomies. It is similar to Matthews and Hearne's approach in that it maximizes the predictive ability of the clustering.

Finally, we can notice that the typicality of the instances is not given explicitly in these systems. (I am not sure about Nagel's system, though.)

5.3 Conclusions

As we have seen, the issue of concept acquisition, in contrast to functional and representational issues, is more elaborated in AI than in Cognitive Psychology. In fact, for all the psychological models presented in this section, there exist corresponding AI methods. For instance, one can compare the association theory with backpropagation learning, the hypothesis testing theory with Gross' system, and the exemplar strategy with Kibler and Aha's experiments on instance-based learning. Thus, the theories about human concept acquisition have already been tested as AI methods, implying that there is not much we can gain by studying these psychological models. However, one result of the study is, as pointed out several times before, that approaches to concept acquisition by autonomous agents are constrained by several demands, they must:

- be incremental
- be able to learn many concepts at the time
- apply several methods simultaneously

To what extent have these demands been met by existing AI systems? In early machine learning research, learning systems were typically non-incremental. In recent years however, many incremental systems have been constructed.

There exist systems that learn many concepts at a time, for instance conceptual clustering systems and some non-traditional learning-from-example systems. However, in traditional learning-from-example systems, knowledge about known categories and taxonomies are typically not used to constrain the hypothesis space. How this should be done seems like an important area of research (if one is interested in the metaphysical functions of concepts).

Still, the requirement that several methods of learning must be applied simultaneously indicates where the greatest need for more research can be found. It may be true that there already exist systems that integrate two or more learning methods. However, most of these systems integrate learning from examples and explanation-based learning, see for instance (Lebowitz, 1990).

The types of learning that are adequate to integrate depends heavily on the environment in which the agent works. As mentioned earlier, there are two possible scenarios for an autonomous agent. It can either be alone in its environment or be among other agents which it can communicate with.

An agent that is alone can, of course, have preprogrammed knowledge about concepts (direct implanting of knowledge). Apart from this, it seems to be limited to learning by observation. The kind of knowledge it can learn in this way is mainly epistemological and to some degree inferential. But since the agent is restricted to perceptual information it can hardly learn any metaphysical knowledge.

In the case where other agents exist, it may be possible for the agent to learn from examples in addition to learning by observation and direct implanting of knowledge. This is done by letting some other agent act as a teacher. Thus, for this kind of agent, an integration of learning from examples and learning from observation may be fruitful.

One of the key problems for an algorithm that integrates learning from examples and learning by observation is to decide when to create a new concept. It needs to know when it encounters an instance of an unknown category. Somewhat surprisingly, this demand radically constrains the choice of representation. An (implicit) assumption that is often made when learning from examples, is that all categories in the universe are exemplified in the learning set. This assumption has led to the construction of algorithms that learn to *discriminate* between categories. By concentrating on the differences between the categories rather than the categories themselves, they just learn the boundaries between categories. Moreover, they partition the entire description space into regions, so that every region belongs to a certain category. Thus, when instances of unknown categories are encountered the algorithm cannot detect this fact, and the instances are categorized in a rather unpredictable manner. This problem is treated by Smyth and Mellstrom in (Smyth and Mellstrom, 1992) where they take decision trees and multi-layer neural networks as examples of discriminative models. As a solution to this problem they suggest *generative* or *characteristic* (Dietterich and Michalski, 1981) models, which are intended to discriminate the instances of the category from *all* other possible instances. These kind of models concentrate on the similarities between the members of the category, so that category boundaries are just an implicit by-product. Examples of such models are logic-based (depending on the learning algorithm, both discriminate and characteristic exist, see (Michalski, 1977)) and instance-based representations. Moreover, Smyth and Mellstrom make quite a provoking statement: "In fact one could even conjecture that *only* generative models can be truly adaptive and that discriminative models are impossible to adapt in an incremental on-line manner. This is certainly true in the general case for the class of discriminative models which includes decision trees and fixed-structure neural networks."

What about learning from discovery then? The experiments conducted so far have shown that such systems might work in a small, well understood, and predictable domain, but that it is very hard to make such systems suitable for real-world domains. Thus, despite the fact that learning by discovery is a very powerful learning method, it seems that (at least at the present stage of research) autonomous agents will have to manage without it.

The algorithms that learn from examples and by observation seem more adequate for learning bottom-up categories than top-down categories, whereas explanation-based learning algorithms are, more or less, designed to learn top-down concepts. Thus, a problem-solving agent may benefit from using EBL. However, as pointed out earlier, EBL do not form any new categories. The actual category formation step is when the high-level description is created during problem solving. This is, I believe, a not very well studied topic. EBL is then used to get a representation that can support the epistemological function. However, as Lebowitz has pointed out (Lebowitz, 1986), it is questionable whether there exist real-world situations where it can be applied, where the agent possesses all the background knowledge that is required to make the transformation into a low-level description.

A problem that I have not addressed is *noise* in the input. Noise is an inescapable problem in most real-world domains and has been addressed by some learning-from-examples systems. The solutions are often based on the assumption that the members of a category are rather similar. The problem with this assumption is that it is not compatible with the existence of atypical instances, in the sense that it becomes impossible to discriminate noise-laden instances from atypical ones.

6 Summary of Conclusions

The goal of this paper is, besides the reviewing of research in cognitive psychology and AI on different aspects of concepts, to investigate the issue whether psychological theories of concept acquisition can help us in constructing algorithms for concept acquisition by computer-based autonomous agents. However, as is evident from the research reviewed, it is from more fundamental aspects than acquisition, that influence from psychology has the potential of being most fruitful.

For instance, the functional aspects of concepts are hardly ever discussed within the AI society. Nevertheless, it seems clear that it is necessary to make a distinction between, at least, the epistemological, the metaphysical and the inferential function, whereas the other functions emerge, more or less, automatically.

AI researchers also have a very simplified view of the nature of categories. An autonomous agent in a real-world environment has to deal with real categories, not artificial ones. Furthermore, it is important to make a distinction between natural and derived categories since they must be acquired in different ways. Natural categories (natural kinds in particular) often arise merely when observing the world, whereas derived categories arise during problem solving activities. Concepts for representing natural categories are probably best learned by a similarity-based algorithm, whereas derived categories need a top-down algorithm. EBL is, in a sense, a top-down approach, but does not address the problem of *formation* of concepts.

As for the representation of concepts, we can conclude that a single and simple structure does not suffice to account for all the functions that concepts might have. Thus, an autonomous agent must have a complex (composite) concept representation. A suggestion for such a structure that supports the most important functions was presented in Section 4.3. It has an epistemological representation for perceptual (normal) categorization and an optional metaphysical representation for more "scientific" categorization. As we have seen, it seems that some kind of prototype-based representation is the best alternative for the epistemological representation, whereas a logic-based classical representation is the most appropriate for the metaphysical. To be able to reason and make predictions about the category and its members, the agent needs a large amount of encyclopedic knowledge. This is stored in the "associated glop". How this should be represented has not been discussed in detail, but some kind of logic-based representation seems appropriate. Moreover, to support stability and linguistic functions, the structure also includes an internal and an external designator.

As for the actual acquisition, it seems that the agent has to rely on learning from examples (if there some kind of teacher available), learning by observation and some method for forming derived (top-down) concepts. Learning from discovery seems too difficult for an agent in a real-world domain. Moreover, the learning must be incremental and not only concern one concept at the time. However, the most urgent topic for research is the integration of the different acquisition methods that already exist. The most interesting combination is perhaps learning from examples and learning from observation. Another demand on the learning algorithms is that they should learn characteristic concept representations, not discriminative. This demand disqualifies several popular algorithms such as TDIDT and back-propagation.

Finally, as has been pointed out earlier, the input to the learner in present AI-systems is usually *descriptions* of instances, consequently they deal with linguistic descriptions of the real world. Thus, the observations are on the linguistic level. Autonomous agents, on the other hand, have to deal with reality itself, making observations also on the perceptual level.[41] Especially, agents that are alone rely heavily on such observations, whereas communicating agents also make observations on the linguistic level.

[41] How these observations actually should be made is a problem that normally is studied within other fields such as *computer vision* (see for instance (Fischler and Firschein, 1987)). However, as the problem of concept acquisition is approached in this paper, it overlaps to some extent with these fields.

Acknowledgements

I wish to thank Eric Astor, Bertil Ekdahl, Peter Gärdenfors, Paul Hemeren, Måns Holgersson and Olof Tilly for helpful comments and suggestions.

References

Agre, P. and Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *AAAI-87*, pages 268–272.

Aha, D., Kibler, D., and Albert, M. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Amsterdam, J. (1988). Some philosophical problems with formal learning theory. In *AAAI-88*, pages 580–584.

Atkinson, R., Atkinson, R., Smith, E., and Hilgard, E. (1987). *Introduction to Psychology, ninth edition*. Harcourt Brace Jovanovic Publishers.

Balkenius, C. and Gärdenfors, P. (1991). Nonmonotonic inferences in neural networks. Technical Report LUCS 3, Cognitive Science, Lund University, Sweden.

Barsalou, L. (1986). Are there static category representations in long-term memory? *Behavioral and Brain Sciences*, 9:651–652.

Bergadano, F., Matwin, S., Michalski, R., and Zhang, J. (1992). Learning two-tiered descriptions of flexible concepts: The POSEIDON system. *Machine Learning*, 8(1):5–43.

Carpenter, G. and Grossberg, S. (1986). Neural dynamics of category learning and recognition: Attention, memory consolidation, and amnesia. In Davis, J., Newburgh, R., and Wegman, E., editors, *Brain Structure, Learning, and Memory*, pages 239–286. AAAS Symposium Series.

Corter, J. (1986). Relevant features and statistical models of generalization. *Behavioral and Brain Sciences*, 9:653–654.

de la Maza, M. (1991). A prototype based symbolic concept learning system. In *Eighth International Workshop on Machine Learning*, pages 41–45.

DeJong, G. and Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning*, 1(2):145–176.

Dietterich, T. and Michalski, R. (1981). Inductive learning of structural descriptions. *Artificial Intelligence*, 16(3):257–294.

Edelman, G. (1989). *The Remembered Present: A Biological Theory of Consciousness*. Basic Books.

Fischler, M. and Firschein, O., editors (1987). *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*. Morgan Kaufmann Publishers.

Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172.

Fisher, D. (1988). A computational account of basic level and typicality effects. In *AAAI-88*, pages 233–238.

Fisher, D. and Langley, P. (1985). Approaches to conceptual clustering. In *IJCAI-85*, pages 691–697, Los Angeles, CA.

Fodor, J. (1975). *The Language of Thought*. Thomas Y. Crowell.

Gärdenfors, P. (1990). Frameworks for properties: Possible worlds vs. conceptual spaces. *Language, Knowledge, and Intentionality, Acta Philosophica Fennica*, 49:383–407.

Gärdenfors, P. (1992). Three levels of inductive inference. Technical Report LUCS 9, Cognitive Science, Lund University, Sweden.

Ginsberg, M. (1987). *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann Publishers.

Gluck, M. and Corter, J. (1985). Information, uncertainty, and the utility of categories. In *Seventh Annual Conference of the Cognitive Science Society*, pages 283–287. Lawrence Erlbaum Associates.

Gross, K. (1988). Incremental multiple concept learning using experiments. In *Fifth International Conference on Machine Learning*, pages 65–72, University of Michigan.

Hanson, S. and Bauer, M. (1989). Conceptual clustering, categorization, and polymorphy. *Machine Learning*, 3:343–372.

Harnad, S. (1987). Category induction and representation. In *Categorical Perception*, pages 535–565. Cambridge University Press.

Holyoak, K. and Nisbett, R. (1988). Induction. In Sternberg, R. and Smith, E., editors, *The Psychology of Human Thought*. Cambridge University Press.

Jansson, C. (1987). *Taxonomic Representation*. PhD thesis, The Royal Institute of Technology, Stockholm, Sweden.

Kibler, D. and Aha, D. (1987). Learning representative exemplars of concepts. In *Fourth International Workshop on Machine Learning*, pages 24–30, Irvine, CA.

Kirsh, D. (1986). Second-generation AI theories of learning. *Behavioral and Brain Sciences*, 9:658–659.

Kruskal, J. and Wish, M. (1978). *Multidimensional Scaling*. Sage.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What categories reveal about the mind*. The University of Chicago Press.

Langley, P., Zytkow, J., Bradshaw, G., and Simon, H. (1983). Three facets of scientific discovery. In *IJCAI-83*, pages 465–468, Karlsruhe.

Lebowitz, M. (1986). Concept learning in a rich input domain: Generalization-based memory. In *Machine Learning: An AI Approach, Volume II*, pages 193–214. Morgan Kaufmann.

Lebowitz, M. (1990). The utility of similarity-based learning in a world needing explanation. In *Machine Learning: An AI Approach, Volume III*, pages 399–422. Morgan Kaufmann.

Lenat, D. (1976). *AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search*. PhD thesis, Stanford University.

Lenat, D. (1977). Automated theory formation in mathematics. In *IJCAI-77*, pages 833–842, Cambridge, MA.

Matthews, G. and Hearne, J. (1991). Clustering without a metric. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(2):175–184.

Medin, D. and Smith, E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35:113–138.

Michalski, R. (1977). A system of programs for computer-aided induction: A summary. In *IJCAI-77*, pages 319–320, Cambridge, MA.

Michalski, R. (1980). Pattern recognition as rule-guided inductive inference. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(4):349–361.

Michalski, R. (1987a). Concept learning. In *Encyclopedia of Artificial Intelligence, Volume 1*, pages 185–194. John Wiley and Sons.

Michalski, R. (1987b). How to learn imprecise concepts: A method for employing a two-tiered knowledge representation in learning. In *Fourth International Workshop on Machine Learning*, pages 50–58, Irvine, CA.

Michalski, R. and Chilausky, R. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing expert systems for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, 4(2).

Michalski, R. and Larson, J. (1978). Selection of most representative training examples and incremental generation of VL hypotheses: The underlying methodology and the description of programs ESEL and AQ11. Technical Report 877, Computer Science Dept., University of Illinois, Urbana.

Michalski, R. and Stepp, R. (1983). Learning from observation: Conceptual clustering. In *Machine Learning: An AI Approach*, pages 331–363. Springer-Verlag.

Mitchell, T. (1977). Version spaces: A candidate elimination approach to rule learning. In *IJCAI-77*, pages 305–310, Cambridge, MA.

Mitchell, T., Keller, R., and Kedar-Cabelli, S. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, 1(1):47–80.

Murphy, G. and Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.

Musgrove, P. and Phelps, R. (1990). An automatic system for acquisition of natural concepts. In *ECAI-90*, pages 455–460, Stockholm, Sweden.

Nagel, D. (1987). *Learning Concepts with a Prototype-based Model for Concept Representation*. PhD thesis, Rutgers, The State University of New Jersey.

Neisser, U., editor (1987). *Concepts and Conceptual Development: Ecological and intellectual factors in categorization*. Cambridge University Press.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press.

Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press.

Quine, W. (1969). *Ontological Relativity and other Essays*. Columbia University Press.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Rendell, L. (1989). Comparing systems and analyzing functions to improve constructive induction. In *Sixth International Workshop on Machine Learning*, pages 461–464.

Rey, G. (1983). Concepts and stereotypes. *Cognition*, 15:237–262.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B., editors, *Cognition and Categorization*, pages 28–49. Erlbaum.

Rosch, E. and Mervis, C. (1975). Family resemblances. studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Rosch, E., Mervis, C., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol.1: Foundations*. MIT Press.

Russell, B. (1912). *The Problems of Philosophy*. Oxford University Press.

Schank, R., Collins, G., and Hunter, L. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences*, 9:639–686.

Shepard, R. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 39:373–421.

Smith, E. (1986). Category differences/automaticity. *Behavioral and Brain Sciences*, 9:667.

Smith, E. (1988). Concepts and thought. In Sternberg, R. and Smith, E., editors, *The Psychology of Human Thought*. Cambridge University Press.

Smith, E. and Medin, D. (1981). *Categories and Concepts*. Harvard University Press.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11:1–74.

Smyth, P. and Mellstrom, J. (1992). Detecting novel classes with applications to fault diagnosis. In *Ninth International Workshop on Machine Learning*, pages 416–425.

Solso, R. (1991). *Cognitive Psychology, third edition*. Allyn and Bacon.

Stepp, R. and Michalski, R. (1986). Conceptual clustering: Inventing goal-oriented classifications of structured objects. In *Machine Learning: An AI Approach, Volume II*, pages 471–498. Morgan Kaufmann.

Subramanian, D. (1989). Representational issues in machine learning. In *Sixth International Workshop on Machine Learning*, pages 426–429.

Towell, G., Shavilik, J., and Noordewier, M. (1990). Refinement of approximate domain theories by knowledge-based neural networks. In *AAAI-90*, pages 861–866.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84:327–352.

Utgoff, P. (1986). Shift of bias for inductive concept learning. In Michalski, R., Carbonell, J., and Mitchell, T., editors, *Machine Learning: An AI Approach, Volume II*. Morgan Kaufmann.

Vaina, L. and Jaulent, M.-C. (1991). Object structure and action requirements: A compatibility model for functional recognition. *International Journal of Intelligent Systems*, 6:313–336.

Watanabe, S. (1969). *Knowing and Guessing: A Quantitative Study of Inference and Information*. John Wiley and Sons.

Winston, P. (1975). Learning structural descriptions from examples. In *The Psychology of Computer Vision*, pages 157–209. McGraw-Hill. Also in Readings In Knowledge Representation, ed. R. Brachman and H. Levesque, Morgan Kaufmann, 1985.

Wittgenstein, L. (1953). *Philosophical Investigations*. Basil Blackwell.